Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces

Mohan M. Trivedi, Kohsia S. Huang, and Ivana Mikic

Computer Vision and Robotics Research Laboratory
University of California at San Diego
La Jolla, California, USA 92093
http://cvrr.ucsd.edu/
[mtrivedi/khuang]@ucsd.edu

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, tele-conferencing or performance spaces. The paper includes an overall system architecture to support design and development of intelligent environments. Details of panoramic (omni-directional) video camera arrays, calibration, video stream synchronization, and real-time capture/processing are discussed. Modules for multicamera based multi person 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.

Keywords: Ambient Intelligence, Smart rooms/spaces, human-machine interfaces, multi-camera systems, real-time vision, active vision, person tracking, face detection/recognition, body modeling, event analysis, activity summarization,
1. 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 human 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.

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. Figure 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 spatial domain. We focus on objects, which are defined using video sensory modality. Examples of such objects would be “person” or “face”.

3) **Event**: This is a “pattern” defined in spatial-temporal domain. We consider events using both audio and video modalities. Examples of events can be “person entering/leaving a room”, or “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 “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.

---

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

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, 3, 4, 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, 13]. Integrated sensory modalities of audio and video [14, 15, 16, 17, 18, 19] are also been 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 [20]. We are currently embedding distributed video networks in rooms, laboratories, museums, and even outdoor public spaces.
in support of experimental research in this domain [21]. 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 in distance learning, teleconferencing, entertainment, and smart homes.

In this paper we present 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 a computational framework to help design distributed video arrays 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.

2. Distributed Video Arrays for Intelligent Environments

Distributed video arrays, (“DIVA”), in an intelligent environment should be able to detect the presence of people, track their movements, recognize them and understand their actions. For recognition of complex actions, high-resolution images of the human body (or even body parts) are necessary. A system capable of monitoring a large space should therefore be able to acquire such high-resolution video anywhere in the environment to allow unconstrained human movement. For that reason, some computer vision groups have equipped their laboratories with large numbers of static cameras [22, 23] with the idea of obtaining very detailed information about the monitored space.

However, 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, depending on the type of activity being sensed. Detecting a person’s presence or recognizing whether they are sitting or standing requires less detailed information than estimating the direction the person is pointing their finger to. Based on this observation, we propose a system that continuously monitors the environment at low resolution – detecting only the presence of people and their location. 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 the focuses of attention of the system. Equipped with a few static wide-angle view cameras, the low resolution monitoring of the environment can be achieved. With a small number of active (PTZ – pan/tilt/zoom) cameras, multiple simultaneous focuses of attention could be maintained. Using this approach, robust monitoring of the entire environment can be achieved with fewer cameras and computational resources.
In this section we discuss development of distributed video arrays, which support a wide range of tasks of intelligent environments. Key features of these "smart" video arrays are:

- Ability to derive semantic information at multiple levels of abstraction.
- Ability to be "attentive" to specific events and activities.
- Ability to actively shift the focus of attention at different "semantic" resolutions.
- 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 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 3D at multiple resolutions are essential parts of the proposed system. At low-resolution level, locations of all people in the environment will be continuously monitored. 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.

Figure 2. Distributed video arrays for tracking, human identification, and activity analysis.
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 people’s 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 Bayesian net [24]. 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 generate 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 speed, accuracy, 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 multi-level abstraction can be further generalized to include many other applications such as object recognition, facial expression recognition, 3D human body modeling and tracking [25], and even intention estimation and prediction.

3. 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. In this section we discuss the experimental system architecture of our intelligent space complex.

3.1 Intelligent Environment Research Complex

The intelligent environment research complex at the CVRR Laboratory of UCSD is shown in Figure 3. It includes two separate but connected rooms appropriately instrumented and suitable for a wide range of experimental research. The first one is called “AVIARY” (for Audio Video Interactive appliances, rooms and systems) was designed to be a small conference room. The second space called “MICASA” (for Multimodal Interfaces, and Context Aware Spaces) was designed to be a classroom or a performance chamber. We present a brief overview of these testbeds below.
Figure 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, 8 omnidirectional cameras, 8 PTZ cameras, and eight microphones, which are embedded in the room.

