Intelligent Environments and Active Camera Networks

Mohan Trivedi, Kohsia Huang, Ivana Mikic
Department of Electrical and Computer Engineering
University of California, San Diego

Abstract

Intelligent environments provide challenging research problems for natural and efficient interfaces between humans and computers as well as between humans. In this paper we present a multimodal sensory intelligent system testbed based on some general requirements for developing intelligent environments. We also present rigorous experimental investigations on the processing and control modules for the active camera networks and the microphone array which are embedded in the intelligent room. An integrated intelligent system is developed utilizing four basic modules for visual and audio processing. The integrated system has the functionality of human tracking, active camera control, face recognition, and speaker recognition. This system is demonstrated to be suitable for teleconferencing type of applications.

Introduction

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, and play. Recent research on intelligent environments provides numerous new challenges in the fields of machine perception. In computer vision [1], distinct progress in face recognition [2, 3], people tracking [4], and gesture recognition [5] has been made in the last decade. For audio, much progress has been made in speaker and speech recognition [6] and source localization [7, 8]. Integrated sensory modalities of audio and video [9, 10, 11] 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 [12].

Our team is pursuing investigations for systematic development of intelligent environments where networks of cameras and microphone arrays serve as the sources of multimodal sensory information [8]. The research challenges are to make computers intelligent so that they can:

- Develop and maintain an awareness of their 3-D environment,
- Acquire and respond to the voice and visual inputs from the users in a robust manner,
- Adapt to the dynamic changes in their surroundings, and
- Interact in a natural and flexible manner with the users.

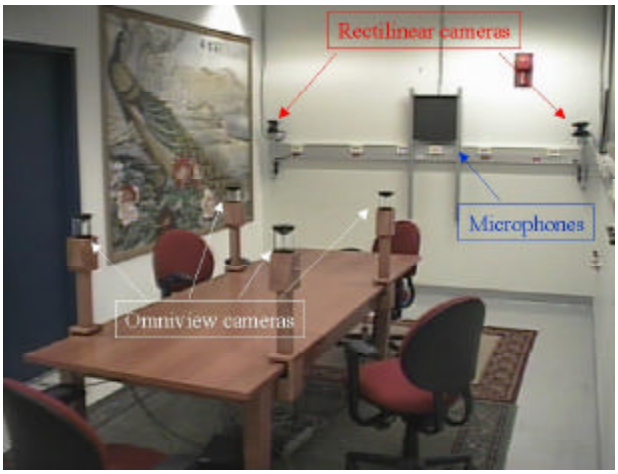

Figure 1 The AVIARY: Audio-Video Interactive Appliances, Rooms and sYstems. This room is built for experimental development and evaluation of the intelligent room systems utilizing four rectilinear cameras, four omnidirectional cameras, and eight microphones which are embedded in the room.

Based on this scenario, we built a multi-purpose intelligent room testbed called AVIARY, Audio-Video Interactive Appliances, Rooms and sYstems, as shown in Figure 1. Currently, a network of four omnidirectional cameras, four pan-tilt-zoom (PTZ) and four static rectilinear cameras, and eight microphones is installed in the room. This room 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. Therefore to meet these requirements, systems need to be equipped with the following capabilities:

1. **Self-Calibrating:** Systems need to be able to self-calibrate with least technical expertise from the users when installed on different locations.
2. **Adaptive:** Systems must have the ability to adapt. This provides the systems the ability to deal with changes that take place in the environment.
3. **Multimodal**: Systems must be equipped with proper information processing capabilities. This includes not only superior unimodal processing capability, but also well integrated multimodal sensory information processing capabilities. This will result in more robust systems with broader range of functionality as well as unobtrusive interfaces to the users.
4. **Interactive:** Systems must have the ability to direct their attention to "interesting" events and make appropriate reactions. This can be achieved by utilizing a semantic event databases to guide human-environment interactions. Semantic event databases store abstracted past events of the intelligent environment with the embedded active sensory networks. These events can be flexibly queried for decision making to inform humans or send commands to the active sensory networks and the processing algorithms.

