Manuscript Number: 27

Omni and Rectilinear Video Arrays for Real-Time Person Tracking

***Abstract*--Real-time 3D person tracking is an important requirement for intelligent rooms. Trackers based on multiple networked cameras estimate human positions and heights more accurately and robustly than single camera or binocular trackers. The goal of this paper is to present a comparative evaluation of two types of 3D tracking systems built using omnidirectional- and rectilinear- video array trackers (VAT). The OD-VAT uses N-ocular algorithm to estimate planar location of a person. Height is then measured after the planar location. The networked omnicamera system then use the 3D tracks of person to perform head and face tracking. The R-VAT uses a calibrated rectilinear camera network to estimate human location and height, and keeps track of human path by a Kalman filter. Experiments on the two trackers are conducted to evaluate tracking accuracy in terms of mean offset and standard deviation as a function of the number of people. Performance is also compared on other indices including speed, computational complexity, robustness to environment, smoothness of head/face tracking, repeatability, calibration efficiency, and system reconfigurability. Possible fusion schemes are also suggested to combine the advantages of the two video networks.**

***Keywords*--Omnidirectional vision, video camera arrays, 3D person tracking, head and face tracking, intelligent environments, performance evaluation.**

I. Introduction

THIS research is performed as a part of a larger project associated with the design, development, and applications of intelligent environments. The intelligent environment considered is an indoor area, similar to a class room, meeting or conference room, a theater or a laboratory. An important requirement is to let the humans do their activities *naturally*. In other words, we do not require humans to adapt to the environments but would like the environment to adapt to the humans. This design guideline places some challenging requirements on the computer vision algorithms. Multiple video streams, from wide range of perspectives and resolutions, need to be accurately and efficiently analyzed. The processing and analysis algorithms need to cover a wide spectrum, from the low signal levels to higher semantic and context levels. Accurate, robust, and efficient means for tracking multiple people in these environments as well as tracking of the head and faces of people is a critical requirement. This motivates us to examine, develop, and evaluate a range of alternate multiple video based tracking algorithms.

There are basically two somewhat opposite ways in which the indoor spaces can be visually captured. These are:

(1) *Outside-in-coverage*: can be realized by placing an array of multiple rectilinear cameras on the walls and ceilings.

(2) *Inside-out-coverage*: can be realized by placing an array of cameras to capture a wide area panoramic images from some no-obtrusive vantage points within the environments. An array of omnidirectional cameras seems to be most in this regards.

Intelligent environments are sensor-based rooms or spaces which automatically derive and continuously maintain an awareness of the space and activities taking place in them [12]. To carry these capabilities, trackers play an important role in tracking human motions and activities from multiple sensory inputs. Trackers using networked multiple cameras have better accuracy and more capabilities than single or binocular camera systems since spatial redundancy between camera coverage is utilized [1] [2] [3] [4] [11] [14]. In this paper, we investigate real-time trackers of human 3D positions using multiple Omni-Directional Vision Sensors (ODVSs) [5] [11] and multiple rectilinear cameras [6] [7] [12]. Algorithms, experimental and empirical performances are compared between these two types of tracking systems.

II. OD-VAT: Omnidirectional Video Array Tracker

The main advantage of the omnidirectional vision sensors (ODVS) is the coverage [8,9,11]. It provides the maximum (360 degrees) coverage using a single camera. The main disadvantage is low resolution. We propose utilization of an array of multiple odvs to develop an effective 3-D tracker of human movements and faces. The OD-VAT we have developed can be considered an extension of the 2-D person tracking system presented in [11].

The 3D person tracker utilizing four ODVSs is shown in Figure 1. The four ODVSs are calibrated in advance on their locations, heights, azimuth directions, and internal parameters. Each ODVS video is unwrapped into a panoramic view. Background subtraction is performed on the panoramas. As shown in Figure 2, first an 1D profile of current panorama frame is formed by accumulating pixel differences to background in each column of the panorama. The background profile is formed by the maximum value for each column of the 1D profiles in 30 frames. Also mean and variance for each pixel are evaluated and the statistics of brightness distortion α and chrominance distortion CD for each pixel are evaluated to determine thresholds for shadow detection,

$$a = \frac{1}{\left(\frac{m_R}{s_R}\right)^2 + \left(\frac{m_G}{s_G}\right)^2 + \left(\frac{m_B}{s_B}\right)^2} \left(\frac{I_R m_R}{s_R^2} + \frac{I_G m_G}{s_G^2} + \frac{I_B m_B}{s_B^2} \right) \quad (1)$$

