Image and Vision Computing Journal, June 2002.

An Integrated Surveillance System—Human Tracking and View Synthesis using Multiple Omni-Directional Vision Sensors

Kim C. Ng[♠], Hiroshi Ishiguro[♦], Mohan Trivedi[♥], and Takushi Sogo[♣]

♠ AST–La Jolla Lab, STMicroelectronics Inc., San Diego, CA, USA
♦ Department of Computer and Communication Sciences, Wakayama University, Japan
♥ Computer Vision and Robotics Research Laboratory, ECE Department, University of California–San Diego, USA
♣ Department of Social Informatics, Kyoto University, Japan

Abstract

Accurate and efficient monitoring of dynamically changing environments is one of the most important requirements for visual surveillance systems. This paper describes the development of an integrated system for this monitoring purpose. The system consists of multiple omni-directional vision sensors and was developed to address two specific surveillance tasks: (1) Robust tracking and profiling of human activities, (2) Dynamic synthesis of virtual views for observing the environment from arbitrary vantage points.

Keywords: human tracking, multiple camera stereo, N-ocular stereo, human activity profiling, dynamic view synthesis, visual modeling, surveillance and monitoring, and intelligent environments.

1 Introduction

Networks of cameras are required to provide wide scene coverage for many surveillance tasks. In designing such networked camera systems, it is important to consider practical aspects such as cost, complexity and robustness. This paper discusses the associated issues and describes development of a multiple camera system, which allows remote observers to view dynamic environments from arbitrary vantage points efficiently and realistically.

Requirements for such a system were discussed in [i]. The multiple camera system utilizes redundant visual information for robust monitoring tasks in large scenes. Several vision sensors observe a common area and provide redundant information. This redundant observation contains rich information for robust vision functions. This concept follows the one of distributed vision proposed by Ishiguro [ii] with more generality and practicality. Key attributes of the system are summarized as follows:

- The system covers a large area to observe dynamic events happening in the environment.
- The system tracks dynamic events in real time.
- The system synthesizes views for visualization at arbitrary viewpoints.
- The system enables us to develop an integrated information framework that can access both the real world and the virtual world through the computer network.
- The system allows inquiries of activities analysis in the scene.

Based on the above specifications, we have developed a vision system using multiple Omni-Directional Vision Sensors (ODVS) [iii]. In our development of the integrated surveillance system, much importance is placed on ensuring *seamless* coverage of wide observation areas. Also of importance is the effective integration of the network of sensors with the network of distributed computers utilizing the most appropriate communication links. In the future, we believe, such networks of distributed cameras/computers will be part of the infrastructure, making the integrated surveillance system quite practical.

Our system differs from other multiple camera approaches [iv][v],mainly in the configuration and utilization of many cameras. These alternate approaches used rectilinear images for stereo matching and their cameras were arranged densely with a short baseline. The scene they were looking at was restrained to a small area; therefore, their methods are not readily applied to our problem for generating views, which is especially important in the coverage of a large area. Further, our images are omni-directional and have their own unique properties. New algorithms had to be developed or modified to accomplish our tasks.

There are other types of multiple camera systems from the computer graphics community. These systems deal with reconstructing visual information by approximating *plenoptic* functions [vi]. *Light-field rendering* [vii] and *Lumigraph* [viii] are the two most prominent methods in approximation. These methods are generally simpler than the methods derived from the computer vision community. They do not require matching between images, and a large number of cameras placed in close proximity are needed to acquire many image samples.

All of these multiple camera systems aim to develop "practical" systems for generating views. However, they either require 3-D models or plenoptic functions. Their methods also do not show the possibility and scalability for utilization in a large area, where dynamic objects prevail.

Recently, systems monitoring wide areas with practical purposes have been proposed especially for visual surveillance [ix][x][xi][xii][xiii]. Some systems have proceeded further to analyze the tracked objects' behaviors [xiv][xv]. The work reported in this paper has a similar purpose for visual surveillance, but our approach is slightly different. Works that use multiple cameras to cover wide areas seldom take advantage of data redundancy from multiple cameras as we do. The large overlapping viewing area from multiple cameras provide the data redundancy.

Note that this paper aims at introducing a new promising multimedia system. Two of the most fundamental system components—tracking and view synthesis, are presented. In the following sections, we first introduce a real-time human tracking system using the ODVS. Tracking is described as one of the functions for surveillance. Further, we show that the system has the capability of synthesizing novel virtual views at arbitrary viewpoints.

2 Real-time human activity tracker

As an essential function of the integrated surveillance system, a real-time human tracker is developed with the use of multiple Omni-Directional Vision Sensors (ODVS). The low-cost and compact ODVS (shown in **Figure 1**) were originally designed and fabricated in house [iii]. These ODVS, used as a cluster of cameras, are placed at a fixed height of approximately 1m, although the configuration can be arbitrary in a global sense with different clusters of ODVS set at different heights. The tracker detects people and measures azimuth angles to localize people by triangulation as illustrated in **Figure 2**(a). The triangulation method is commonly known as "stereo," or more precisely omni-directional stereo in this case. To track people with stereo, the following problems should be considered:

- Correspondence problem among multiple targets
- Measurement precision of target locations
- Treatment of deformable human bodies

These problems occur in both conventional multiple baseline stereo [xiv][xvi][xvii] and omni-directional stereo. However, it is more difficult to verify the correspondence of targets accurately in omni-directional stereo. The main reason is due to the fact that the baselines among wide field-of-view ODVS are configured much more arbitrarily in our work than that of a conventional stereo. The measurement precision of a target location can become very low when the target locates along the baseline of two ODVS [xviii]. For example, the measured location of target B in **Figure 2**(a) is unstable because the target is on the baseline of sensors 1 and 2.

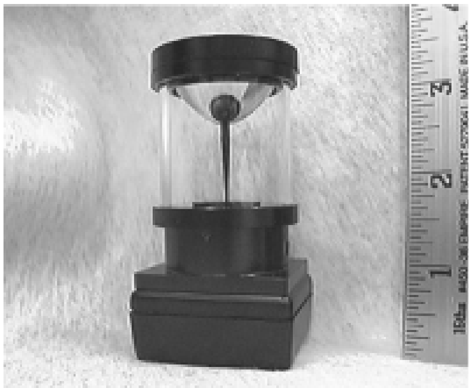

Figure 1 Developed compact ODVS.