The audio-video sensory suite for the AVIARY includes a network of four omnidirectional cameras, four pan-tilt-zoom (PTZ) and four static rectilinear cameras, and eight microphones is installed in the room. It is equipped with four omnicameras or ODVSs 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 pan-tilt-zoom 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 pan-tilt-zoom (PTZ) dynamic rectilinear cameras are installed at the four corners about 1.4m 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 Figure 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 recall. 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 interface systems of these two modalities can be very natural and effortless for the users. Robustness to 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.

MICASA is a much larger testbed. 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 Figure 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 the eight overlap each other by approximately a $2m \times 3m \times 2.5m$ 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 2D 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 6 of the 8 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 side of the room for classroom setup, as shown in Figure 3.

![Figure 4: MICASA static rectilinear camera array placement.](image)

### 3.2 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 coordinate system [32]. The calibration is carried out in advance and parameters are stored in the computers.

The calibration of ODVSs 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 needs 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.

### 3.3 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, two 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. This by nature synchronizes frames from the four cameras. 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.

![Figure 5: The architecture for synchronous video capturing using quad.](image)

The second approach is to synchronize the image grabbing on the computers. This way guarantees full frame capturing for high-resolution demands. 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, as shown in Figure 6. 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 tri-state 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 8 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 Figure 7.
Figure 6: The primary-secondary architecture for full-frame synchronous capture of multiple video streams.

Figure 7 Workstations that perform synchronous grab from multiple cameras.
3.4 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 Figure 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 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 multi-threaded programming, and the thread synchronization is shown in the timing diagram in Figure 8. Tracking could be running on one computer and the trajectories are communicated to other machines through network sockets. 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 multi-state models. Those processes achieve minimum delay and optimal efficiency by carefully synchronized multi-threading.

Figure 8: System processes and multi-thread synchronization for active event capturing.

The features of the system architecture of our current intelligent complex testbed are summarized in Table 1.

Table 1: Summary of the intelligent complex setup.

<table>
<thead>
<tr>
<th></th>
<th>AVIARY</th>
<th>MICASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>6.7m × 3.3m × 2.9m</td>
<td>6.7m × 6.6m × 2.9m</td>
</tr>
<tr>
<td>Video Array</td>
<td>4 Omnicams</td>
<td>4 Omnicams</td>
</tr>
<tr>
<td></td>
<td>4 Rectilinear</td>
<td>8 Rectilinear</td>
</tr>
<tr>
<td></td>
<td>4 PTZ</td>
<td>4 PTZ</td>
</tr>
<tr>
<td>Video Synchronization</td>
<td>Quad synchronized</td>
<td>Quad or hardware synchronized</td>
</tr>
<tr>
<td>Audio Array</td>
<td>2 microphone arrays</td>
<td>Lapel</td>
</tr>
<tr>
<td>Processors</td>
<td>1 for tracking</td>
<td>1 for tracking</td>
</tr>
<tr>
<td></td>
<td>1 for event analysis</td>
<td>6 for voxelization and event</td>
</tr>
<tr>
<td></td>
<td>1 for video &amp; audio archiving</td>
<td>analysis</td>
</tr>
</tbody>
</table>
4. Tracking and Analysis of Humans in Intelligent Environments: Experimental Studies

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 multi-person tracking as well as for human body analysis and give experimental results.

4.1 Study 1: Multi-Person Tracking using Video Arrays

We have developed a real-time intelligent room system, the Networked Omni Video Array (NOVA) system as shown in Figure 9 which utilizes the omnican array for tracking, face capture, and face recognition [6]. The 3D 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. A $64 \times 64$ face video is then extracted from the perspective view to be identified, as shown in Figure 10. Demonstration clips of person and face tracking on the ODVS array is available at http://cvrr.ucsd.edu/pm-am/demos/index.html. 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.