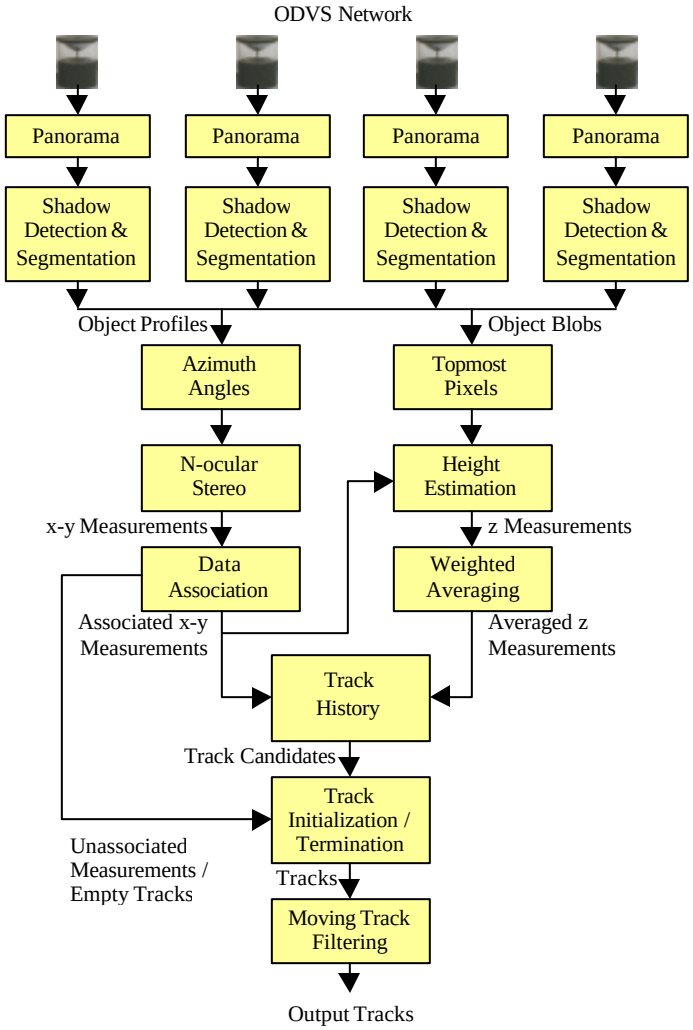


Figure 1: OD-VAT: Omnidirectional Video Array Tracker.

$$CD = \sqrt{\left(\frac{I_R - a\,m_R}{s_R}\right)^2 + \left(\frac{I_G - a\,m_G}{s_G}\right)^2 + \left(\frac{I_B - a\,m_B}{s_B}\right)^2} \qquad (2)$$

Foreground and shadow can be differentiated by the values of a and CD at each pixel [10]. Each panoramic column corresponds to an azimuth angle viewing from ODVS. Then if a person or an object presents, it will be detected from the background profile of each ODVS as the white histogram profile in Figure 2. Shadow detection and background subtraction are performed in the detected range to extract foreground objects as the marked silhouettes. The azimuth angle of the person relative to an ODVS is estimated as the center of the detected object. Knowing the locations of the four ODVSs, the x-y planar location of the person or object can be determined by triangulation method as illustrated in the lower-left window in Figure 2. This localization mechanism is called N-ocular and is detailed in [11]. Error compensation measures are applied to handle ghost situations and people near the baselines of two ODVSs. The measured x-y locations is then associated to the nearest human track.

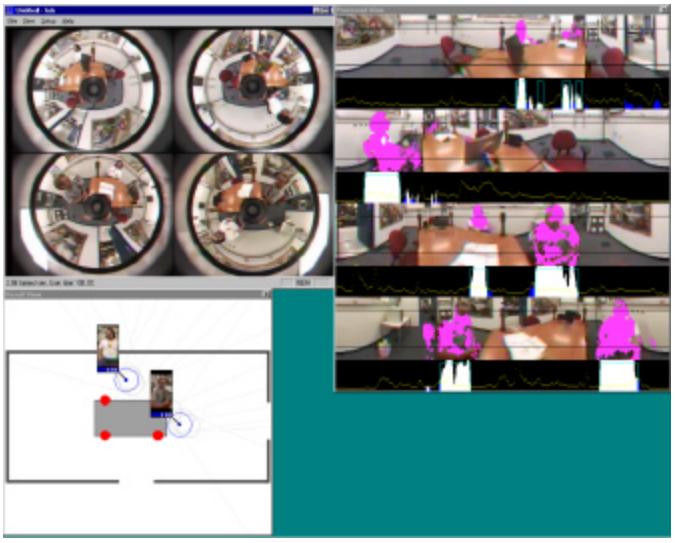


Figure 2: Real-time 3D human tracker based on four ODVS video streams.

