Machine Vision and Applications (2003) 14: 103–111
Digital Object Identifier (DOI) 10.1007/s00138-003-0106-5

Machine Vision and Applications

Video arrays for real-time tracking of person, head, and face in an intelligent room

Kohsia S. Huang, Mohan M. Trivedi

Computer Vision and Robotics Research (CVRR) Laboratory, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0407, USA; e-mail: {khuang,mtrivedi}@ucsd.edu

Abstract. Real-time three-dimensional tracking of people is an important requirement for a growing number of applications. In this paper we describe two trackers; both of them use a network of video cameras for person tracking. These trackers are called a rectilinear video array tracker (R-VAT) and an omnidirectional video array tracker (O-VAT), indicating the two different ways of video capture. The specific objectives of this paper are twofold: (i) to present a systematic comparison of these two trackers using an extensive series of experiments conducted in an 'intelligent' room; (ii) to develop a real-time system for tracking the head and face of a person, as an extension of the O-VAT approach. The comparative research indicates that O-VAT is more robust to the number of people, less complex and runs faster, needs manual camera calibration, and the integrated omnidirectional video network has better reconfigurability. The person head and face tracker study shows that such a system can serve as a most effective input stage for face recognition and facial expression analysis modules.

Keywords: Multi-camera systems – Real-time vision – Omnidirectional video arrays – 3D person and face tracking – Intelligent rooms

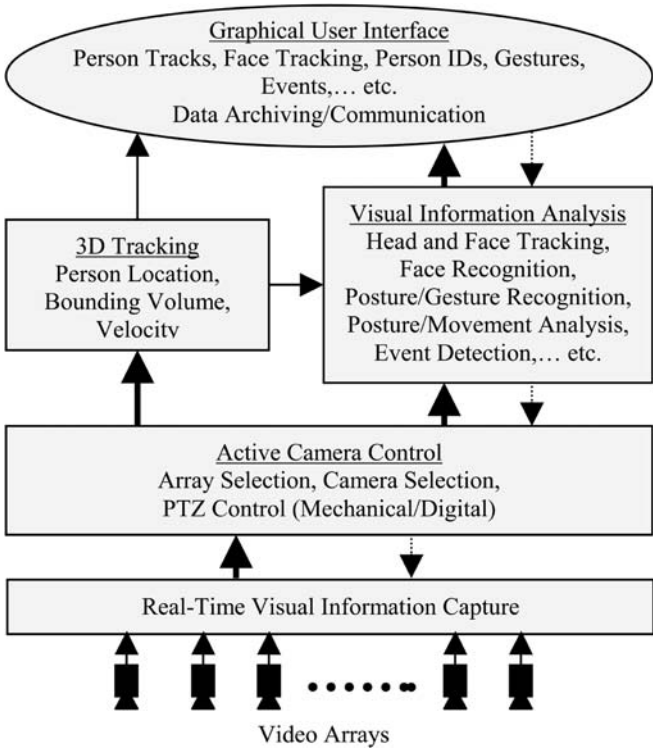

Fig. 1. General architecture and tasks of visual information capture and analysis system (VICAS) for intelligent-room systems

1 Introduction

Tracking of people using video sensors in real-time is an important requirement for a wide variety of applications. Researchers in the computer-vision community have recognized this importance and also the challenges associated with the development of accurate, reliable, robust, and practical algorithms and systems. In this paper, we describe two types of trackers for real-time tracking of people in indoor environments. These environments can be considered 'intelligent' environments, as our objective is to maintain an awareness of all the dynamic events and activities taking place in them.

The general structure and tasks of intelligent-room systems are illustrated in Fig. 1. The system acquires visual information both broadly by wide-coverage video arrays and specifically by zooming in on human details. The most important functionality in the system is three-dimensional (3D) tracking of people. Accurate and robust 3D tracking boosts system performance on generating close-up videos of human heads, faces and other body parts for face recognition [5, 7] and posture and movement analysis [8]. This motivates us to examine and evaluate a range of 3D tracking systems on video arrays. For active camera control, pan, tilt, and zoom (PTZ) can be controlled either mechanically or electronically. The mechanical approach typically provides higher-resolution video; however it requires better calibration. It also has slower performance and a limited number of simultaneous focuses of

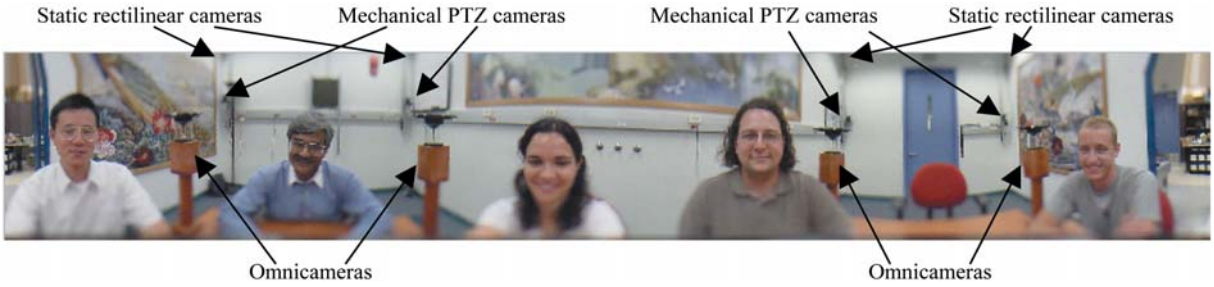

Fig. 2. The AVIARY intelligent-room test bed. Panorama unwrapped from an ODVS image shows the placement of multiple-camera networks. The ODVS network is mounted 1.2 m above ground on the corners of a $1.3\,\mathrm{m} \times 0.9\,\mathrm{m}$ meeting table around which people are sitting. Rectilinear arrays are mounted at the corners of the room. The dimensions of the room are $6.6\,\mathrm{m} \times 3.3\,\mathrm{m} \times 2.9\,\mathrm{m}$