Figure 9: Functional blocks of the NOVA intelligent room system.
The omnicam-based person tracker is described in [27, 6]. We evaluated the accuracy of the 3-D multiple omnicamera tracker by tracking people on a specified path. The tracking ground truth is a rectangular walking path designated on the floor of the AVIARY testbed. We have tested up to 4 people, both children and adult volunteers. The logged tracking paths are later retrieved, analyzed, and plotted offline for comparison.

The accuracy of 3D tracking is evaluated in terms of offsets of the tracks from the ground truth and the corresponding standard deviation. Those indices are experimented on single person, multiple-2 people, multiple-3 people, and multiple-4 people cases. For each case, children (except multiple-4 people) and adults are tested. An example of the tracking result of tracking two children is presented in Figure 11, and the offsets and standard deviations for all experimented cases are listed in TABLE 2.

From the experimental results, the O-VAT 2D standard deviations and track offsets are almost constant after 2 people, and the magnitude of change from 1 to 2 people is quite small. The O-VAT height estimation is excellent and height standard deviation is small and independent of the number of people. Therefore O-VAT is robust to the number of people. We also evaluated the static rectilinear video array tracker during the experiments. It turns out that the 2D standard deviation increases rapidly after 3 people, especially for adult cases, and the height estimate is inaccurate and degrades rapidly with the number of people in the rectilinear cases. It is due to the fact that for outside-in rectilinear coverage, the chance of people occluding each other increases rapidly with the number of people. This situation is less likely to happen on O-VAT because the inside-out ODVSs are standing upright and people walking around them can be easily distinguished in the ODVS images.

Figure 10: Left video: face tracking by ODVS perspective view generation. Right video: face video extraction and face recognition.

Figure 11: Accuracy evaluation of the O-VAT. The tracking targets are two subjects (small children). The floor plan on the left shows the 2D tracking accuracy. Red dash lines are the designated walking paths on the floor. Different tracks of the volunteers are color-coded. The height tracking on the right is plotted against time. The actual heights of the volunteers are shown as red dash lines and the height estimates of the subjects are plotted.
Table 2: Summarization of tracking accuracy in terms of track offsets and standard deviations. 2D tracking and height estimates are evaluated on single person and multiple people cases. Height estimate is not valid for an adult since they are taller than the camera coverage.

<table>
<thead>
<tr>
<th>Tracking Accuracy (cm)</th>
<th>x-y (2D)</th>
<th>Height z</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>∆µ</td>
<td>σ</td>
</tr>
<tr>
<td>Single subject</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Adult</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Multiple Subjects</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Adult</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Child</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Adult</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult</td>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

O-VAT runs at approximately 20 frames per second on a platform of dual Pentium III 866MHz, 256MB RAM, Windows NT 4.0. Due to the moving-average track filtering, there is a delay of about 300 ms. This delay can be fixed by using Kalman filter. Therefore it is very suitable for real-time applications.

In term of system flexibility, ODVS network has good reconfigurability because with the same ODVS network, the system not only allows tracking but also allows electronic pan-tilt-zoom (PTZ) for higher level processing simultaneously. Note that electronic PTZ does not require mechanical control, which leads to delay and damping issues, as will see next. 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 is 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 meeting room setup.

4.2 Study 2: Face, Body and Movement Analysis

Head Tracking and Face Capture

Results of tracker are used to control the face capture module. As shown in Figure 9 and Figure 10, 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 Figure 12. The O-VAT uses the ODVS array on the ceiling in the MICASA testbed to track people in 3D. The location of head is then estimated and used to drive a PTZ rectilinear camera. From the video sequence in Figure 12, it can be seen that mechanical PTZ would have some control delay issue, and the human motion speed needs to be limited.
Figure 12: Visual servoing for face capture on a mechanical PTZ camera driven by the 3D omni-video array tracker. 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.

**Single-Frame Face Detection and Recognition**

The captured video is then processed to detect the face and extract the face video, as shown in Figure 13. From the head tracking output, skin tone segmentation is first used to fine the face candidates. Possible face images are cropped from the skin tone blobs. Those images are then classified to reject non-faces. A simple eigenface method is used for both face classification and single-frame face recognition [2]. The outputs of this module are the stream of projection vectors of the face video in the eigenface feature subspace as well as the stream of face recognition identity. The stream of feature vectors can be further processed to estimate the face orientations and recognize person over the frames.