After the instantaneous x-y measurement of person is available, height z of the person can be estimated. For each panorama, the topmost pixel of person's segment is detected. Since the planar location of the person is estimated previously, the distance of the person to the corresponding ODVS can be computed. Then height of person H_{person} can be estimated by similar triangle as

$$\frac{(y_{blob} - y_{horizon})H_{pixel}}{R_{panorama}} = \frac{H_{person} - H_{upper\ focus}}{d_{person\ to\ ODVS}} \qquad (3)$$

where y_{blob} is the topmost point of person's blob, $y_{horizon}$ is the horizon on the panorama, H_{pixel} is the pixel height of panorama, $R_{panorama}$ is the radius of cylindrical screen of the panorama, $H_{upper\ focus}$ is the physical height of the upper focus of ODVS hyperboloid mirror, and $d_{person\ to\ ODVS}$ is the estimated planar distance between the person and ODVS. The final estimate of person's height is a weighted sum of the estimates from the four ODVSs. The weight is the inverse of the distance between person and ODVS. Thus the x-y-z location is measured and associated to a human object.

The tracks are kept by the human objects of the tracker. A new track is initialized if there exists an unassociated measurement. If no new measurements are associated to it for a period, the track is terminated. The track is displayed if new measurements are added to it for several frames, as shown in Figure 2. The estimated human height is displayed in centimeters under the clipped human image in the result window. The final 3D output track is a moving average of the track positions in past 0.5 seconds.

Based on the 3D tracker, an ODVS network system [5] is designed upon the ODVS tracker as in Figure 3. The 3D ODVS tracker running on one computer detects and tracks people and send the tracks to another computer. Single ODVS signal can be captured in full frame on the second computer. Utilizing the tracking information, the Active Camera

Selection (ACS) and Dynamic View Generation (DVG) modules latch on a human head by a perspective view generated from the full-frame ODVS image. The face image can be tracked within the perspective view and then be identified. As shown in Figure 4, a perspective view is generated electronically from the full-frame ODVS video. It can be used to latch on person's head and face both manually or automatically by tracker. Demonstration clips of person and head and face tracking on ODVS network is available at http://cvrr.ucsd.edu/pm-am/demos/index.html.

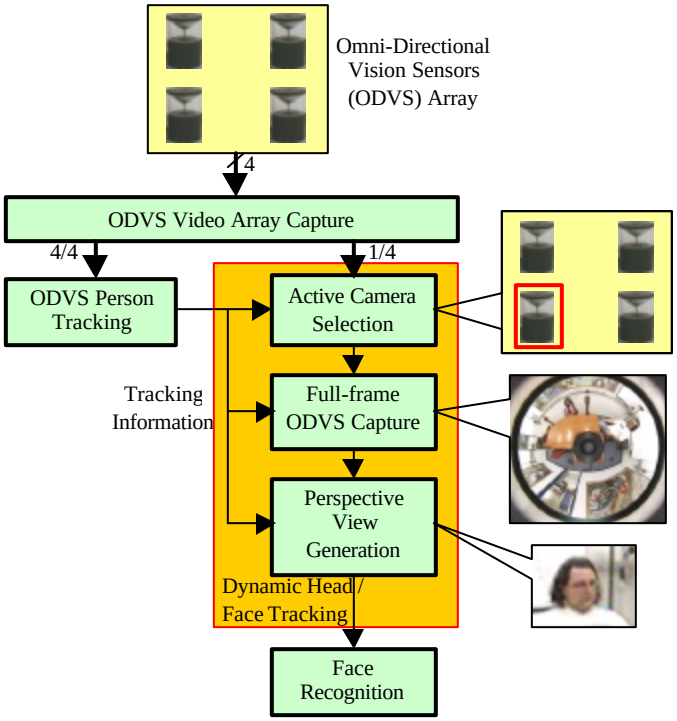


Figure 3: The ODVS network tracking system.