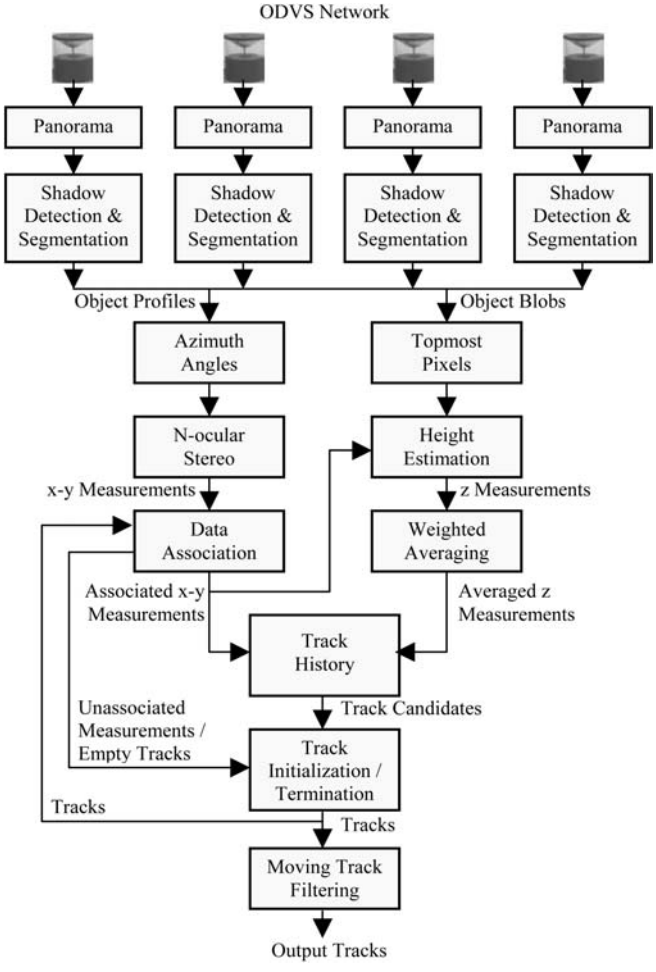

Fig. 3. O-VAT: omnidirectional video array tracker

attention. The fully electronic approach is free from these limitations, but with a tradeoff on resolution.

There are basically two somewhat opposite ways in which the indoor spaces can be visually captured, as demonstrated in Fig. 2.

1. *Outside-in-coverage*: can be realized by placing an array of multiple rectilinear cameras on the walls and ceiling.
2. *Inside-out-coverage*: can be realized by placing an array of cameras to capture wide-area panoramic images from non-obtrusive vantage points within the environment. An array of omnidirectional cameras seems to be most effective in this regard.

In this paper, we consider two camera network systems for the above competing approaches: the networked omnidirectional video array (NOVA) system [5], which utilizes an inside-out omnidirectional vision sensor (ODVS) [15] array, and the intelligent meeting room (IMR) system [7], which utilizes outside-in static and pan–tilt–zoom rectilinear camera arrays. It is important to compare their real-time 3D trackers, omnidirectional and rectilinear video array trackers (O-VAT and R-VAT) since 3D trackers directly influence the overall system performance [5, 7–9, 14, 15, 17]. It is not feasible to compare other works [11, 12, 16, 19] because they do not exploit camera networks or do not perform 3D tracking. The specific objectives of this paper are twofold: (i) to present a systematic comparison of these two trackers using an extensive series of experiments conducted in our intelligent room test bed; (ii) to develop a real-time system for tracking the head and face of a person, as an extension of the O-VAT approach. First, algorithms and real-time performances are compared between O-VAT and R-VAT. These experimental comparisons are quite unique, as we are able to compare and determine the strengths and limitations of the ODVS and rectilinear arrays on the same tasks simultaneously in exactly the same test bed. We then study the necessary perspective transform and active camera control schemes of the ODVS array for capturing and tracking faces of walking or sitting people. A series of experimental evaluations is presented to demonstrate the fully integrated ODVS person, head, and face tracking system.

2 O-VAT: omnidirectional video array tracker

O-VAT is the 3D tracker of the omnidirectional camera network system [5]. The main advantage of the omnidirectional vision sensor is the coverage [10, 12, 15]. It provides the maximum (360°) coverage using a single camera. The main disadvantage is low resolution. We propose utilization of an ODVS array to develop an effective 3D tracker of human movements and their faces. The O-VAT we have developed is an extension of the two-dimensional (2D) person tracker using the N-ocular algorithm [14, 15].

The 3D person tracker on four ODVSs is shown in Fig. 3. Each of the four ODVSs is calibrated in advance on the location and height of the ODVS optical center, the horizon on the ODVS panorama, azimuth direction, and internal parameters.

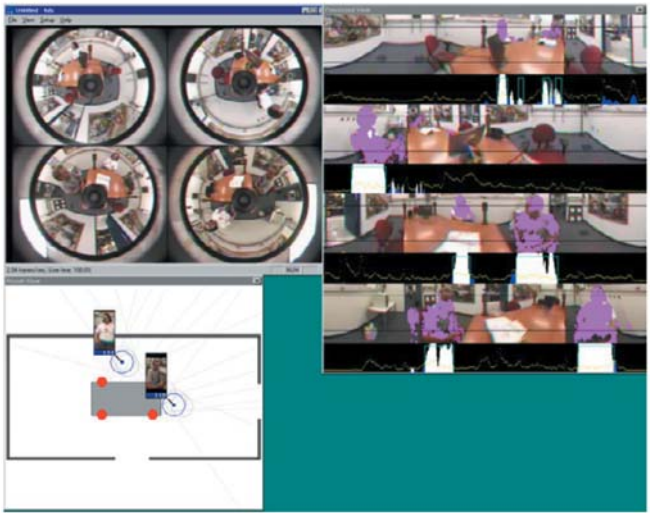

Fig. 4. The real-time 3D O-VAT. The *upper-left window* shows the four-source ODVS videos. The *upper-right window* shows the unwrapped panoramic videos with human-detection processing. The 1D profiles below the panoramas are for detecting the azimuth range of humans and the human blobs in the panoramas are for shadow detection. The *lower-left window* is the floor plan with the estimated human locations and heights in centimeters

The ODVSs are set upright. Location and height of the optical center are measured physically with respect to a preset origin in the room. To calibrate the horizon on the ODVS panorama (cf. Fig. 4), first the ODVS height is marked on the walls. Then the center of the ODVS image is trimmed so that the marks align into a row of the panorama. This is necessary for an accurate human-height estimation. The azimuth direction α of the ODVS is calculated by the relative location of a known object in the image with respect to the ODVS as

$$\alpha = \tan^{-1}\left(\frac{o_y - c_y}{o_x - o_x}\right) - \tan^{-1}\left(\frac{y_1 - y_0}{x_1 - x_0}\right), \qquad (1)$$

where (c_x, c_y) is the center of the ODVS image, (o_x, o_y) is the image coordinate of the object, (x_0, y_0) is the horizontal location of the mirror focus, and (x_1, y_1) is the horizontal location of the object. Multiple object points help to increase the accuracy of α, likewise the horizontal tracking accuracy. Internal parameters, i.e. the geometry of the hyperboloidal mirror, camera focal length, and CCD pixel geometry, are supplied by the manufacturer.