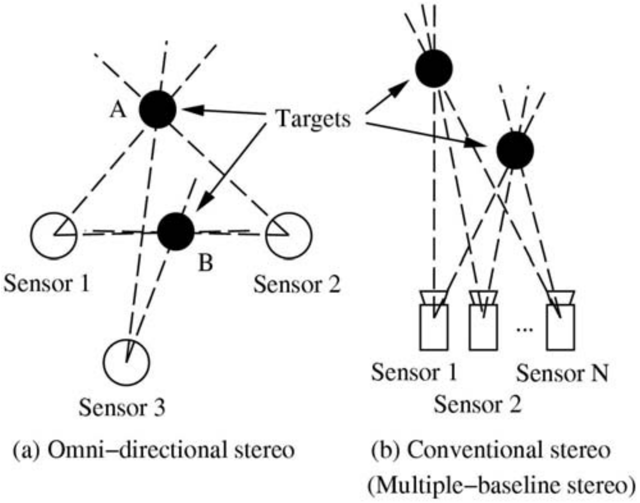

Figure 2 Omni-directional stereo and conventional stereo.

In order to solve the above problems, we have extended trinocular stereo [xix][xx]. The extended method, called *N-ocular stereo*, verifies correspondence of multiple targets without visual features. In addition, we have developed several compensation methods for observation errors in order to measure target locations more robustly and accurately.

The following subsection explains N-ocular stereo, which measures target locations using multiple ODVS. The N-ocular stereo is later simplified and error compensation methods are introduced for real-time processing. We show experimental results of people tracking in real time.

2.1 Multiple ODV stereo

Passive 3-D extraction approaches are most suitable for this application domain [xxi][xxii]. The passive stereo system, which we currently employ, is made of many ODVS. One of the merits of using the ODVS vision system is their ability to self-identify and self-localize [xxiii]. Based on the precise measurement of the camera positions, we perform multiple camera stereo.

Since each ODVS is fixed in the environment, the system can detect targets in omni-directional images by background subtraction. Then, it measures the azimuth angles of the targets. If the locations and the orientations of the ODVS are known, the locations of the targets can be measured from the azimuth angles by triangulation (see **Figure 3**).

In triangulation, multiple possible targets in the environment make the correspondence problem difficult. In **Figure 3**, for example, there are two targets (the black circles indicate actual target locations). However, from azimuth angles observed by sensors 1 and 2, it is estimated by triangulation that the targets may exist at A through D. In general, a correspondence problem can be solved using visual features of the targets. Nevertheless, because a target can be seen with different visual features at various viewpoints, it is difficult to verify the correspondence of the target with visual features alone. To make the system more robust, three or more sensors are used for the verification of the target correspondence. This technique is known as *trinocular stereo* [xix][xx]. In **Figure 3**, the locations C and D are verified and eliminated with sensor 3, since sensor 3 does not observe the targets in those directions.

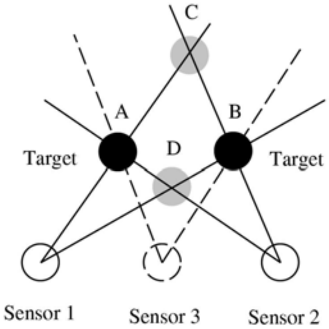

Figure 3 The correspondence problem and trinocular stereo.

2.2 Localization of targets by N-Ocular stereo

2.2.1 Basic algorithm

In trinocular stereo, three vision sensors are used to measure the target location and to verify the correspondence. On the other hand, in N-ocular stereo, more than three vision sensors are used. This is based on the idea that having more visual information reduces observation errors.

The basic process of our multiple camera stereo is as follows:

Step 1. Measure the location of a target from azimuth angles that are detected by every possible combinatorial pair of arbitrary vision sensors (**Figure 3**).

Step 2. Check if another sensor observes the target at the location measured with (N-1)-ocular stereo. If so, the location is considered a result of N-ocular stereo (see A and B in **Figure 3**). Iterate this step from N=3 to N=(the number of sensors in the ODVS cluster).

Step 3. The locations measured with only two sensors (C and D in **Figure 3**) are considered wrong matches. Erase them from the list of candidates.

2.2.2 Localization of targets and error handling

For the multiple camera stereo, we have employed a model-based method. Generally, the model-based stereo is more stable than the feature-based stereo in the case when the target is known. However, appearance of a human body frequently deforms in the image by changes in pose and viewing directions. Therefore, we have approximated the human body with a circle (the diameter is 60 cm) on the horizontal plane, which is parallel to the floor. The errors in the model matching can be handled with the following two parameters:

- α : Detection errors of the right and the left sides of a target
- β : An error of the human model, i.e., the error of the circle's radius

With the parameters α and β, the center of the circle is localized within the hexagon as shown in **Figure 4**. It is computed as follows: suppose that a target C with a radius r is observed from the sensor, and the detection error of the right and left side of the target is α, as shown in **Figure 4**. First, a circle C_- with a radius $(r-\beta)$ is considered. The black region in **Figure 4** indicates a possible region for the center location of the circle C_-, on the condition that the right and the left sides of the circle C_- are projected within $\pm\alpha$ from those of the target C, respectively. Here, the straight lines l and m are parallel to l' and m', respectively, and the black region indicates only the upper half of the possible region for the circle C_-. In the same way, the dark-gray region indicates a possible region for the center location of a circle C_+ with a radius $(r+\beta)$. Here, the straight lines l'' and m'' are parallel to l and m, respectively. Hence, the center of the circle whose radius is from $(r-\beta)$ to $(r+\beta)$ exists in the merged region of the black, the dark -gray, and the light-gray regions (**Figure 4** shows only the upper half of the region). This light-gray region indicates the location of the target C.

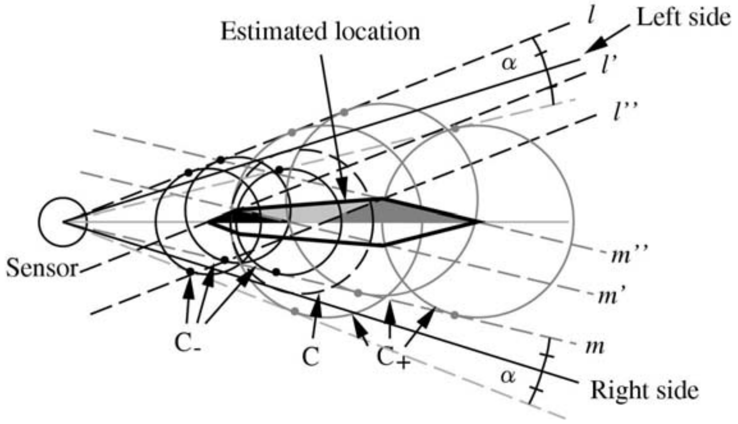

Figure 4 Localization of a target considering observation errors.