Note that handheld or head-mounted microphones are not appealing if mobility of the user is allowed. Multiple microphones as in AVIARY setup offer an attractive alternative. However, mobility comes with a price, and poses these challenges:

- The received audio signal is distorted by the room acoustic properties, i.e., reverberations. This not only degrades the quality of the audio signal, but also interferes any post-processing such as speech recognition. Recovering the speech signal requires complicated room acoustics modeling and a difficult deconvolution problem. More importantly, the room acoustical properties are dependent on the position and gesture of the users, so it can vary widely and hard to account for in advance. Self-calibration is essential for the system to function in such diverse environments. Moreover, ambient noise is not stationary both spatially and temporally, hence requiring systems to be adaptable.
- Vision system has to operate efficiently and robustly to conditions and events of the dynamic world. Novel approaches for real-time 3D modeling, visualization, and rendering are required. The tracking of the speaker also has to be robust and fast.

Basically the interactive systems developed in the AVIARY operate in two modes: **(1) initialization mode** and **(2) active mode.** The system initialization refers to the phase when the system is learning its environment by carrying out self-calibration to prepare for its run-time functions. This initialization phase contributes to system robustness by orienting system model to environmental variables. During the active mode operations, the system implements useful functions such as acquiring and interpreting audio-visual signals thereby enabling effective interactions and providing a convenient user interface.

Omnidirectional camera network and audio processing modules

The omnidirectional vision sensor (ODVS) network embedded in the AVIARY has proven to be most useful for a number of tasks required in the system initialization and active mode operations. We have developed a multiple-, wide-baseline stereo system for accurate geometric modeling of the room [13]. We have also developed a visual modeling system, where accurate 3-D range information and color information are simultaneously extracted [14]. The ODVS network also allows us to track multiple people [15] and to dynamically generate the views associated with the movement of the persons in the room [16, 17], as shown in Figure 2. The audio information is acquired using an array of eight microphones and robust techniques for room acoustic modeling and speaker localization are being developed [8].

Rectilinear camera network and multimodal person identification

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

1. A robust and efficient human tracking module which utilizes a four rectilinear camera network.
2. An active camera control for capturing frontal view of a person moving in the room. The module performs camera selection for best view and also automatic panning, tilting, and zooming for taking a close-up image.