**Face Orientation Estimation**

From the single-frame feature vectors, face orientation can be estimated robustly across frames [28]. We construct a continuous density hidden Markov model (CDHMM) to model the dynamic within the feature vector stream. The $N$-state Markov chain has a linear left-right topology. The states represent the underlying face orientations from left facing to right facing. The state transition is restricted to be linear since the facing angle should not change abruptly with a large magnitude. Each state has an observation density function, which is a mixture of $M$ multi-dimensional Gaussian densities. From the PCA-based feature vectors, the first $d$ components are used by the Gaussian densities. So given a sequence of captured face images, we can estimate the face pose sequence over frames from the state sequence, which is obtained by MAP estimation or by Viterbi algorithm [10].
There are several reasons to take this approach. Due to inaccuracies of face tracking and face detection, the face is usually not aligned in the center of the captured image, and the illumination is not constant since the person is walking around in the space. In some extreme cases half of the face may be too bright or too dark. Also during some frames the face may not be detected due to occlusion, turn-away, and camera noise issues. To overcome these issues, we have to take streaming-type estimation scheme instead of single-frame based methods. By accumulating the likelihoods of the frames and interpolating between frames, a smoothed sequence of face orientation is achieved. This smoothing is carried out by the Markov chain within the HMM. The need for robust face orientation estimation is to assess the direction of attention of people in the intelligent environment. This information can be used in the estimation of the behaviors of people.

In our experiments, we first collect sequences of captured face video with the frame-to-frame ground truth of face orientation. Some of the sequences are used for training the CDHMM, the others are used to test the accuracy. We train the CDHMM on different number of states $N$, Gaussian mixtures $M$, and utilized dimension $d$ of the feature vector and pick the set that gives the minimum root-mean-squared error (RMSE) between the estimated face orientation sequence and the ground truth sequence. The training initialization of observation density is by $k$-means vector quantization of the training feature vectors and assign them to the $N$ states according to the corresponding facing angle from left to right. The covariance of the Gaussian densities is initialized to a identity matrix with diagonal values $\sigma$. After CDHMM training, the facing angle of the $N$ states are determined by inputting the training sequences to the model and average the corresponding ground truth angles for each single state in the output state.
sequence. Then on the testing phase only the captured face video is fed to the model to estimate the face orientation sequence.

**Streaming Face Recognition**

Similar concept can be applied to streaming face recognition [29]. 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 Figure 14, accumulating the confidence across frames in the face video stream will enhance the recognition robustness. As shown in Figure 15, the captured face video is partitioned into segments and the single-frame subspace analysis outputs both recognition identities and feature vectors of the frames. These outputs are classified by three schemes. The majority rule decides on the highest occurrence of the single-frame recognition identities in the segment, the discrete HMM (DHMM) rule decides on the sequence pattern of the single-frame recognition identities, while the continuous density HMM (CDHMM) rule decides on the sequence pattern of the single-frame feature vectors in the subspace.

Figure 14: 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.

Figure 15: The streaming face recognition scheme and the geometric interpretation in feature subspace.

The parameters of the decision rules include the number of states \( N \) in the DHMM and CDHMM, the number of Gaussian mixture \( M \) for CDHMM, utilized dimension \( d \) of the PCA feature vector with full dimension \( D=135 \), and sequence length \( L=49 \). 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 sequences of all sequences in the streaming cases. Table 3 compares the best OCPs of the recognition schemes. This outcome justifies the streaming type processing schemes.
because accumulating a segment of distribution across frames would provide a better match to a pre-trained class in the feature subspace than only one single frame, as illustrated in Figure 15. It also suggests that the decision of classification should be postponed to the final stage, as in the case of CMD rule. This delayed decision has been adopted in the design of all the streaming face processing modules in the NOVA system.

Table 3: Comparison of the best OCPs of the single-frame face recognition and the SFR rules.