In the above method, target matches can be verified by checking if the hexagons overlap one another. Then, in the first step of N-ocular stereo, the target is localized at the overlapped region of two hexagons as shown in **Figure 5**. In the second step, the target is localized at the overlapped region of N hexagons. When α and β become smaller, the overlapped region also becomes smaller. In an ideal case, the overlapped region finally becomes a point, and that point can be considered the location of the target.

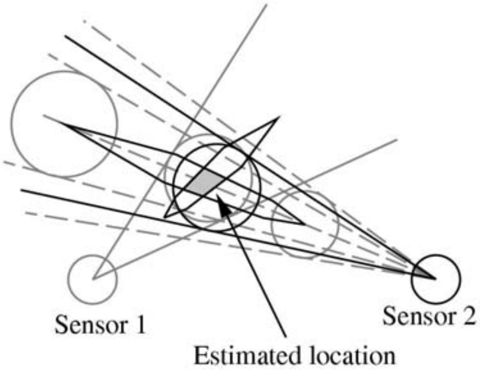

Figure 5 Localization of a target by binocular stereo.

2.2.3 False matches in N-Ocular stereo

N-ocular stereo can solve the correspondence problem of multiple targets in most cases. However, there is a particular arrangement of targets that it cannot deal with. In **Figure 6**, for example, it is estimated by N-ocular stereo that targets exist at up to four locations of A through D, including a false one (called a ghost target). In general, the ghost target cannot be eliminated, except to wait for the target to move away from that intersection [xxiv].

The false match in N-ocular stereo occurs when an azimuth angle of a target is associated with multiple locations (in **Figure 6**, an azimuth angle observed by the sensor 1 is used for the locations B and D). Therefore, if all of azimuth angles, which are used for measuring a target location, are also used for other locations, it is estimated that the location may be a false match (the location D in the case of **Figure 6**).

In the implementation, the false matches are considered in the process of target tracking. In the process, each of the currently measured locations is the nearest one to the previously measured locations, and the locations of false matches are detected after those of the true ones.

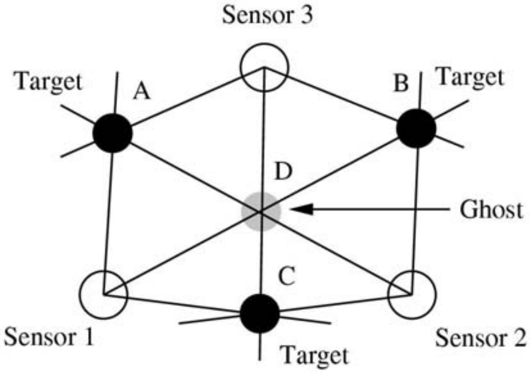

Figure 6 False matches in N-ocular stereo.

2.3 Implementing N-Ocular stereo

2.3.1 Simplified N-Ocular stereo

In N-ocular stereo described in the previous section, the verification costs of overlapped regions of hexagons and that of convergent operations are very high, and it is difficult to perform in real-time. Therefore, we have simplified N-ocular stereo as follows:

1. In the first step (binocular stereo), place a circle at the intersection of the azimuth angles detected by two arbitrary sensors, and assume that the circle is at the target location (see three black circles shown in **Figure 7**). Here, the radius of the circle is assumed to be 30cm since the targets are people.
2. In the second step (*N*-ocular stereo), check if the circles overlap. This verifies if the N^{th} sensor observes the target. If the circles overlap, place a new circle with a radius of 30cm at the center of gravity of the circles. It is considered as the target location measured with *N* sensors.

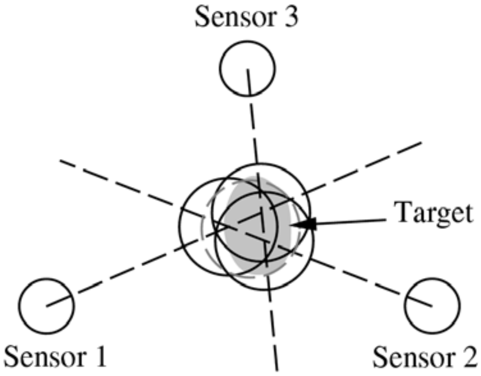

Figure 7 Simplified N-ocular stereo.

2.3.2 Error handling in the simplified N-ocular stereo

In the simplified N-ocular stereo, errors α and β described in **Section 2.2.2** are handled as follows.

α : Detection Errors of the Right and the Left Sides of a Target

Binocular stereo has a low-precision problem with respect to targets located along the baseline of the sensors [xviii]. In the simplified N-ocular stereo, this problem concerns the following situations. **Figure 8**(a), (b) and (c) show examples (in the first step of binocular stereo matching) where there is a target but no circle can be estimated since there is no intersection from the two lines that are associated with the azimuth angles. **Figure 8**(d) shows another example (in the second step of *N*-ocular stereo matching) where the target cannot be localized since the circles, which are placed in the step of *(N-1)*-ocular stereo matching, do not overlap.

Here, we introduce the following techniques to cope with the above problems.

- **When the two lines do not intersect on the account of observation errors:** If the angle between the baseline *l* of the two sensors and each of azimuth angles detected by the sensors (let these be θ_1 and θ_2) are equal to or less than α [see **Figure 8**(a) and (b)], consider that a target exists on the baseline *l*. Then, locate the target in such a way that the ratio of the distances between the target and each sensor (let this be $d_1 : d_2$) matches that of the apparent sizes of the target observed by the sensors. Consider that a target exists on the line representing the other azimuth angle (θ_1). If one of the azimuth angles (let this be θ_2) is equal to or less than α [see **Figure 8**(c)], correct the azimuth angle (θ_2) with $\Delta\theta\left(\Delta\theta \le \alpha\right)$ and locate the target in such a way that the ratio of the distances ($d_1 : d_2$) is close to that of the apparent sizes of the target.
- **When the two circles do not overlap:** The circles are considered overlapped, if they become overlapped after the adjustment of one of the azimuth angles with $\Delta\theta\left(\Delta\theta \le \alpha\right)$ [see **Figure 8**(d)].

β : An error of the human model

After the target is localized, the apparent size of the target reflected on each sensor can be computed from the distance between the sensor and the measured target location. If it differs by more than β from the actual size observed by the sensor, the measured location is considered a false match.