For human detection, each ODVS video is first unwrapped into a panoramic view. Segmentation is performed on the panoramas. As shown in Fig. 4, first a one-dimensional (1D) profile is formed by accumulating the pixel differences between the current frame and the pre-acquired background frame in each column of the panorama. Mean and variance of each background pixel are also acquired so that shadow detection [13] can be performed on the pixels of the current frame. Since each panoramic column corresponds to an azimuth angle, the azimuth range of a human can be detected from the 1D profile for each ODVS as in Fig. 4. Knowing the locations of the four ODVSs, the x–y horizontal location of the person can be determined by a sophisticated triangulation algorithm called N-ocular stereo [14, 15]. With increasing N, the number of ODVSs in the array, N-ocular is able to localize

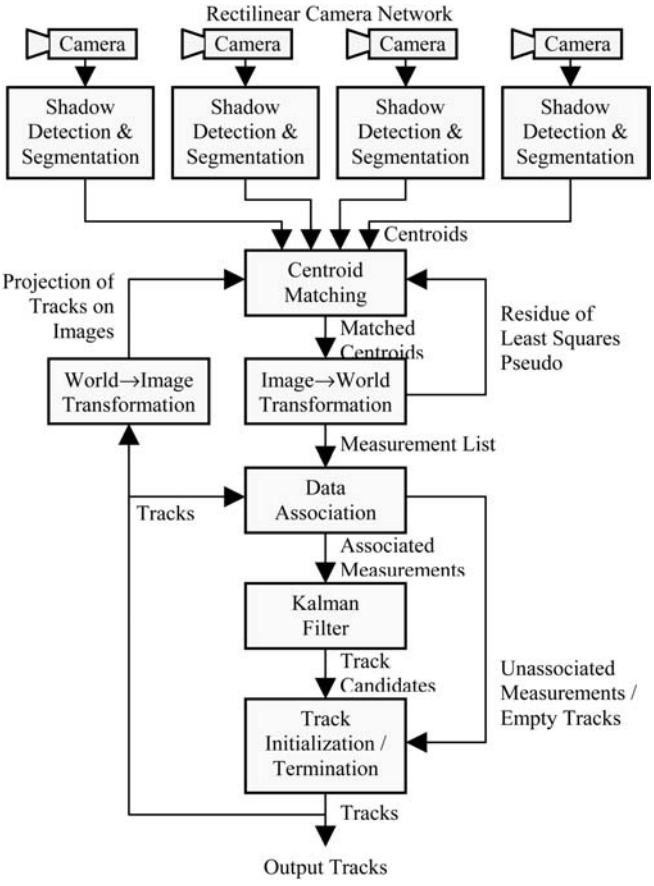

Fig. 5. Block diagram of R-VAT

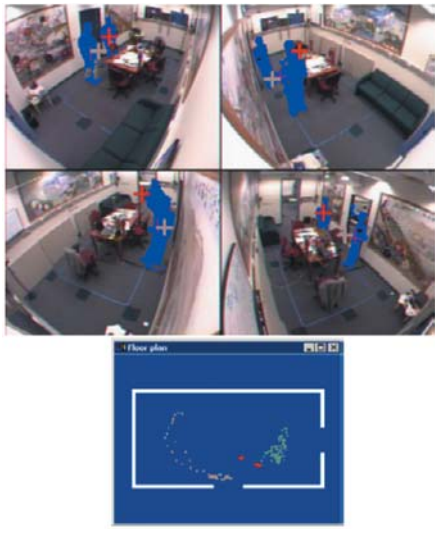

Fig. 6. The real-time 3D R-VAT. The *upper window* shows the source rectilinear videos of outside-in coverage. Note the occlusion situations on the human segments which drive away human centroids. The *lower floor plan* displays the color-coded human tracks

humans more precisely and reject more ghost locations. The measured x–y locations are then associated with the nearest human track registered by the O-VAT.

After the x–y measurement is available, the height z of the person can be estimated. First, the horizontal distance of the person to an ODVS is computed. Then, on the panorama, the

topmost pixel of the person's blob is detected. Thus the height of the person H_{person} can be estimated by similar triangles as

$$\frac{(Y_{\text{blob}} - Y_{\text{horizon}})\, H_{\text{pixel}}}{R_{\text{panorama}}} = \frac{H_{\text{person}} - H_{\text{upper focus}}}{d_{\text{person to ODVS}}}, \qquad (2)$$

where Y_{blob} is the topmost pixel of the person's blob, Y_{horizon} is the horizon on the panorama, H_{pixel} is the pixel height of the panorama, R_{panorama} is the radius of the cylindrical screen of the panorama, $H_{\text{upper focus}}$ is the physical height of the upper focus of the ODVS hyperboloidal mirror, and $d_{\text{person to ODVS}}$ is the estimated horizontal distance between the person and the ODVS. The final estimate of the person's height is a weighted sum of the estimates from the four ODVSs. The weight is inversely proportional to the distance between the person and the ODVS. Thus the x–y–z location is measured and associated with a registered human track.

On track registration, a new track is initialized if there exists an unassociated measurement. If no new measurements are associated with it for a period, the track is terminated. A human video of the track is displayed if the track has been registered for several hundred milliseconds, as shown in Fig. 4. The estimated human height is also displayed in centimeters. The 3D output track of the O-VAT is a moving average of the x–y–z measurements in the past 0.5 s.

3 R-VAT: rectilinear video array tracker

Rectilinear cameras are commonly used in 2D [3, 19] or 3D [1, 2, 4] human-tracking applications. However, few of them are based on wide-angle multiple rectilinear cameras for person-position tracking. In order to compare with the networked ODVS tracker, we choose a networked rectilinear camera tracker [7, 8, 17].