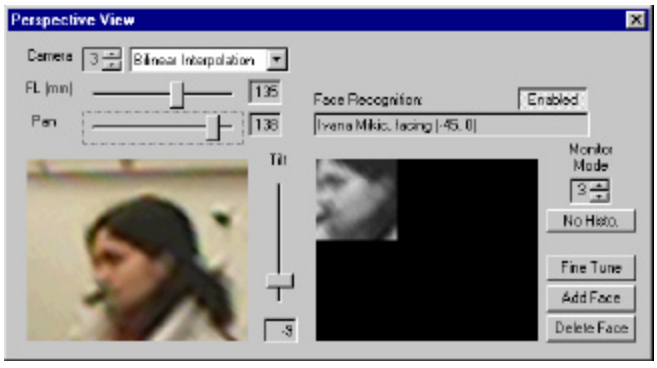


Figure 4: Perspective view generation for head and face tracking.

III. R-VAT: Rectilinear Video Array Tracker

Rectilinear cameras are commonly used in 2D [3] [14] 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 [6] [7] [12].

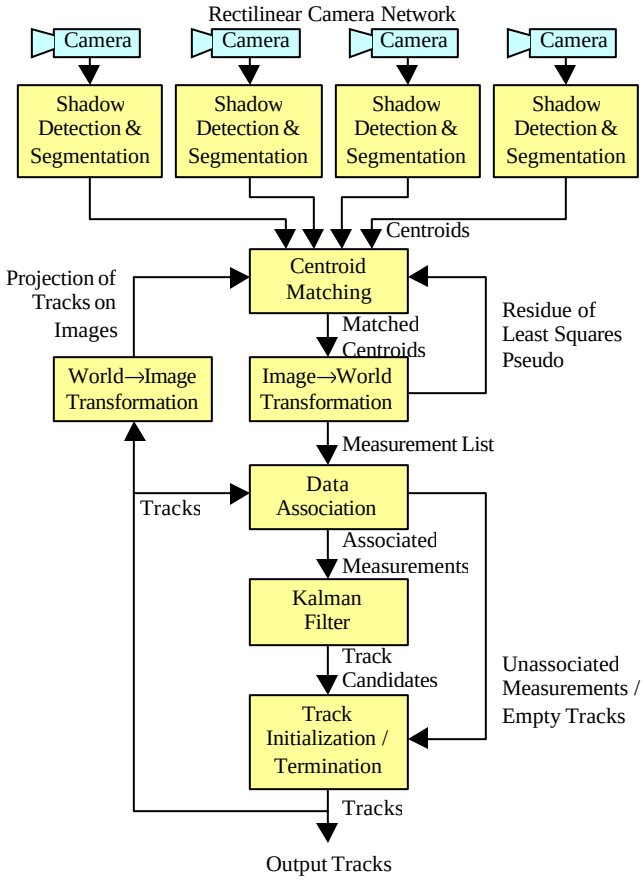


Figure 5: Block diagram of R-VAT.

The block diagram of the networked rectilinear tracker is shown in Figure 5. The four fixed focal length CCD cameras are installed at the four higher corners of the room. Each camera cover the entire view of the room. The cameras are calibrated in advance by Tsai's model [13] for internal and external parameters. Person is segmented from the camera images by background subtraction. Same shadow detection of equations (1) and (2) is performed to remove shadow interference. A forgetting factor is also applied on the background modeling so the segmentation can update its background. Centroids of the segments are then matched between the cameras with reference to the predicted current track. For each centroid, the following equation is generated,

$$\begin{cases} (r_7 o_1 - r_1)x + (r_8 o_1 - r_2)y + (r_9 o_1 - r_3)z = T_x - T_z o_1 \\ (r_7 o_2 - r_4)x + (r_8 o_2 - r_5)y + (r_9 o_2 - r_6)z = T_y - T_z o_2 \end{cases} \quad (4)$$
$$\Leftrightarrow \quad \mathbf{A}_{2\times 3}\mathbf{x}_{3\times 1} = \mathbf{b}_{2\times 1}$$

where the r's and T's are available from Tsai's calibration algorithm, o's are computed from the image coordinates of the centroid. The 3D location $\mathbf{x} = [x\ y\ z]^T$ can be estimated by putting together the equation pairs of the centroids and take pseudo inverse as $\mathbf{x} \approx (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$. A combination of the centroids is accepted and added to the measurement list if the residue of pseudo inverse, $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2$, is small enough. The measured 3D position is then associated to an existing track q by maximizing the likelihood $\Lambda(q)$ as

$$\arg\max_q \Lambda(q) = \arg\max_q p(Z \mid q) = \arg\max_q \prod_{i=1}^{M} p(z_i \mid q) \quad (5)$$

where Z is the measurement list, M is the number of measurements, and $p(z_i \mid q)$ is a Gaussian probability with respect to the Mahalanobis distance of the measurement z_i to the predicted location of track q. The assigned measurement is used to update a Kalman filter which models the track,