<table>
<thead>
<tr>
<th>Decision Rules</th>
<th>Best OCP</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-Frame FR</td>
<td>75.9 %</td>
<td></td>
</tr>
<tr>
<td>SFR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAJ</td>
<td>81.7 %</td>
<td></td>
</tr>
<tr>
<td>DMD</td>
<td>89.7 %</td>
<td>N=14</td>
</tr>
<tr>
<td>CMD</td>
<td>99.0 %</td>
<td>N=1, M=1, d=8</td>
</tr>
</tbody>
</table>

**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 Figure 6. Within the volume, voxel reconstruction of human body can be carried out in real-time. As shown in Figure 16, the 2D 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 3D 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 2D silhouettes has a pixel corresponds to it through camera calibration model. After the 3D voxel reconstruction is available, first an ellipsoid is searching through the surface voxels to fit the head, then torso and the limbs are fitted afterward. 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 [25]. The demonstration video clips are available at http://cvrr.ucsd.edu/pm-am/index.html. Some of the joint angles of the body model are plotted in Figure 17 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.
Figure 16: Body modeling system components. Video from four cameras is segmented and 3D 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.

Figure 17: Comparison of multiple joint angle patterns of the body model when tracking two walking subjects.
5. Human actions and interactions in an Intelligent Meeting Room

Intelligent meeting rooms (IMR) are spaces, which support efficient and effective interactions among their human occupants. 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.

Figure 18. 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.

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:

- between active participants – people present in the room
- between the system and the remote participants
- 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. Other two 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 spatiotemporal 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 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 recognize and capture. We identified three: a person located in front of the whiteboard, a lead presenter speaking and 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. Other two 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:

a) 3D tracking of centroids using static cameras with highly overlapping fields of view
b) Person identification (face recognition, voice recognition and integration of the two modalities)
c) Event recognition for directing the attention of active cameras
d) Best view camera selection for taking face snapshots and for focusing on the face of a current speaker
e) Active camera control
f) 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 3D 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 [30, 31].

The system block diagram is shown in Figure 19. 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 [32]. Two PC computers are used. One performs 3D 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 3D 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.
Event Recognition for Directing the Attention of Active Cameras

This module constantly monitors for events described in the section 2. When a new track is detected in the room, it is classified as person or object depending on the dimensions of the bounding cylinder. This classification is used to permanently label each track. If classified as object, the camera closest to it takes the snapshot. If classified as 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 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.

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 (Figure 20). Center of the face is taken to be 20cm 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 the person not in front of the whiteboard speaks. In these three situations, we use different assumptions in estimating the direction the person is facing.

Figure 20. 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)
Figure 21. Person standing close to the whiteboard draws attention from one active camera

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 (Figure 21). 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. Due to the hairline, the ellipse fitted to the skin pixels changes orientation as person turns from far left to far right (Figure 22). We use skin-tone detection algorithm. If skin pixels are regarded as samples from a 2D Gaussian distribution, the eigenvector corresponding to the larger eigenvalue of the $2 \times 2$ covariance matrix describes the orientation of the ellipse. A lookup table based on a set of training examples (Figure 23) is used to determine the approximate angle between the direction the person is facing and the direction connecting the person and the camera that the whiteboard image was taken with. These angles are not very accurate, but we have found that this algorithm works quite reliably for purposes of best view camera selection. In the third case, where person elsewhere in the room is speaking, 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. The first image obtained with the chosen camera is processed using the algorithm described in the previous paragraph and the camera selection is modified if necessary.

Figure 22. Face orientation estimation for best view camera selection

Figure 23. Lookup table for face orientation estimation computed by averaging across training examples.
**Active Camera Control**

Pan and tilt angles needed to bring the point at 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 be corrected by the pan and tilt angles between the center of the image and the zoom center. 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 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 [33]. Two images taken with two different zoom values are compared by shrinking the one taken with the larger zoom using the Equation 1. The value of magnification (will be smaller than 1) that achieves best match between the two images is taken to be the inverse of the magnification between the two images. The algorithm was written for outdoor cameras where objects present in the scene are more distant from the camera than in the indoor environments. Therefore, instead of comparing images at different zooms to the one taken at zero zoom as done in [33], we always compare two images that are one zoom step apart. The absolute magnification for a certain zoom value with respect to zero zoom is computed by multiplying magnifications for smaller zoom steps. However, we could not reliably determine the location of the zoom center using this algorithm. Instead, we determine its coordinates 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.