The block diagram of the networked rectilinear tracker is shown in Fig. 5. The four static CCD cameras are installed at the four upper corners of the room. Each camera covers the entire room, as shown in Fig. 6. The cameras are calibrated in advance by Tsai's algorithm [18] for the internal and external parameters. A person is segmented from the camera images by background subtraction. Shadow detection [13] is performed to segment humans accurately. A forgetting factor is also applied so that the background image is updated continuously. Centroids of the human blobs are then matched between the cameras with reference to the registered tracks, as shown in Fig. 6. For each human centroid, the following equation is derived from Tsai's algorithm:

$$\begin{aligned}
&(r_7 o_1 - r_1)\, x + (r_8 o_1 - r_2)\, y + (r_9 o_1 - r_3)\, z = T_x - T_z o_1 \\
&(r_7 o_2 - r_4)\, x + (r_8 o_2 - r_5)\, y + (r_9 o_2 - r_6)\, z = T_y - T_z o_2 \\
&\Leftrightarrow \boldsymbol{A}_{2\times 3}\boldsymbol{x}_{3\times 1} = \boldsymbol{b}_{2\times 1},
\end{aligned} \qquad (3)$$

where the r's and T's are available from Tsai's calibration algorithm and the o's are the image coordinates of the centroid. The 3D location $x = [\boldsymbol{x}\ y\ z]^T$ can be estimated by taking the pseudo-inverse of (3) as $x \approx (\mathbf{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{b}$. The measurement $\boldsymbol{x}$ is accepted and added to the measurement list if the $\|\boldsymbol{b} - \mathbf{A}\mathbf{x}\|_2$ residue is small enough. The measurement list $Z = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\}$ is assigned to the registered tracks by the data-association rule θ. The best measurement-track association rule is determined by maximum likelihood as

$$\arg\max_{\theta} p\,(Z|\theta) = \arg\max_{\theta} \prod_{i=1}^{M} p\,(\boldsymbol{x}_i|\theta), \qquad (4)$$

where $p(\boldsymbol{x}_i\|\theta)$ is a Gaussian density of the Mahalanobis distance from the measurement $\boldsymbol{x}_i$ to the predicted location of a track. Each track utilizes a Kalman filter for human tracking. The associated measurement is used to update the Kalman filter of the track. The Kalman filter predicts the next location of the track and the prediction is fed back to centroid matching (3) to enhance the matching process. On track registration, if a measurement in the measurement list has no track associated with it, a new track is started and validated after several frames. A track is terminated if no measurement is associated with it for several frames. Kalman-filter states of the valid tracks are the final output of the R-VAT.

4 Performance comparison of O-VAT and R-VAT

The ODVS array, static rectilinear video array, and PTZ rectilinear video array are installed in our $6.6\,\text{m} \times 3.3\,\text{m}$ AVIARY intelligent-room test bed, as shown in Fig. 2. The ODVS array is installed at the four corners of a $1.3\,\text{m} \times 0.9\,\text{m}$ meeting table standing in the middle of the room, providing an inside-out coverage for O-VAT. The static and PTZ rectilinear arrays are installed at the four corners of the room to provide an outside-in coverage. R-VAT uses static rectilinear array videos as input. O-VAT and R-VAT are running on separate computers of the same specification when comparing the performances.

The emphasis of experimental design is on comparing the accuracy of O-VAT and R-VAT. Other issues on real-time efficiency and system flexibility are also considered. Therefore, we define the following indices for the comparison:

- 3D tracking accuracy
- Speed and complexity
- Calibration efficiency
- System reconfigurability

A. 3D tracking accuracy

The 3D accuracy of tracking walking people by O-VAT is compared to that of R-VAT. A rectangular walking path is designated on the floor in our test-bed meeting room. The ODVS network is mounted 1.2 m above ground on a meeting table to perform tracking and also to be closer to faces of sitting people. However, heads of standing or walking adults cannot be covered by the ODVSs. If the ODVSs are mounted higher, face details of the meeting participants will be lost. Therefore we invited a group of children to walk on the designated path. However, adults were also tested to compare 2D tracking of O-VAT and R-VAT. We tested up to three children and four adults.

O-VAT and R-VAT run on two independent computers and log 3D tracking estimate data as people are walking. Later the tracking estimates are retrieved, analyzed, and plotted off-line for comparison.

Table 1. Summary of tracking accuracy comparisons of O-VAT and R-VAT in terms of track offsets and standard deviations. 2D tracking and height estimates are evaluated on single-person and multiple-person cases. The height estimate is not valid for O-VAT on adult cases

Tracking accuracy (cm)			O-VAT x–y (2D)		O-VAT Height z		R-VAT x–y (2D)		R-VAT Height z	
			$\Delta\mu$	σ	$\Delta\mu$	σ	$\Delta\mu$	σ	$\Delta\mu$	σ
Single person		Child	10	12	0	3	10	3	15	12
		Adult	12	12	–	–	10	5	30	12
Multiple people	2	Child	10	12	0	3	10	5	15	15
		Adult	20	15	–	–	10	5	35	15
	3	Child	10	15	0	3	10	10	20	25
		Adult	20	15	–	–	10	12	35	30
	4	Adult	20	15	–	–	15	20	35	40

Note 1: $\Delta\mu$ is the offset of the track from the designated path; σ is the standard deviation of the track

Note 2: in the 3-children case, O-VAT height estimates of the tallest child are excluded from the accuracy calculation because the top of the child's head in the ODVS images was chopped off occasionally

The accuracy is evaluated in terms of the number of tracking targets. We organize the experiments as 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. Horizontal (x–y) accuracy as well as height (z) accuracy are compared between O-VAT and R-VAT. The accuracy indices are *offsets* of the tracking estimates from the designated walking path and *standard deviations* of the tracking estimates. Note that for adult cases only 2D tracking is available on O-VAT. The child-case tracking results are compared in Fig. 7, and the offsets and standard deviations for all cases are listed in Table 1.