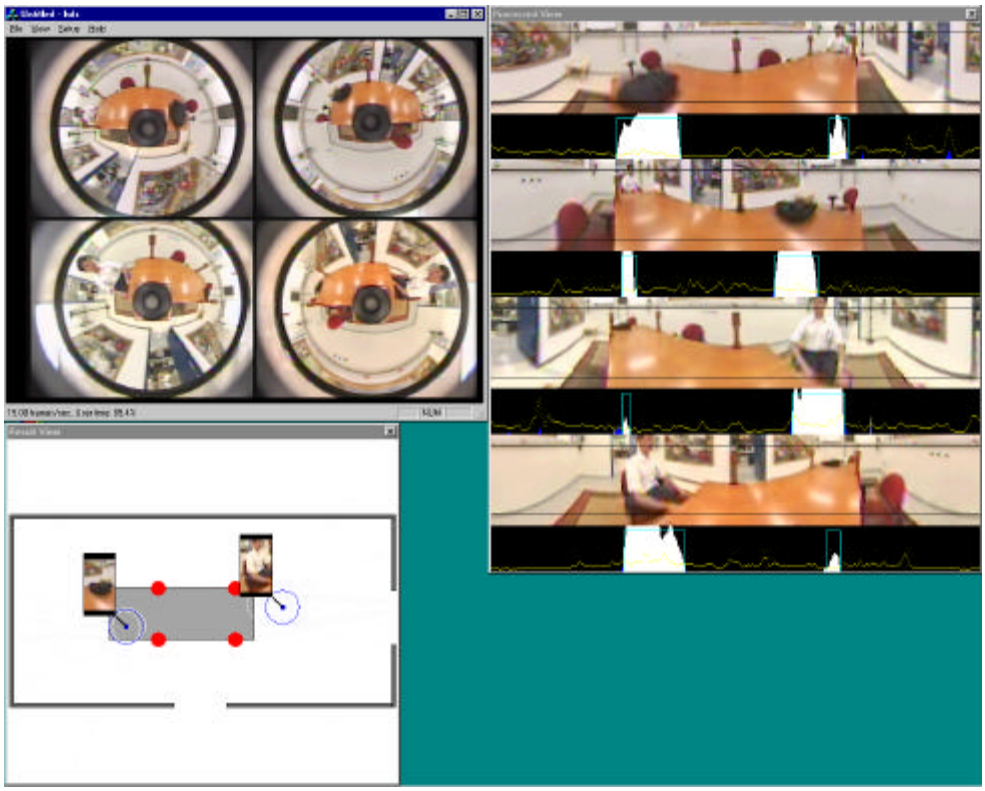

Figure 2 The ODVS tracker. The upper left window shows the raw image from four ODVS cameras. Next to it on the right are the unwrapped images and histograms for object/person detection. The bottom left window is the planar view of AVIARY room and the dynamic views of the detected object/person.

3. A face recognition module that takes the close-up image to recognize the face in the image.
4. A speaker recognition module which utilizes the microphone array and operates in parallel with the face recognition module for robust and active person identification.

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

Video Segmentation Module. The segmentation is based on background subtraction. First and second order statistics for background pixels are continuously updated. Due to the use of a forgetting factor, background model is adaptable to slow changes. Foreground pixels are segmented using the Neyman-Pearson test and grouped into blobs. Blob centroids for all cameras are computed and serve as input to the tracking algorithm. See [18] for details of the segmentation algorithm.

The Multi-Camera 3-D Tracking Module. The role of the tracker is to track multiple objects in 3D using segmentation results from different cameras with highly overlapping fields of view. The cameras are calibrated using Tsai's algorithm. For details of the algorithm, see [19]. The tracker is capable of tracking multiple objects simultaneously. It maintains a list of Kalman filters, one for each object in the scene. The real-time nature of the system requires the tracker to produce updated and predicted positions of each object for the current frame. Also, the availability of up-to-date prediction allows us to feed back the information to the segmentation algorithm, which can increase its sensitivity in the areas where objects are expected to be present.

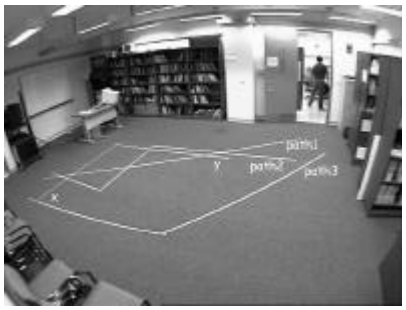

Figure 3 The world coordinate system and the measured paths.

We evaluated the accuracy of the tracking algorithm on real data. Three different paths were measured and marked on the floor of a room as illustrated in Figure 3. First set of sequences was recorded with three different persons walking each of the paths. For the second set, two people would walk at the same time on two different paths.

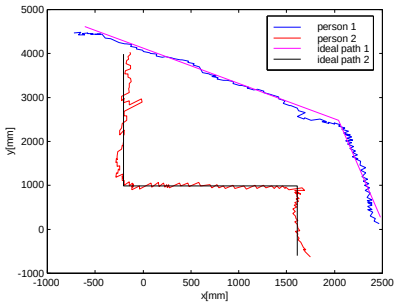

Figure 4 xy plots of tracks for two people in the scene.

Figure 4 shows ideal and measured paths for one experiment with two people. The tracks very accurately followed the ideal paths with maximum error in all experiments being around 200 mm and average error around 30 mm. That is a very good accuracy, especially if it is taken into account that the error in these experiments is influenced by calibration and segmentation errors and the errors in measuring the path, drawing it on the floor and walking on it.

Active Camera Control Module. Four pan/tilt/zoom cameras (Canon VC-C3), controlled through RS-232 links, are used in this system. Since the cameras are calibrated, it is trivial to compute pan and tilt parameters that bring an object at a known location to the center of view. Calibration of the zoom was done using the method described in [20], which enables us to find the zoom value that makes the size of the face in the image sufficient for recognition.

The current use of the PTZ capabilities of cameras is for taking snapshots of objects and of people's faces for face recognition and archiving. We also plan to use cameras for focusing of attention of the video-processing part of the system to "interesting" objects and locations. As a new track appears in the room, it is classified as object or person based on its shape and size. If person is detected, the location of the head is estimated to be at the top fifth of the height. Best view camera is chosen to be the one for which the negative inner product of the viewing direction and average person's velocity in the last few frames is the largest. In other words, the camera that the person was walking toward in the recent frames is chosen to be the one that the person is most likely facing, as shown in Figure 5. If the track is classified as object, the snapshot of the whole object is taken and stored.