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{F}\mathbf{x}_k + \mathbf{v}_k \\ \mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{w}_k \end{cases} \quad (6)$$

where the state $\mathbf{x} = [x\ y\ z\ x'\ y'\ z']^T$ and output $\mathbf{z} = [x\ y\ z]^T$, $\mathbf{F}$ and $\mathbf{H}$ are derived using Newton's laws of motion, and $\mathbf{v}$ models random acceleration of human motion and $\mathbf{w}$ models measurement error. Kalman filter predicts the next location of the track and the prediction is fed back to centroid matching to accelerate the matching process. 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 networked rectilinear tracker.

IV. Performance Evaluation

In an intelligent room application, the accuracy of 3D tracker determines robustness of the overall system. In this section we are evaluating system performance in experiments on tracking people both walking and sitting, and tracking heads and faces of people. Besides quantitative accuracy, other performance indices include speed, computational complexity, robustness to environments, repeatability, calibration, and system reconfigurability.

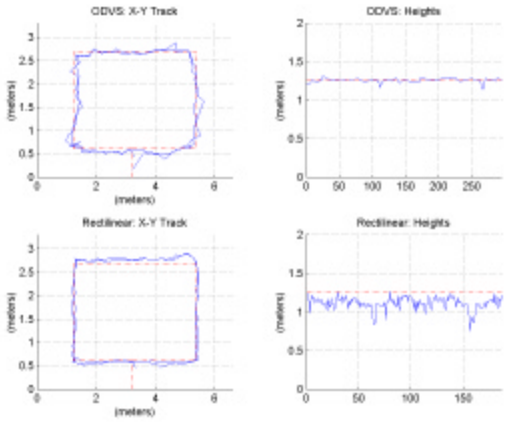
Both the networked ODVS tracking system and the networked rectilinear tracking system are embedded in our 6.7m-by-3.3m meeting room testbed called AVIARY. The four camera ODVS network is installed on the four corners of a rectangular meeting table sitting in the middle. The four rectilinear cameras are equipped with wide angle lenses and are installed at the four higher corners of the room so that each rectilinear camera covers the entire room.

A. Accuracy

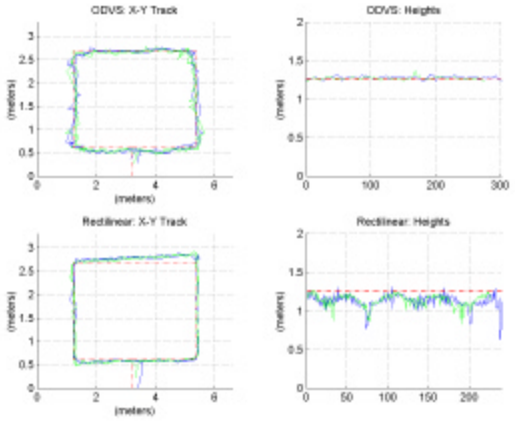
The accuracy of the 3D ODVS tracker is evaluated on tracking walking people and is compared to 3D rectilinear tracker [6] [7] [12]. A rectangular walking path is defined in our testbed around the meeting table. The ODVS network is mounted 1.2 meters above ground to cover sitting people but cannot cover standing adults. Therefore to evaluate the trackers we invited a group of children to walk over the test pattern in the laboratory (Figure 6.).

Figure 6: Volunteers walking over the test tracks during a two-person tracking experiment.

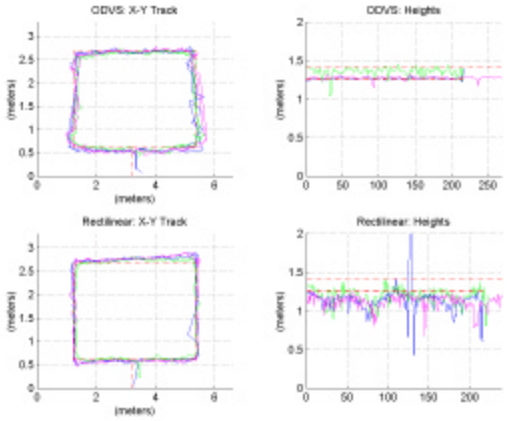
The experiments are conducted to evaluate 2D x-y tracking for single person and multiple people. Both children and adults are evaluated. For 3D x-y and z tracking, single child and multiple children are evaluated. Adults cannot be compared for 3D tracking because height is not available for ODVS tracker. The results are compared in Figure 7. the 3D tracks are plotted in x-y planar tracks and heights for both ODVS tracker and rectilinear tracker. The mean offset and standard deviation of the two trackers on different number of people are listed in Table 1.