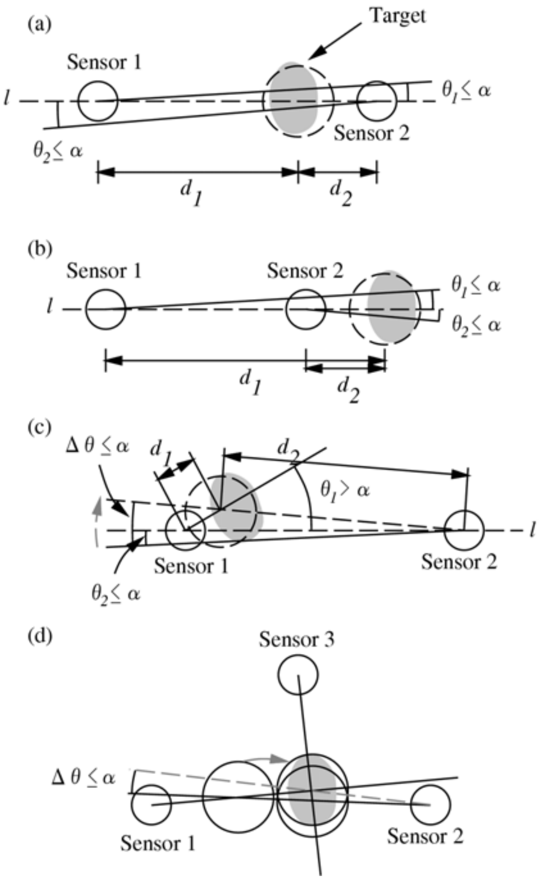

Figure 8 Error compensation.

3 Dynamic view synthesis

Another important function that supports the integrated surveillance system is the ability to generate views from arbitrary viewpoints. By integrating with real-time human tracking, we synthesize mainly two types of virtual view: one is the view that observes/follows a walking person and the other is the view that is seen from the walking person's perspective.

3.1 View synthesis for observing a walking person and from the perspective of a walking person

For synthesizing views that observe a walking person (**Figure 9**) and synthesizing a walking person's perspective view (**Figure 10**), we can utilize the walking trajectories provided by the human tracking algorithms along with the information of the ODVS locations. The pose of the walking person can also be roughly estimated from the walking vector of the profile. A virtual view is synthesized at a constant distance away from the walking person for observation purposes. The center of the virtual view points against the walking vector of the profile. The center viewing frustrum passes through the centroid of the walking person. While synthesizing the walking person's perspective view, the virtual viewpoint is at the viewpoint of the walking person and the center viewing frustrum points in the same direction as the walking vector of the profile.

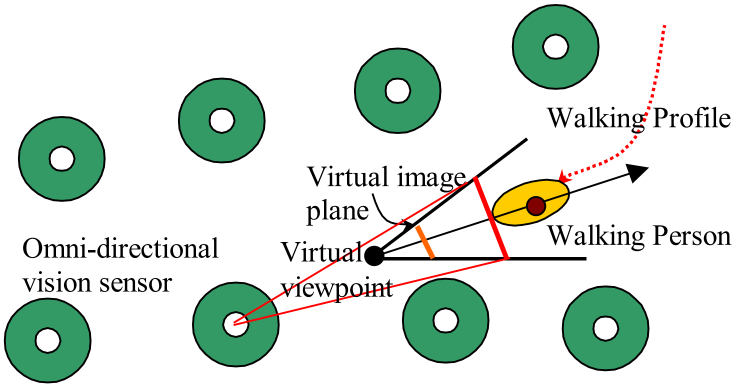

Figure 9 View synthesis for observing a walking person.

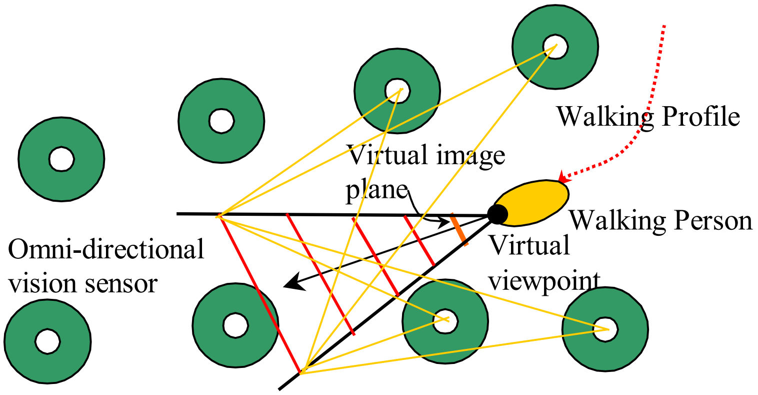

Figure 10 View synthesis from the perspective of a walking person.

3.2 Range-space approach for virtual view synthesis

For synthesizing a virtual view, we have developed a range-space approach. The general idea is similar to the multiple baseline stereo proposed by Okutomi [xvi], but we have modified it for the *general camera configuration*. Generally, cameras used for multiple camera stereo are arranged in a line; and the search for finding corresponding feature points is performed in the disparity/image space based on the image plane structure. However, if we suppose a general camera arrangement, we have difficulty in performing the search in the disparity space, especially if the cameras are not closely positioned and their field of view is large (such as of the ODI).

Range-space approach directly searches /matches /renders along the virtual rays, starting from the virtual viewpoint in the range space. The searching method is equivalent to what is called "Forward Projection" in computer graphics terms, and it is the inverse mechanism of common 3D accelerator cards. Forward projection processing time is independent of the complexity of the scene composition, and its processing time is in the same order as the number of virtual pixels. The searching interval is non-uniform, depending upon the image resolution and the camera arrangement. The incremental voxels on the virtual rays are backward-projected to the cameras, searching for the best correspondence. This part is the inverse mechanism of the common stereo matching practice in the computer vision community.

Figure 11 shows a view of four ODVS in a multiple baseline stereo configuration. The starting point for the range-space search is at the virtual viewpoint. The search intervals along the virtual 3-D line are determined based on the pixel resolution of the nearest ODVS to preserve maximum range resolution during the search, as illustrated in **Figure 12**. When the range resolution is preserved, feature points in the image will not be missed during matching. Note that pixels in the ODI cover the 3-D space with different resolutions. The pixel at the image center covers the largest 3-D space compared to the one at the outer radius.