**Graphical Summarization/User Interface Module**

The tracks, identities, and events are logged into a database as shown in Table 1 and the audio and video are also recorded for later retrieval. A summarization interface as shown in Figure 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.
Experimental trials confirm satisfactory performance of the system. Person tracking module performed with maximum errors around 200mm. These experiments included a 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 seconds. The results are very good for clips with low silence percentage even for shorter clips, but gets erroneous when silence is more than 50% of the clip. However, there is a delay of 1-5 seconds 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.

If the person faces the direction they are walking, the camera selection for acquisition of face snapshots also works with perfect accuracy. It would, of course, fail if person turned their head while walking. 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. The systems stores all information needed to access appropriate parts of the video that correspond to the events the user selects from the interface. From the interface, the user can view identities and face snapshots of people associated with different tracks by clicking on the corresponding
colored shape. 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 summarization graph.

**Table 4: Event log database for activityarchiving and recall. Entries are logged when there is change of the states. (K=Kohsia, I=Ivana; and M=Mohan)**

<table>
<thead>
<tr>
<th>Time Stamp</th>
<th>Person ID (Location)</th>
<th>Speech Activity</th>
<th>IMR state / Context (# Occupants)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Vacant</td>
</tr>
<tr>
<td>1</td>
<td>K (3.3, 0.2)</td>
<td>0</td>
<td>Occupied (1)</td>
</tr>
<tr>
<td>2</td>
<td>K (2.5, 0.7)</td>
<td>0</td>
<td>Occupied (2)</td>
</tr>
<tr>
<td>3</td>
<td>K (1.5, 1.3)</td>
<td>0</td>
<td>Occupied (2)</td>
</tr>
<tr>
<td>4</td>
<td>K (0.5, 2.1)</td>
<td>0</td>
<td>Occupied (2)</td>
</tr>
<tr>
<td>5</td>
<td>K (0.4, 2.0)</td>
<td>K: Presenting I: Listening</td>
<td>Presentation (2)</td>
</tr>
<tr>
<td>8</td>
<td>K (0.4, 2.1)</td>
<td>K: Presenting I: Listening M: Listening</td>
<td>Presentation (3)</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>11</td>
<td>K (0.4, 2.1)</td>
<td>K: Listening I: Speaking M: Listening</td>
<td>Discussion (3)</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>15</td>
<td>K (0.4, 2.0)</td>
<td>K: Presenting I: Listening M: Listening</td>
<td>Presentation (3)</td>
</tr>
<tr>
<td>20</td>
<td>K (0.5, 2.2)</td>
<td>0</td>
<td>Occupied (3)</td>
</tr>
<tr>
<td>22</td>
<td>K (0.6, 2.1)</td>
<td>K: Listening I: Listening M: Speaking</td>
<td>Discussion (3)</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

**6. 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 (omni-directional) video camera arrays,
calibration, video stream synchronization, and real-time capture/processing are discussed. Modules for multicamera based multi person 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.

Slowly but certainly the spaces which humans live in are getting increasingly more technologically sophisticated. Sensors, computers, web appliances, communication devices, and displays/monitors are commonly found in our work places, entertainment spaces, hospitals, and homes. Intelligent environments offer the new generation of interfaces between humans and the spaces they live in.

Acknowledgements

Our research is supported by a number of sponsors, including the California Digital Media Initiative and UC Discovery Grants projects, Technical Support Working Group of the U.S. Department of Defense, Sony Electronics Corp, Compaq Computers, Daimler Chrysler Corporation. We are thankful for their support. The authors would also like to thank their colleagues form the Computer Vision and Robotics Research Laboratory, especially Mr. Nils Lassiter, Dr. Kim Ng, Mr. Rick Capella, Mr. Shinko Cheng for their valuable contributions and collaboration.

References