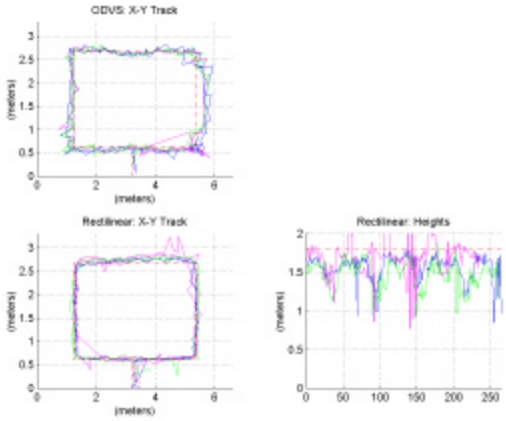
(a) One Child Experiment



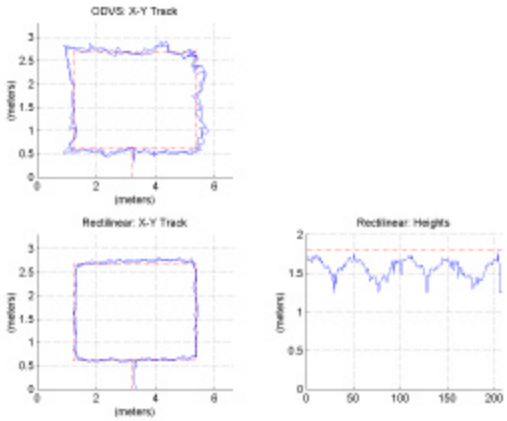
(b) Two Children Experiment



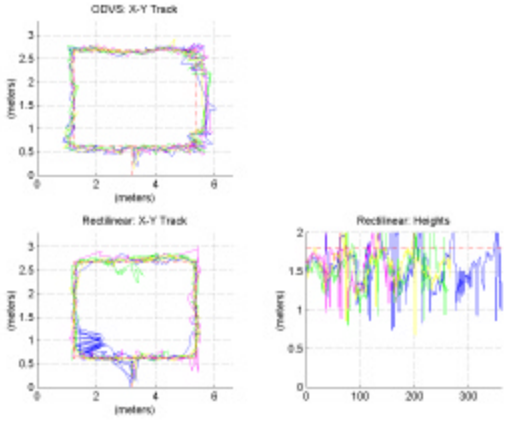
(c) Three Children Experiment



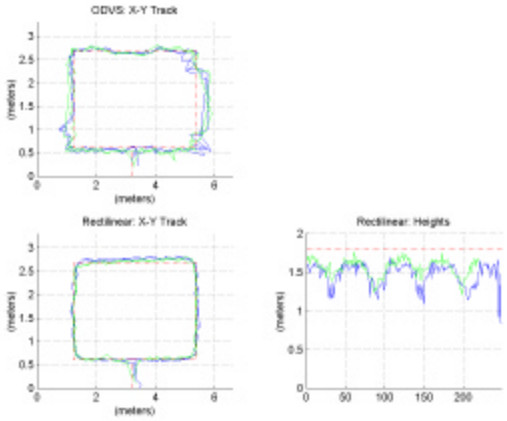
(f) Three Adults Experiment



(d) One Adult Experiment



(g) Four Adults Experiment



(e) Two Adults Experiment

Figure 7: Results of the comparative evaluation of the OD-VAT (upper row) and R-VAT (lower row) trackers. Part (a) (b) and (c) show results with small children as volunteers. This allows for proper assessment of the OD-VAT height estimator. Parts (d), (e), (f) and (g) shows results of experiments with adult volunteers. In this situation only the R-VAT height estimates are valid.