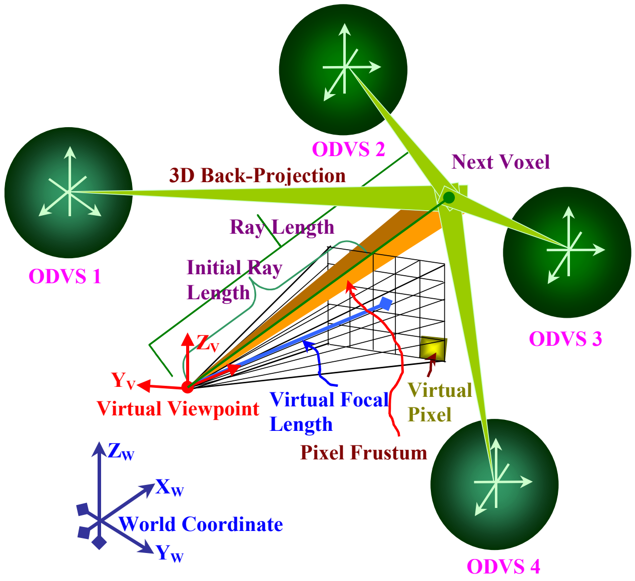

Figure 11 Illustration of Range-Space Approach.

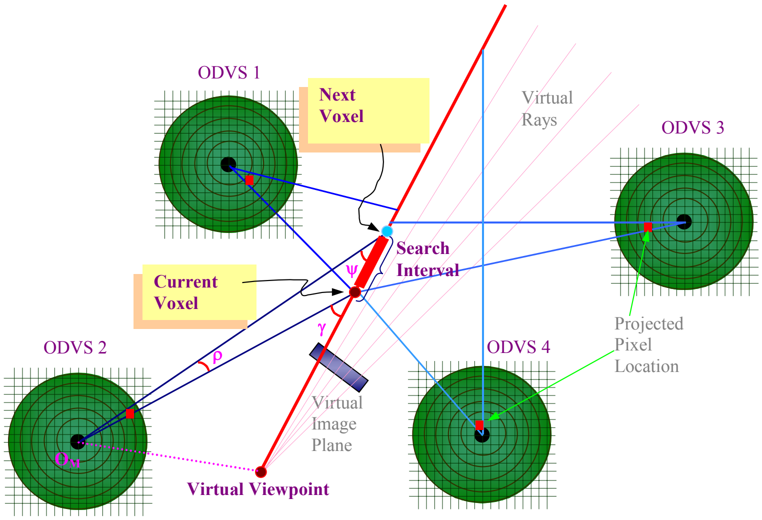

Figure 12 Illustration of range-space search. The criterion is to preserve the maximum depth resolution. In this illustration, ODVS 2, which contains the least visual information, decides where the next voxel is.

Range-space approach has the characteristics of not needing a 3D model and having high parallelism for every virtual pixel. The input data to the algorithms are raw omni-directional video images rather than the 3D model. Raw images are much easier to transmit through the network in comparison with the complete 3D model. Most importantly, it performs exactly at the user-specified viewpoint. This is what we call "*View-on-Demand*." The process is different from how geometrically valid virtual views have been generated in past years. In the past, 3D models were first created for the entire scene and then voxels were projected to the small window of new virtual camera's location. In integrated Surveillance System setup, the cameras' coverage is large and the scene is dynamic but the virtual viewing area is small. Following this conventional method of recovering depth for every physical pixel and then re-projecting voxels to the virtual view is too time-consuming. However, when range-space approach is introduced, we calculate only the necessary 3D along the virtual rays. These voxels are a much smaller subset of the entire wide-angle view. Consequently, range-space approach is highly promising for real-time view generation of dynamic scenes for dynamic observers. Viewers' computer can generate virtual views by the means of Distributed Computing.

The range-space approach brings in the capability of performing wide-baseline stereo on omni-directional videos with an arbitrary camera arrangement. The ability to perform wide-baseline omni-directional stereo allows for a reduction in the number of cameras required to cover a large scene. There are a total of five major challenges when dealing with the wide-baseline stereo. These challenges are scaling effect, foreshortening effect, window cutoff, specular highlight, and occlusion. The scaling effect becomes an issue due to the fact that when cameras are placed far apart, objects appear much larger in closer cameras than in the farther ones. This makes it difficult to perform the correspondence in the disparity space even if there is no occlusion. In the range-space stereo approach, the matching template will be created larger for a closer camera while smaller for a farther camera. A pixel in the farther camera covers the space larger than a pixel in the closer camera. Therefore, it is not a one-to-one matching process; that is one pixel in the farther camera will compare with multiple pixels in the closer camera. When searching in the range space, we also have the full control of the orientation of the 3D-matching template. The matching templates in different images are the projection of a common 3D-matching template (volumetric matching template) in the range-space. The volumetric matching template has the center positioned at the voxel to be investigated and the radius specified based on the image resolution and the camera configuration. By rotating the 3D-matching template in various directions, in principle, we can accommodate the foreshortening effect. Window cutoff, that is when a voxel is inside the field of view of one camera but out of view of the other, can be easily detected when the back-projected pixel location of the voxel is out of bound of its image sensor array.

Further, for finding the best match, we have modified the computation of template matching for multiple camera stereo. Suppose that three ODVS observe a common object but one other ODVS is occluded from that common object. If standard template matching is performed, the result will suffer significant errors. Our idea is to minimize the error by choosing the best-suited ones among these ODVS employing the technique in robust statistics. The matching templates are represented as points in one-dimensional parameter space. In this parameter space, outlier detection is performed using standard deviation [xxv] and rejects ODVS, which are on the extreme values. This method dynamically selects ODVS that provides the best match and avoids the occlusion problem that happens especially in cases where the ODVS are placed at arbitrary locations.

Our purpose is not to acquire precise range information in this work, but to robustly provide smooth virtual view sequences for monitoring dynamic events. Large templates having the same resolution as the desired virtual image are used for robust matching. This range-space search accommodates the effects of scaling, foreshortening, and window distortion which larger baseline stereo inherits, and that simple rectangular template in disparity space cannot trivially solve. I don't understand this sentence.

4 The integrated surveillance system

4.1 System configuration

We have developed the integrated surveillance system by using the proposed techniques of real-time human tracking and view synthesis. The system uses four ODVS in the laboratory as shown in **Figure 13**. ODVS is mounted 1.3m from the floor. The images (**Figure 14**) taken by the ODVS are sent to a quadrant image unit and its output (standard video signal) is sent to a computer. The computer has Pentium 400MHz CPU, 128Mbyte memory and a standard image capture board (Matrox).

In **Figure 14**, a person is seen by the four ODVS at the same time instant. The layout of the images is in the same order as the sensor layout shown in both **Figure 13** and **Figure 15**. **Figure 15** shows an example screen shot of the developed system.