From the experimental results, the O-VAT 2D standard deviations and track offsets are almost constant after two people, and the magnitude of change of the offsets and deviations from one person to two people is quite small. The O-VAT height estimation is excellent and the height standard deviation is small and independent of the number of people. Therefore O-VAT is relatively robust to the number of people. For R-VAT, the 2D standard deviation increases rapidly after three people, especially for adult cases. The R-VAT height estimate is inaccurate and degrades rapidly with the number of people. This is due to the fact that, for outside-in rectilinear coverage, the chance of occlusion increases rapidly with the number of people, as shown in Fig. 6. This situation is less likely to happen on O-VAT because the inside-out ODVSs are standing upright and people walking around can be easily distinguished in the ODVS panoramas.

B. Speed and complexity

When the two trackers are tested on the same platform (dual Pentium III $\sim$ 866 MHz, 256 MB RAM, Windows NT 4.0) and one person is being tracked, R-VAT runs at about three frames per second and O-VAT runs at about five frames per second. Therefore R-VAT is approximately 1.7 times slower than O-VAT. This is because R-VAT needs additional computations

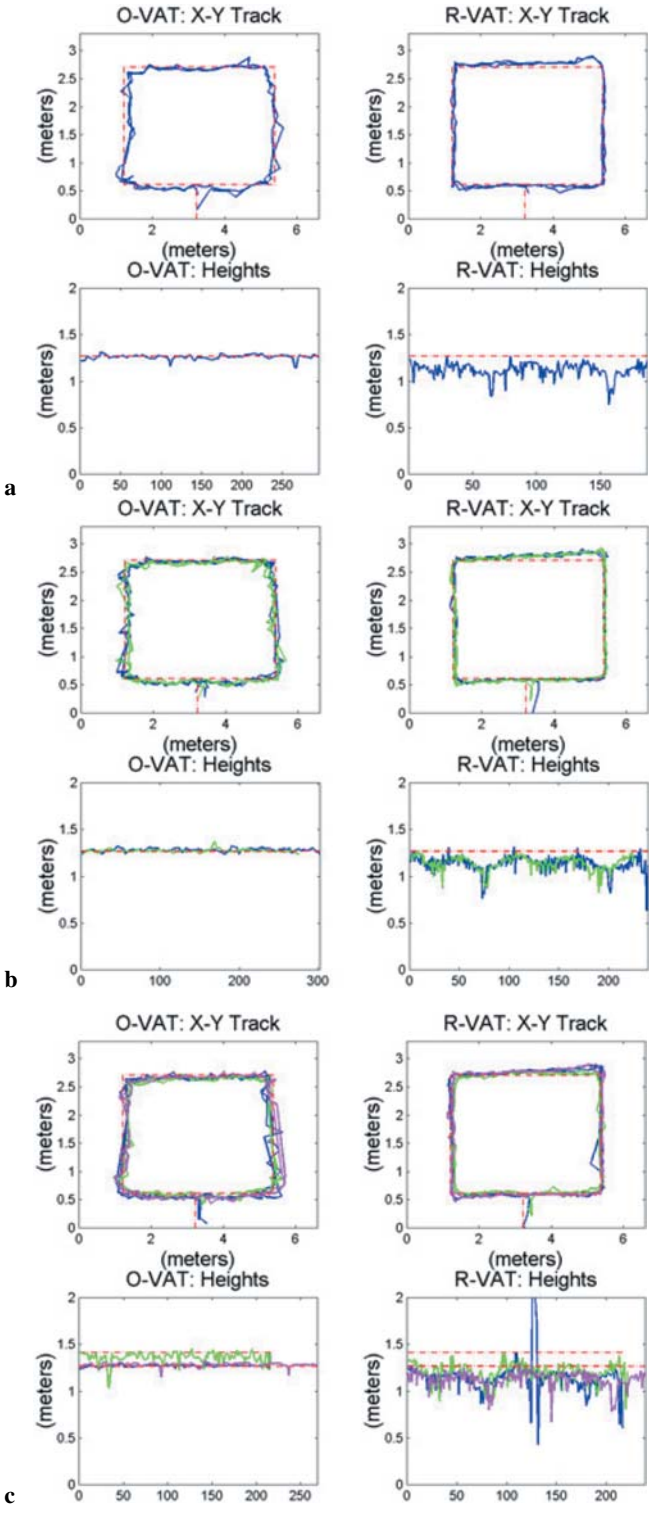

Fig. 7. Examples of accuracy comparison of O-VAT (*left columns*) and R-VAT (*right columns*): **a** Single-person tracking. **b** Simultaneous tracking of two people. **c** Simultaneous tracking of three people. **a**, **b**, and **c** respectively show tracking results with 1, 2, and 3 small children as volunteers. 2D x–y plots show the floor plans. *Red dashed lines* are the designated walking path on the floor. Different tracks of the volunteers are *color coded solid lines*. The heights are plotted against time. The actual heights of the volunteers are shown as *red dashed lines* and the height estimates of the children are plotted as *color coded solid lines*

for massive centroid matching, matrix pseudo-inverse, statistical data association, and Kalman filtering. On the other hand, O-VAT does not need centroid matching and matrix pseudo-inverse by its nature. Hence, if statistical data association and Kalman filtering are utilized, O-VAT should outperform R-VAT in speed and be at least as good in accuracy.

C. Calibration efficiency

Both O-VAT and R-VAT need careful calibration to yield accurate results. Tsai's calibration algorithm [18] is commonly applied to calibrate both external and internal parameters of rectilinear cameras. The process is automatic and gives accurate calibration. On the other hand, currently no such calibration procedure exists for hyperboloidal ODVSs. O-VAT calibration is performed manually as described in Sect. 2 for approximate parameters. However, it still gives satisfactory accuracy according to the previous accuracy comparison.

D. System reconfigurability

For any two video networks, we say one has higher reconfigurability than the other if it allows more functionalities. In this sense, the ODVS network has higher reconfigurability because, with one set of ODVS network, the system not only allows tracking but also allows electronic pan–tilt–zoom (PTZ) simultaneously. On the other hand, for a rectilinear camera network, if PTZ cameras are absent, the static rectilinear network can only perform tracking. Electronically zooming into a person's face is unsatisfactory because the face is too small and is obscured in the wide-angle static rectilinear cameras.