From the experimental results, the ODVS xy planar standard deviation does not drop with the number of people after 2, and so are the maximum mean offsets. The height standard deviation is independent of the number of people. Height estimation of the ODVS tracker is excellent because it is unbiased and deviation is very low. Note that in Figure 7(c) the height of the taller child has larger deviation because the child is a little higher than the ODVS coverage and at some places the top was chopped off. For the rectilinear tracker, the planar accuracy is good if there are less than 3 people. After 3 people, the planar accuracy drops rapidly, especially for adults. The height estimation is much less accurate than ODVS tracker since both mean offsets and standard deviations

	OD-VAT				R-VAT			
	x-y Plane		Height z		x-y Plane		Height z	
	$\Delta\mu_{max}$	σ	$\Delta\mu_{max}$	σ	$\Delta\mu_{max}$	σ	$\Delta\mu_{max}$	σ
1 Child	10	10	0	3	10	3	15	12
2 Children	10	15	0	3	10	5	15	15
3 Children	10	15	0	3	10	10	20	20
1 Adult	15	12	N/A	N/A	10	3	30	12
2 Adults	20	15	N/A	N/A	10	5	35	15
3 Adults	20	15	N/A	N/A	10	12	35	30
4 Adults	20	15	N/A	N/A	15	20	35	40

Note: $\Delta\mu_{max}$ is the maximum mean offset; σ is the standard deviation

Table 1: Mean offsets and standard deviations (in centimeters) estimates in a comparative evaluation of the OD-VAT and R-VAT trackers.

are very large. Also it degrades rapidly with the increase of people. It is due to the fact that as the number of people increases, the chance that one person occlude another will increase rapidly in a small room. This situation is less likely to happen on ODVS tracker because the ODVSs are standing straight and looking around. People walking in circles can be distinguished easily by ODVSs.

In summary, height estimation of ODVS tracker outperforms rectilinear tracker. The rectilinear tracker performs better when there are less people in the room. However, rectilinear tracker accuracy degrades rapidly with the increasing number of people. After 3 people, the ODVS tracker performs better than the rectilinear tracker. The ODVS performance could be further improved if Kalman filter is included like the rectilinear tracker.

B. Speed

When the two trackers are tested on the same platform (dual Pentium ~866MHz, 256MB RAM) and one person is being tracked, the rectilinear tracker runs at about 3 frames per second and the ODVS tracker runs at about 4 frames per second. Therefore the rectilinear tracker is approximately 1.3 times slower than the ODVS tracker.

C. Computational Complexity

From Figure 1 and Figure 5, the rectilinear tracker is equipped with shadow detection, maximum likelihood data association, Kalman filter track modeling, and background updating. The ODVS tracker is equipped with shadow detection, nearest neighborhood data association, and 0.5 sec track moving average. The rectilinear tracker is more complicated computationally and is verified by the execution speed.

D. Robustness to Environmental Changes

For ODVS tracker, background is updated manually by user command. The rectilinear tracker can update the background automatically with a forgetting factor. Thus although more complex, the rectilinear tracker is more robust to environmental changes. However, the drawback of the forgetting factor is that a person will be absorbed into background if he stays for too long.

E. Repeatability

The two trackers are embedded in the same room and the illumination is constant over time. We have operated the two systems several hundred times and overnights. The performance of the two trackers is very consistent over these experiments.

F. Calibration

Both the ODVS and rectilinear network trackers need careful calibration to yield accurate results. Tsai's calibration algorithm [13] is commonly applied to calibrate both external and internal parameters of rectilinear cameras. On the other hand, currently there is no such an efficient calibration procedure exist for hyperboloid mirror ODVSs. External parameters for ODVS network are measured manually. The accuracy could be within one centimeter. Internal parameters are supplied by the manufacturer and the accuracy depends on the manufacturing.

G. System Reconfigurability

System reconfigurability can be defined as: Upon the same type of infrastructure, a system has higher reconfigurability if it allows more functionality than the other system. In this manner, ODVS network has higher reconfigurability because with the same set of ODVS network, the system not only allows tracking but also allows electronic pan-tilt-zoom simultaneously. On the other hand, the rectilinear system, without adding pan-tilt-zoom cameras, can only do tracking on the static camera network. Zooming into person's face is unsatisfactory because face is too small and obscure in the wide-angle static rectilinear cameras.

It should also be noted that as compared to rectilinear pan-tilt-zoom cameras, the ODVS network does not require mechanical motion of the cameras at all. Thus the speed and system reconfigurability are increased. In addition, since the ODVSs are placed in the midst of the meeting participants, omnicameras have the advantageous viewing angles of the faces from a closer distance. Unobtrusive electronic pan-tilt-zoom can generate perspective views of people's faces. Therefore ODVS network is very suitable for meeting room setup.

In summary, the comparisons between the ODVS tracker and the rectilinear tracker are listed in Table 2.