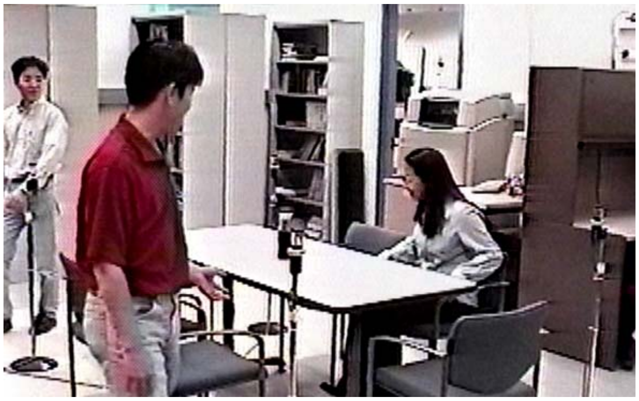

Figure 13 System overview.

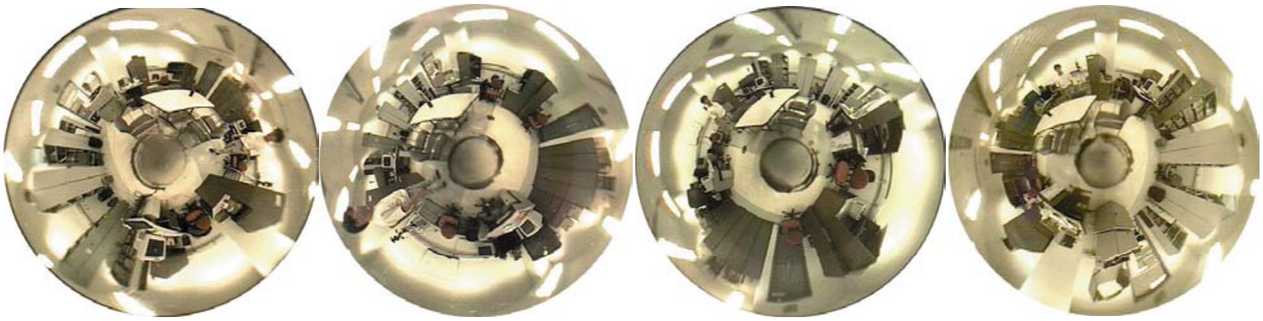

Figure 14 ODI taken by ODVS.

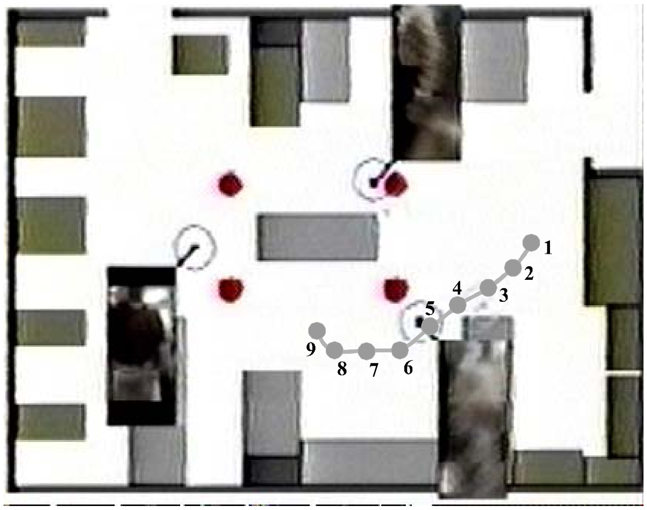

Figure 15 Screen shot of the developed system.

In addition to the real-time tracking, the system can show the best view for watching the walking person. In the figure, the four gray circles around the table indicate the ODVS positions. The three circles and small images attached to them are peoples' locations and the views observing the walking person, respectively. The system covers the room of 7×9m with four ODVS. The resolution of the image plane for each ODI is 320×240 pixels and the generated view is 40×80 pixels. Although this is enough for human behavior tracking and simple visual surveillance, it can be combined with standard pan-tilt cameras for acquiring better views. The trajectory numbered from 1 to 9 is the profile of a walking person. The locations with numbers will be discussed in following sections. .

4.2 Real-time tracking of walking people

The method proposed here robustly tracks a walking person in real time with four ODVS. In every frame, background subtraction is performed for detecting people in the omni-directional images. Then each of the locations measured by N-ocular stereo is related to the nearest one of the previously measured locations. **Figure 16** shows a set of trajectories of a walking person. The system was continuously acquiring video streams for about 3 minutes. The system is computing the average location of each person with the duration of ½ second for smoothing. Resulting tracks indicate high localization accuracy, robustness in tracking as well as the basic real-time performance. The system could track the person without loosing sight in real time. We have confirmed that the trajectory matched the marked trajectory on the floor.

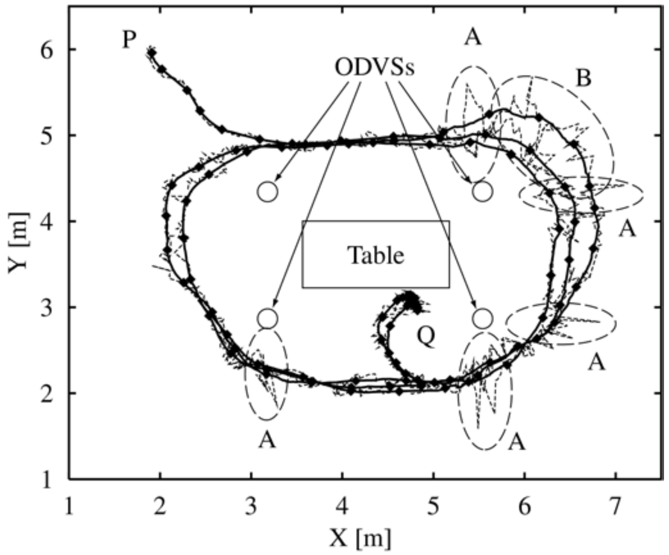

Figure 16 Real-time trajectory detection of a walking person.

The broken lines in **Figure 16** indicate the person's locations at every frame before smoothing. There are large errors around A and B. This is because (1) binocular stereo using ODVS has a low-precision problem with respect to targets located along the baseline of the sensors (this corresponds to A in **Figure 16**), and (2) the result of background subtraction becomes noisy if the color of person's clothes is similar to that of the background (this corresponds to B in **Figure 16**). In the latter case, general noise filtering techniques such as the Kalman filter may not be able to successfully eliminate the noise, since the noise is different from white noise. It is effective to add additional sensors to cope with this kind of noise.