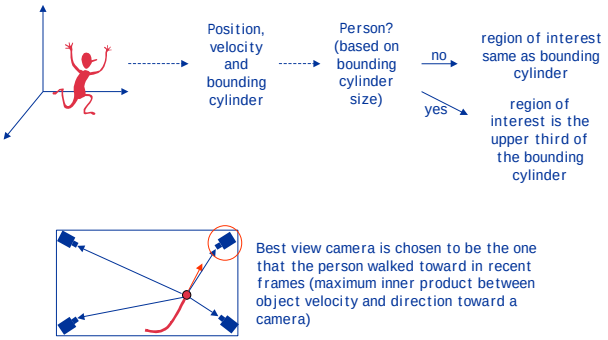

Figure 5 Best view camera selection for taking face snapshots.

Joint visual-acoustic person recognition modules

Integration of audio and video modalities increases robustness of person recognition through the redundancy of sensor information and the complementary functionality.

Face Recognition Module. Face recognition is a visual biometric modality. Eigenface recognition algorithm [21] is currently utilized in the face recognition module. Human face is extracted from the snapshot image of camera network by skin color detection [22]. Face images of known people on certain facing angles are stored into the training face database. The training faces are then used to span the eigenface space by the eigenvectors of the correlation matrix of the training face vectors. This is done on system initialization. During active mode, the test face image is projected into the eigenface space and compared to the training faces in terms of distances in the eigenface space. The test face is then classified as a certain person of the minimum distance or score if the minimum distance is smaller then a recognition bound. The recognition bound is estimated based on Bayesian rules to compromise between the probabilities of false acceptance and false rejection. False acceptance happens when the recognition bound for the class is too big that an incorrect person is included into the class, and false rejection stands for the converse case that a correct person is rejected. Other face recognition algorithms like independent component analysis also apply, and are reported to have better recognition performances [2, 23, 24].

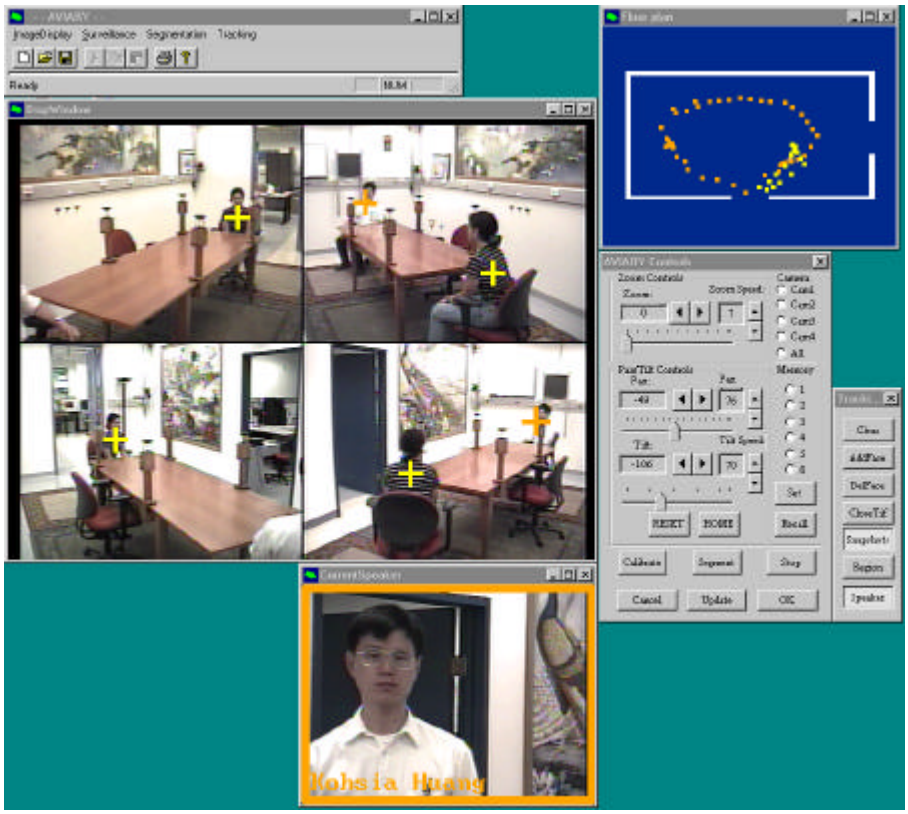

Figure 6 Integrated performance: Two people having a conversation in a room. Upper left window shows the views from four cameras and the crosshairs correspond to the 3D tracks projected back to the image planes. Upper right window shows the projections of the 3D tracks of the participants onto the floor plane. Bottom window is the current speaker recognized by their voice, and the snapshot of their faces is displayed with the identity from joint audio-video recognition.