It should also be noted that, as compared to rectilinear PTZ cameras, the ODVS network does not require mechanical control to do pan–tilt–zooming. Electronic PTZ of the ODVS is performed by a perspective view [5]. Also, multiple perspective views looking at different objects can be generated from the same ODVS image at the same time. Thus the dynamic speed and system reconfigurability of the ODVS network are much higher. In addition, since the ODVS array is placed in the midst of the meeting participants, it has the advantageous inside-out coverage of people's faces from a close distance by unobtrusive electronic PTZ views. Therefore the ODVS network is very suitable for a meeting room setup.

In summary, the comparisons between the O-VAT and the R-VAT are listed in Table 2. This indicates that the ODVS network would be preferable for indoor intelligent environments if the O-VAT is improved by statistical data association and Kalman filtering for accurate and robust human tracking.

Table 2. Summary of the performance evaluations between the real-time 3D O-VAT and R-VAT

Comparison index	O-VAT	R-VAT
Accuracy	Better if > 3 people More robust to number of people	Better if ≤ 3 people Degrades rapidly with number of people
Speed/Complexity	Faster/Low	Slower/High
Calibration	Manual, approximate but satisfactory	Automatic, accurate
Reconfigurability	Excellent	Restricted

5 Integrated system: real-time head and face tracking

The integrated system, as illustrated in Fig. 1, needs to have two modes of operation:

1. 3D Tracking of people, and
2. Tracking of the head and face of the persons tracked in mode 1.

Results of mode 1 are used to select the 'best-view' camera to capture the inputs for the mode 2 operation. In this section, we present the details of accomplishing these two steps in an integrated manner using a network of four omnidirectional video sensors. Experimental evaluations of the integrated head and face tracking are presented in the next section.

The ODVS array can be extended to perform real-time head and face tracking as a dynamic system operation. The implementation of head and face tracking is to latch onto the face of a walking or sitting person by an electronic PTZ perspective view generated from a full-frame (640×480) ODVS video. Given the location of a person's head from the 3D O-VAT, the closest ODVS in the array is chosen to generate the perspective view by active camera selection (ACS). If the person moves, ACS switches to a suitable ODVS that faces the person according to the walking direction. The rectilinear camera network did not implement real-time head and face tracking because of the slower damping dynamics of mechanical PTZ cameras.

The perspective view is generated from the ODVS image by the ODVS geometry. The perspective view is a rectilinear screen whose viewing point is at the upper focus of the hyperboloidal mirror of the ODVS, as shown in Fig. 8. The lower focus of the mirror is at the optical center of the CCD lens. We explicitly derive this perspective transform from [12] where the detail is not given. The rectilinear screen can be specified by the pan angle θ, the tilt angle ϕ, and the effective focal length FL for zooming. The normal vector $\boldsymbol{n}$ and unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ of the rectilinear screen can be represented in terms of the 3D x–y–z coordinate system as

$$\mathbf{n} = \mathbf{R}\begin{bmatrix}1\\0\\0\end{bmatrix}, \quad \mathbf{u} = \mathbf{R}\begin{bmatrix}0\\-1\\0\end{bmatrix}, \quad \mathbf{v} = \mathbf{R}\begin{bmatrix}0\\0\\1\end{bmatrix}, \tag{5}$$

where $\mathbf{R}$ is the rotation matrix which rotates the x'–y'–z' coordinates to x–y–z coordinates in Fig. 8,

$$\boldsymbol{R} = \begin{bmatrix} \cos\theta\ \cos\phi & -\sin\theta & -\cos\theta\ \sin\phi \\ \sin\theta\ \cos\phi & \cos\theta & -\sin\theta\ \sin\phi \\ \sin\phi & 0 & \cos\phi \end{bmatrix}. \tag{6}$$

Thus a screen point P in u–v coordinates (u_p, v_p) can be related to the 3D x–y–z system by

$$\begin{bmatrix}x_p\\y_p\\z_p\end{bmatrix} = u_p\boldsymbol{u} + v_p\boldsymbol{v} + \mathrm{FL}\ \boldsymbol{n} = \boldsymbol{R}\begin{bmatrix}\mathrm{FL}\\-u_p\\v_p\end{bmatrix}. \tag{7}$$

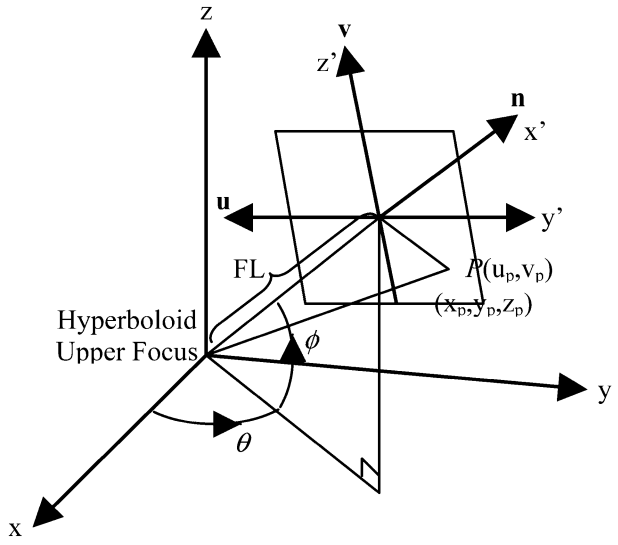

Fig. 8. ODVS perspective-view generation. The viewing point is the upper focus of the ODVS hyperboloidal mirror. A perspective screen in 3D space is specified by a pan angle θ, a tilt angle ϕ, and an effective focal length FL that determines the zooming factor. The associated CCD pixels of the points on the perspective screen can be computed. The generated perspective view is used for head and face tracking

Next the x–y–z coordinates of P can be used to find its associated pixel $(c_x - r_1 x_p/r_p, c_y - r_1 y_p/r_p)$ in the CCD plane of the ODVS, where (c_x, c_y) is the center pixel of the ODVS image, $r_p = \sqrt{x_p^2 + y_p^2}$, and

$$r_1 = \frac{f r_M}{z_M + 2c} = \frac{f r_M}{z_M + 2\sqrt{a^2 - b^2}}, \tag{8}$$

with

$$r_M = \frac{mc + a\sqrt{1 + m^2}}{(a^2/b^2 - m^2)}, \tag{9}$$

$$z_M = m r_M, \tag{10}$$

$$m = \frac{z_p}{r_p}, \tag{11}$$

where a, b are hyperboloidal shaping parameters and f is the focal length of the CCD lens. Equations (8) to (11) can be represented in polar coordinates as in [12]. Head and face tracking is thus carried out by calculating the θ, ϕ, and FL values on the relative 3D location of the human head and the chosen ODVS. It also allows users to manually specify the pan, tilt, and zoom factors to zoom into a human face.