The system's task is not to precisely measure the people's locations, but to robustly track them. It tracks people who are wearing different types of clothes, on the top that people are deformable I don't understand what this. In this implementation, the system could simultaneously track three people at video rate (30 fps). The experimental results show that the system completely tracked one person. It correctly tracked two persons 99% of the time, and it correctly tracked multiple people (up to three people) 89% of the time. The 10% errors mainly occurred when the fourth person was mistakenly detected in the scene due to false matches. Tracking errors occurred in the following cases:

- When two or more people moved closer to each other, the system treated them as a single person.
- When one person moved behind anther person, the system could not measure its location.

An important point is that the system can obtain good performance without referring to the visual feature of the targets, although utilization of features should decrease incorrect matches. In order to solve the above error, a better technique needs be developed for detecting people. Additionally, increasing the number of ODVS employed will increase the robustness of tracking further. It will also allow more people to be tracked in the scene simultaneously. We are planning to implement a system consisting of sixteen ODVS in the future.

4.3 View synthesis

Figure 17 are the ODI of the closest camera to the trajectory showing a walking person from points 7 to 9 as indicated in **Figure 15**. **Figure 18** shows the virtual views for observing a walking person from points 1—9. The virtual camera follows the walking person one-meter away from his detected position. The viewing direction is 30 degrees deviated from the person's heading. Images 1, 2, 3, 8, and 9 are generated from the closest camera to the trajectory, while images 4—7 are from the bottom left camera, based on the criterion of distance and motion direction. Images 4—7 have a grainy look since the views are "digitally zoomed" in from the original images to create the view from the virtual camera location. **Figure 19** shows a virtual image sequence from the tracked person's perspective. The walking path is the same one. All of the views were selected from the bottom right camera, since the camera has the closest distance to the tracked person and the viewing direction aligns best with it. It is easy to observe that the tripod is getting larger as the person walks closer to it and that the person is turning to the right-hand side.

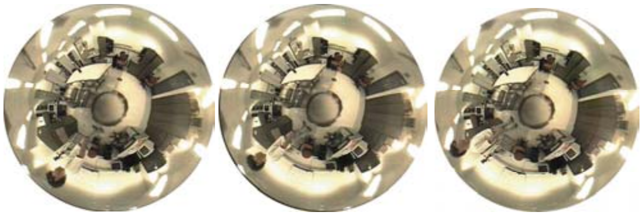

Figure 17 A ODVS observes a walking person at three arbitrary instances.

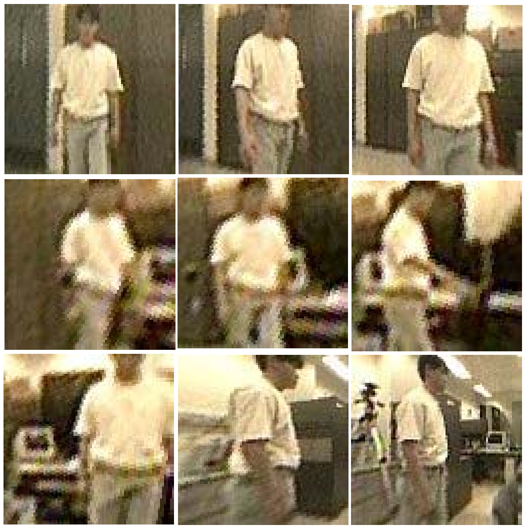

Figure 18 Virtual image sequence (from left to right) for observing a walking person.

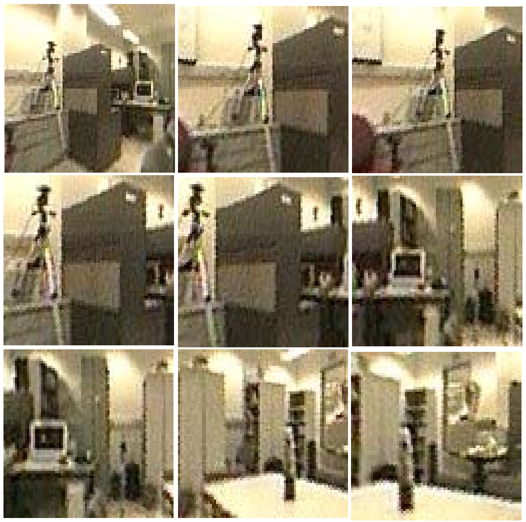

Figure 19 Virtual image sequence (from left to right) of an observer's view.

5 Concluding Remarks

The developed system for visual surveillance is built upon the concept of ubiquitous vision [i]. The system consisting of ODVS is a platform to access the physical and virtual world. We have developed the system based on three key ideas: (1) original design of low-cost and compact ODVS, (2) real time tracking of walking people, and (3) arbitrary view synthesis. We believe this practical system opens up new possibilities and issues grounded in real time monitoring of dynamic environments.

An integrated surveillance system with well-integrated sensors, communication and computers, is feasible and can be made practical for wide area coverage. There are still many interesting and challenging issues to be resolved. These efforts will have to be framed in the context of the utility of the systems in meeting specific needs of humans, i.e. applications. Surveillance can be viewed as one example. Others may include distance learning, collaborative environments, or “tele-touring”. Once the basic technological infrastructure can assure a reliable surveillance system, it will become possible to effectively address research problems of important cognitive and social sciences. Problems where we can systematically investigate how we interact with each other and with sensorized “intelligent” environments, which are quite important, will become manageable in such an infrastructure.

Robustness and generality of the approach was evaluated in some of our other studies. More discussion on the performance of human tracking was presented in [xxvi]. We have also used the range-space approach to synthesize views in a hallway environment [xxvii]. The smooth virtual walkthrough was compared with a real, rectilinear camera walking through the same path. The synthesized views resembled the real ones very closely.

It is also important to point out that the view synthesis technique presented here is different from the method presented in [xxviii]. In [xxviii], the view was generated using a single effective viewpoint that has the virtual image plane perpendicular to the center ray that is connected to the mirror’s focal point. Instead, the virtual view presented in this paper is the perspective transform of a portion of the original ODI onto the new virtual image plane that is perpendicular to the center ray that is connected to the virtual viewpoint. This sentence is difficult. The virtual viewpoint can be anywhere in the 3-D space. The perspective transform foreshortens the ODVS’ view, and thus gives us the feeling of a new view.

We have also extended this line of research to develop the range-space stereo approach to extract depth and color simultaneously for every virtual pixel [xxix]. Every virtual pixel is a composite color from multiple cameras. The synthesized views are perspective views and they are physically valid. This allows overcoming the limitation of object *skewing* in the perspective-transformed view, which is the perspective transform of a portion of the original ODI onto the new virtual image plane, as described in this paper.