Tracker Parameter	OD-VAT	R-VAT
Accuracy	Better if > 3 people	Better if ≤ 3 people
Speed	Faster	Slower
Complexity	Lower	Higher
Robustness	Static	Adaptive
Repeatability	High	High
Calibration	Manual	Automatic
Reconfigurability	Excellent	Restricted

Table 2: Summarization of the performance evaluations between 3D ODVS tracker and rectilinear tracker.

V. CONCLUDING REMARKS

In this paper we presented the ODVS network and rectilinear camera network systems for intelligent rooms. The ODVS and rectilinear network systems are compared experimentally. The ODVS tracker has very good height estimation, better speed, less complexity than the rectilinear tracker, and the ODVS network has very good system reconfigurability. The rectilinear tracker has better calibration efficiency and better x-y planar tracking accuracy when there are less people, but it needs another set of pan-tilt-zoom cameras to achieve head and face tracking as the ODVS network system.

In the future we can fuse the two system operations so that the advantages of the two networks can be combined. 3D measurements from two video networks can be shared between the two trackers. Statistical data association and track modeling can also be applied to the ODVS tracker. For head and face tracking the rectilinear tracker can drive electronic pan-tilt-zoom of the ODVSs, and ODVS tracker can drive rectilinear pan-tilt-zoom cameras. Other fusion schemes are also possible to optimize the two video networks.

ACKNOWLEDGMENTS

Our research is supported by the California Digital Media Innovation Program (DiMI) program. Main sponsors of the research are Sony Electronics, Compaq Computers, DaimlerChrysler, and Caltrans. 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, which took place over a period of five weeks.

REFERENCES

[1] C. Bregler and J. Malik, "Tracking People with Twists and Exponential Maps," *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, pp. 8-15, 1998.

[2] D. Gavrilla and L. Davis, "3-D Model-based Tracking of Humans in Action: A Multi-view Approach," *Proc. IEEE Conf. on CVPR*, p. 73-80, 1996.

[3] I. Haritaoglu, D. Harwood, and L. Davis, "W^4: Who? When? Where? What? A Real Time System for Detecting and Tracking People," *Proc. IEEE Int'l. Conf. on Automatic Face and Gesture Recognition*, pp. 222-227, Apr. 1998.

[4] T. Horprasert, I. Haritaoglu, D. Harwood, L. Davis, C. Wren, and A. Pentland, "Real-Time 3D Motion Capture," *Proc. PUI*, Nov. 1998.

[5] K. Huang and M. M. Trivedi, "NOVA: Networked Omnivision Arrays for Intelligent Environment," *Proc. SPIE Conf. on Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation IV*, Vol. 4479, Jul.-Aug. 2001.

[6] I. Mikic, K. Huang, and M. M. Trivedi, "Activity Monitoring and Summarization for an Intelligent Meeting Room," *Proc. IEEE Workshop on Human Motion*, Dec. 2000.

[7] I. Mikic, S. Santini, and R. Jain, "Tracking Objects in 3D using Multiple Camera Views," *Proc. ACCV 2000*, Taipei, Taiwan, Jan. 2000.

[8] S. Nayar, "Catadioptric Omnidirectional Camera," *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, pp. 482-488, Jun. 1997.

[9] Y. Onoe, N. Yokoya, K. Yamazawa, and H. Takemura, "Visual Surveillance and Monitoring System Using an Omnidirectional Video Camera," *Proc. IEEE 14th Int'l. Conf. on Patt. Recog.*, pp. 588-592, Aug. 1998.

[10] A. Prati, I. Mikic, C. Grana, and M. Trivedi, "Shadow Detection Algorithms for Traffic Flow Analysis: A Comparative Study," *Proc. IEEE Intelligent Transportation Systems Conf.*, Oakland, CA, Aug. 2001.

[11] T. Sogo, H. Ishiguro, and M. M. Trivedi, "Real-Time Target Localization and Tracking by N-Ocular Stereo," *Proc. IEEE Workshop on Omnidirectional Vision*, pp. 153-160, Jun. 2000.

[12] M. M. Trivedi, K. Huang, and I. Mikic, "Intelligent Environments and Active Camera Networks, " *Proc. IEEE Int'l. Conf. SMC.*, pp. 804-809, Oct. 2000.

[13] R. Tsai, "A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses," *IEEE J. of Robotics and Automation*, Vol. RA-3, No. 4, pp. 323-344, Aug. 1987.

[14] C. Wren, A. Azarbayejani, T. Darell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 780-785, Jul. 1997.