6 Evaluation of head and face tracking system

In this section we present a series of experimental evaluations on the performance of the integrated ODVS array tracking system. The experiments will consider two possible scenarios for evaluation of the head and face tracking module: (i) people walking in the room and (ii) people sitting around a conference table.

Table 3. Performance of integrated head and face tracking

Category	Latch-on percentage
Walking people	$\geq 92\%$ if walking slower than 1.6 m/s Otherwise drop-off
Sitting people	100%

The experimental setup is the same as the experiments of the 3D O-VAT. The purpose of head and face tracking is to latch onto the face of a walking or sitting person by a perspective view (176×144) generated from a full-frame (640×480) ODVS video. Head and face tracking is regarded as successful if the human head or face is kept fully within the perspective view by the system. When the person is walking, the head or face of the person can be at least 92% tracked by the dynamically generated perspective view when the person is walking more slowly than approximately 1.6 m/s in our test bed. The instances when the system did not fully latch onto the person's head or face were when the active ODVS was being handed over to another one by hardware switching. The hardware-switching delay is about 0.5 s. If the person walks faster than 1.6 m/s, the tracking system would have a problem latching onto the head or face due to a delay between the moving-average tracking output and the instantaneous human position. When the person is sitting, the face is 100% latched on regardless of the facing angle. These two cases are summarized in Table 3. Examples of dynamic head and face tracking for walking people are shown in Fig. 9. Figure 10 shows head and face tracking for sitting people. Demonstration video clips of automatic person, head, and face tracking on the ODVS array are available at http://cvrr.ucsd.edu/pm-am/demos/index.html.

When a face is being tracked, the face can be identified using a robust streaming face recognition algorithm [6]. Instead of using single-frame images, it boosts the recognition rate to 99% by classifying sequences of face images by a Gaussian mixture model and a maximum-likelihood decision rule. The face video can also be analyzed for facial expressions. Thus the integrated system is more intelligent for applications like video conferencing and visual surveillance.

7 Conclusions

In this paper we compared two real-time 3D person trackers on an ODVS array and on a rectilinear camera array by experimental evaluations. O-VAT is more robust to the number of people in the room, runs faster, and the ODVS array has very good system reconfigurability, yet the calibration is done manually but satisfactorily. If O-VAT is equipped with statistical data association and Kalman filtering, it would be more advantageous than R-VAT in speed and accuracy. For the integrated operations, the ODVS array uses the 3D locations of a person from O-VAT to track the head and face dynamically in real-time. The face image can then be identified by a robust video-based face recognition algorithm or analyzed for facial expression. The experimental results presented in this paper suggest that the ODVS array is preferable for real-time human, head, and face tracking in intelligent-room applications.

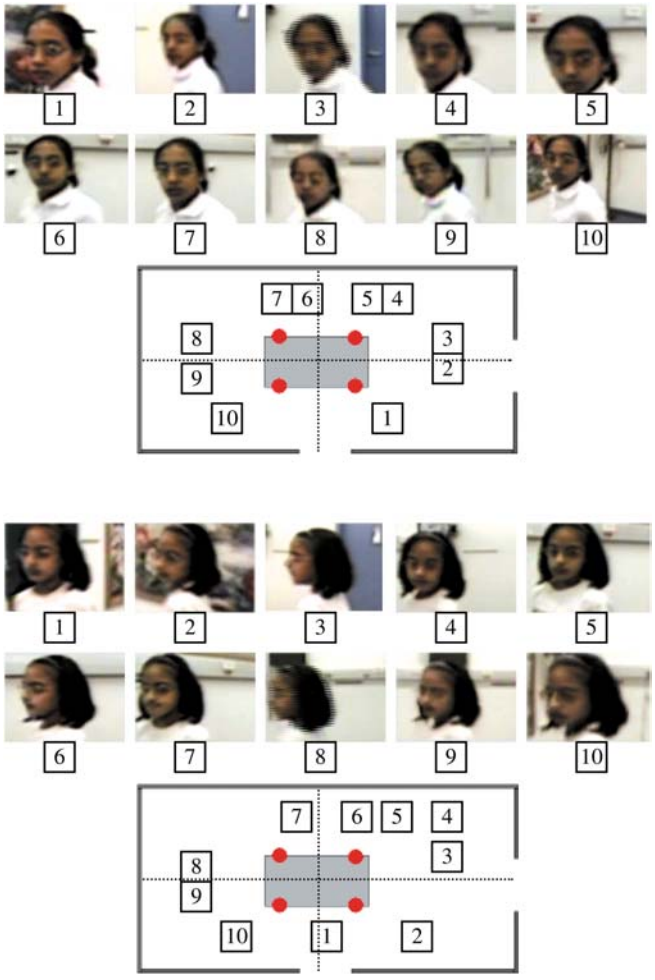

Fig. 9. Head and face tracking of walking people. The participants walked in circles around a meeting table where the ODVS array sits. Relative locations of the images are marked on the floor plans. *Dotted lines* on the floor plans partition the approximate duty regions of each ODVS. Images whose marks fall in one duty region were taken by the associated ODVS