Speaker Recognition module. Speaker recognition is an acoustic biometric modality that uses voice patterns to recognize the speaker. A text independent speaker identification module is used here. It takes the signal from microphone array and operates in parallel with the modules related to the rectilinear camera network. It has two modes of operation. For the first mode, when the camera network takes snapshot of a person, the speech sample is also taken from that person. The recognition results from face recognition and speaker recognition modules are then fused to yield the final recognition result. This fused result is then stored along with the tracking information of that person. For the second mode, when people in AVIARY room are speaking, the speaker recognition module detects speech and takes the speech sample to identify the speaker. The speaker recognition module we used here is VVDetective speaker recognition engine from the IBM ViaVoice SDK for Windows.

Joint Visual-Acoustic Person Recognition. Here the results of face and speaker recognition modules are fused together for robust person identification. Since ViaVoice does not provide access to confidence measures for recognition results, we are not able to make optimal decisions in Bayesian sense. Therefore, we perform the following procedure. Each module gives output only if there is reasonable confidence associated with it. If only one module outputs a valid result, then it is taken as the final decision. If both modules output results, and the two results are the same, then obviously such result is accepted. If the two results are different, the output from face recognition is accepted if confidence is above predetermined high value, otherwise the output from speaker recognition is accepted.

The integrated system of the above modules enables a teleconference type of operation. In the demonstrative experiment, two conduct a conversation in a room. When one person enters the room, the camera network takes a snapshot of the person, and the person speaks a sentence. The snapshot and speech sample are used to identify the person by the face recognition and speaker recognition modules, respectively. Then, the snapshot and the joint audio-video recognition result are stored with the track of that person. When the two people walk to their seats and sit down, the system identifies the

current speaker by the speech when the dialog is on going. The snapshot and identity of the current speaker are then displayed. Example images of the system operation are given in **Figure 6**. The system interface shows the video from the four cameras with projections of the 3D tracks to the image planes shown as colored crosshairs. Also, a floorplan of the room is shown with projections of tracks on the floor plane. If user clicks on one of the tracks in this view, the snapshot of the face/object is shown with the recognition result. Also, at the bottom of the screen in CurrentSpeaker window, the face and name of the current speaker are shown.

Concluding Remarks

Multimodal sensory intelligent environments benefits the users with more natural and efficient mechanisms of interactions, even when they are not sharing the same physical space. The overall system specification and general framework discussed in this paper convey many possible research challenges. The intelligent system developed in this paper with its four functional blocks of human tracking, active camera control, face recognition, and speaker recognition is an experimental investigation toward such intelligent environments. The developed system is demonstrated to be suitable for remote conferencing type of applications.

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

[11] M. Blattner, E. Glinert, "Multimodal Integration," *IEEE Multimedia*, vol. 3, No. 4, pp. 14-25, 1996.

[12] R. Bolt, "Put That There: Voice and Gesture at the Graphic Interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853-869, May 1998.

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

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

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

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

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

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

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

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

[21] M. Turk, A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Conf. Comput. Vis and Patt. Recog., Maui, HI, USA*, pp. 586-591, Jan. 1991.

[22] J. Yang, A. Waibel, "A Real-Time Face Tracker," *Proc. WACV'96, Sarasota, FL, USA*, 1996.

[23] P. Phillips et. al., "The FERET Database and Evaluation Procedure for Face-Recognition Algorithms," *Imag. Vis. Comput.*, vol. 16, pp. 295-306, 1998.

[24] J. Zhang, Y. Yan, M. Lades, "Face Recognition: Eigenface, Elastic Matching, and Neural Nets," *Proc. IEEE*, vol. 85, no. 9, pp. 1423-1435, Sep. 1997.