All of the studies considered so far use multiple ODVS; however, the basic approach is applicable to the conventional rectilinear images. It is also to integrate the low-resolution omni-directional images with high-resolution rectilinear images to produce high visual quality virtual view. Capella recently demonstrated one unique form of image-based rendering [xxx], which does not fall directly into the two methods (light-field interpolation and morphing). A high-resolution image was manually registered and blended seamlessly possible with a low-resolution omni-directional image. The virtual walkthrough was linearly translating between two omni-directional image centers without geometric modeling.

Acknowledgements

The authors would also like to thank Sony Electronics Corporation, Compaq Computers, and the California Digital Media Innovation Initiative (DiMI) for their support.

References

[i] K. C. Ng, M. M. Trivedi, and H. Ishiguro, “3D ranging and virtual view generation using omni-view cameras,” Proc. Multimedia Systems and Applications, SPIE vol. 3528, Boston, November 1998.

[ii] H. Ishiguro, “Distributed vision system: A perceptual information infrastructure for robot navigation,” Proc. IJCAI, pp. 36-41, 1997.

[iii] H. Ishiguro, “Development of low-cost and compact omnidirectional vision sensors and their applications,” Proc. Int. Conf. Information Systems, Analysis and Synthesis, pp. 433–39, 1998.

[iv] P. J. Narayanan, P. W. Rander, and T. Kanade, “Constructing virtual world using dense stereo,” Proc. ICCV, pp. 3–10, 1998.

[v] S. M. Seitz and K. N. Kutulakos, “Plenoptic image editing,” Proc. ICCV, pp. 17–24, 1998.

[vi] E.H. Adelson and E.H. Bergen, “The plenoptic function and the elements of early vision,” Computation models of visual processing (M. Landy and J.A. Movshon eds.), MIT Press, 1991.

[vii] M. Levoy and P. Hanrahan, “Light field rendering,” Proc. SIGGRAPH, pp. 31–42, 1996.

[viii] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, “The lumigraph,” Proc. SIGGRAPH, pp. 43–54, 1996.

[ix] J.E. Boyd, E. Hunter, P.H. Kelly, L.-C. Tai, C.B. Philips, and R.C. Jain, “MPI-Video infrastructure for dynamic environments,” in Proc. IEEE Int. Conf. Multimedia Computing and Systems, pp. 249–54, 1998.

[x] T. Kanade, T. Collins, A.J. Lipton, P. Anandan, P. Burt, and L. Wixson, “Cooperative multi-sensor video surveillance,” In DARPA Image Understanding Workshop, vol. 1, pp.3–10, 1997.

[xi] A.J. Lipton, H. Fujiyoshi, and R.S. Patil, “Moving target classification and tracking from real-time video,” In DARPA Image Understanding Workshop, pp.115–22, 1998.

[xii] R.T. Collins, Y. Tsin, J.R. Miller, and A.J. Lipton, “Using a DEM to determine geospatial object trajectories,” In DARPA Image Understanding Workshop, 1998.

[xiii] T. Mori, Y. Kamisuwa, H. Mizoguchi, and T. Sato, “Action recognition system based on human finder and human tracker,” In Proc. Int. Conf. Intelligent Robots and Systems (IROS), pp. 1334–42, 1997.

[xiv] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee, “Using adaptive tracking to classify and monitor activities in a site,” In Proc. IEEE Computer Vision and Pattern Recognition (CVPR), pp. 22–9, 1998.

[xv] M.M. Trivedi, K. Huang, and I. Mikic, “Intelligent environments and active camera networks,” IEEE System, Man and Cybernetics, October 2000.

[xvi] M. Okutomi and T. Kanade, “A multiple baseline stereo,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, pp.353–63, April 1993.

[xvii] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, “A stereo machine for video-rate dense depth mapping and its new applications,” Proc. IEEE Computer Vision and Pattern Recognition (CVPR), pp.196–202, 1996.

[xviii] H. Ishiguro, M. Yamamoto, and S. Tsuji, “Omni-directional stereo,” IEEE Trans. PAMI, vol. 14, no. 2, pp. 257–262, 1992.

[xix] M. Yachida, “3-D data acquisition by multiple views,” In 3rd international Symposium on Robotics Research (ISRR’85), pp. 11-18, London, 1986.

[xx] E. Gurewitz, I. Dinstein, and B. Sarusi, “More on the benefit of a third eye,” In Proc. Int. Conf. Pattern Recognition (ICPR), pp. 966–68, 1986.

[xxi] S. B. Marapane and M. M. Trivedi, “Multi-primitive hierarchical (MPH) stereo analysis,” IEEE Transactions on PAMI, vol. 16, no. 3, 1994.

[xxii] A. Dalmia and M. Trivedi, **“**Depth extraction using a single moving camera: an integration of depth from motion and depth from stereo,” Machine Vision and Applications, vol.9, (no.2), 1996. pp. 43–55.

[xxiii] K. Kato and H. Ishiguro, “Identifying and localizing robots in a multi-robot system,” Proc. Int. Conf. Intelligent Robots and Systems (IROS), p. 966-972, 1999.

[xxiv] S. Shams, “Neural network optimization for multi-target multi-sensor passive tracking,” Proc. IEEE, 84(10):1442-1457, 1996.

[xxv] Statistical methods in research and production: with special reference to the chemical industry, edited by Owen L. Davies and peter L. Goldsmith. 4th rev. ed. London: Published for Imperial Chemical Industries, ltd. By Longman Group, 1976. xiii, 478 p.: ill.; 24 cm.

[xxvi] T. Sogo, H. Ishiguro, M. M. Trivedi, “N-ocular stereo for real-time human tracking,” Panoramic Vision: Sensors, Theory and Applications, (R. Benosman and S.B. Kang, eds.), Springer Verlag, 2001.

[xxvii] K. C. Ng, H. Ishiguro, and M. Trivedi, “Multiple omni-directional vision sensors (ODVS) based visual modeling approach,” *Conference and Video Proceedings of IEEE Visualization '99*, San Francisco, California, October 1999.

[xxviii]T. Boult, “*Remote reality demonstration*,” In IEEE Conf. Computer Vision and Pattern Recognition, p. 966–967, Santa Barbara, CA, June 23-25, 1998.

[xxix] K. C. Ng, *3D Visual Modeling and Virtual View Synthesis: A Synergetic, Range-Space Stereo Approach Using Omni-Directional Images*, Ph.D. Dissertation, University of California, San Diego, March 2000.

[xxx] R. Capella, *Real-Time System for Integrated Omni-Directional and Rectilinear Image-based Virtual Walkthroughs*, Master Thesis, University of California, San Diego, December 1999.