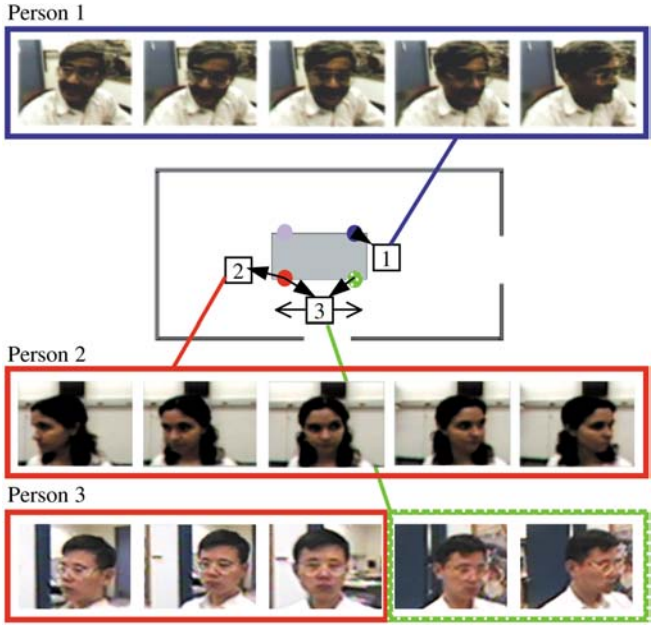

Fig. 10. Head and face tracking of sitting people. Heads and faces are always latched onto independent of the facing angle. Note that the four ODVSs are color coded. Pictures taken by a specific ODVS are bordered by the same color as the ODVS. The first person and the second person are respectively latched onto by two different ODVSs, while the third person is moving around and two ODVS are jointly latching onto that person

Acknowledgements. Our research is supported by the California Digital Media Innovation (DiMI) program. We are pleased to acknowledge the assistance of our colleagues, especially Ms. Ivana Mikic, during the course of this research. The authors are also grateful to the volunteers for their participation during experiments, especially the 4th graders Amruta, Aditi, and Kristina, for their enthusiastic involvement.

References

1. Bregler C, Malik J (1998) Tracking people with twists and exponential maps. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 8–15
2. Gavrilla D, Davis L (1996) 3D model-based tracking of humans in action: a multi-view approach. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 73–80
3. Haritaoglu I, Harwood D, Davis D (1998) W4: who? when? where? what? A real time system for detecting and tracking people. Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp 222–227
4. Horprasert T, Haritaoglu I, Harwood D, Davis L, Wren C, Pentland A (1998) Real-time 3D motion capture. Proceedings of Workshop on Perceptual User Interfaces, pp 87–90
5. Huang K, Trivedi M (2001) NOVA: networked omnivision arrays for intelligent environment. Proceedings of SPIE Conference on Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation IV, Vol 4479, pp 129–134
6. Huang K, Trivedi M (2002) Streaming face recognition for the omni video array intelligent room system. Proceedings of International Conference on Pattern Recognition, Vol 4, pp 213–216
7. Mikic I, Huang K, Trivedi MM (2000) Activity monitoring and summarization for an intelligent meeting room. Proceedings of IEEE Workshop on Human Motion, pp 107–117
8. Mikic I, Trivedi MM, Hunter E, Cosman P (2001) Articulated body posture estimation from multi-camera voxel data. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, Vol 1, pp 455–460
9. Mikic I, Santini S, Jain R (2000) Tracking objects in 3D using multiple camera views. Proceedings of ACCV 2000, Taipei, Taiwan, pp 234–239
10. Nayar S (1997) Catadioptric omnidirectional camera. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp 482–488
11. Boult T, Micheals R, Gao X, Eckmann M (2001) Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. Proceedings of the IEEE, Vol 89, No 10, pp 1382–1402, Oct. 2001
12. Onoe Y, Yokoya N, Yamazawa K, Takemura H (1998) Visual surveillance and monitoring system using an omnidirectional video camera. Proceedings of International Conference on Pattern Recognition, pp 588–592
13. Prati A, Mikic I, Grana C, Trivedi MM (2001) Shadow detection algorithms for traffic flow analysis: a comparative study.

Proceedings of IEEE Intelligent Transportation Systems Conference, pp 340–345
14. Sogo T, Ishiguro H, Trivedi MM (2001) N-ocular stereo for real-time human tracking. In: Benosman R, Kang SB (eds) Panoramic Vision, Springer, Berlin, pp 359–375
15. Sogo T, Ishiguro H, Trivedi MM (2000) Real-time target localization and tracking by N-ocular stereo. Proceedings of IEEE Workshop on Omnidirectional Vision, pp 153–160
16. Stiefelhagen R, Yang J, Waibel A (2000) Simultaneous tracking of head poses in a panoramic view. Proceedings of International Conference on Pattern Recognition, pp 722–725
17. Trivedi MM, Huang K, Mikic I (2000) Intelligent environments and active camera networks. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, pp 804–809
18. Tsai R (1987) A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. IEEE J Robot Autom RA-3(4): 323–344
19. Wren C, Azarbayejani A, Darell T, Pentland A (1997) Pfinder: Real-Time Tracking of the Human Body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7): 780–785

Kohsia Samuel Huang was born in Hsinchu, Taiwan, R.O.C. in 1966. He received his BS degree in Electrical Engineering from Chung-Yuan Christian University in 1988, an MS in Control Engineering from National Chiao-Tung University in 1991, and an MS in Electrical Engineering from the University of Southern California in 1995. From 1996 to 1999, he worked as a software engineer in industry. He is currently a PhD student in the Department of Electrical and Computer Engineering, University of California, San Diego. His research interests include computer vision, multimodal intelligent environments, and learning algorithms.

Mohan Manubhai Trivedi was born in Wardha, India on October 4, 1953. He received a BE (Honors) degree in Electronics from the Birla Institute of Technology and Science in Pilani, India in 1974 and MS and PhD degrees in Electrical Engineering from the Utah State University in 1976 and 1979, respectively. He is a Professor in the Electrical and Computer Engineering Department of the University of California, San Diego (UCSD) where he serves as the Director of the Computer Vision and Robotics Research Laboratory (http://cvrr.ucsd.edu). He and his team are engaged in a broad range of research studies in active perception and novel machine-vision systems, intelligent ('smart') environments, distributed video networks, and intelligent systems including intelligent highways and intelligent vehicles. At UCSD, he also serves on the Executive Committee of the California Institute for Telecommunication and Information Technologies, Cal-$(IT)^2$ (http://www.calit2.net/), 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 (http://www.dimi.ucsb.edu/) (DiMI) program. He serves as the Editor-in-Chief of the journal Machine Vision and Applications. 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. He is a Fellow of the International Society for Optical Engineering (SPIE). He is listed in Who's Who in the Frontiers of Science and Technology, Who's Who in American Education, American Men and Women of Science, Who's Who in the World, and other similar publications. He has published extensively (over 60 journal articles and over 200 conference papers) and has edited over a dozen volumes including books, special issues, video presentations, and conference proceedings. He serves regularly as a consultant to various national and international industry and government agencies.