Monitoring Dynamically Changing Environments by Ubiquitous Vision System

Kim Ng, Hiroshi Ishiguro*, Mohan Trivedi**, and Takushi Sogo***

* Department of Social Informatics, Kyoto University, Japan
** Department of Electrical and Computer Engineering, University of California, San Diego, USA

Abstract

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

Keywords: real-time tracking, multiple camera stereo, human activity profiling, dynamic view synthesis and remote observation

1 Introduction

Networks of large number of cameras are required to provide wide scene coverage for many surveillance tasks. In designing such networked camera systems, considerations of practical aspects, for matters such as cost, complexity and robustness, must be given. This paper discusses the associated issues and describes development of a multiple camera system, called ubiquitous vision system, which allows remote observers to view dynamic environments from arbitrary vantage points efficiently and realistically.

In our system, an original concept of *ubiquitous vision* is brought out. Requirements for such a system have been discussed in [Ng, 1998]. The word "ubiquitous" means "being or seeming to be everywhere at the same time" (from the American Heritage dictionary). The ubiquitous vision system that we develop enables observing the environment from arbitrary viewpoints in real time, as the word "ubiquitous" indicates.

Related works

Comparing with other multiple camera approaches [Narayanan, 1998; Seitz, 1998], our system differs mainly in the configuration and utilization of many cameras. They were using rectilinear images for stereo matching and their cameras were arranged densely with short baseline. The scene where they were looking at was restrained to small area; therefore, their methods are not readily applied to our problem for generating views, which especially requires the coverage of large scene area. Besides, our images are omni-directional and having their own unique properties. New algorithms have to be developed or modified to accomplish our tasks.

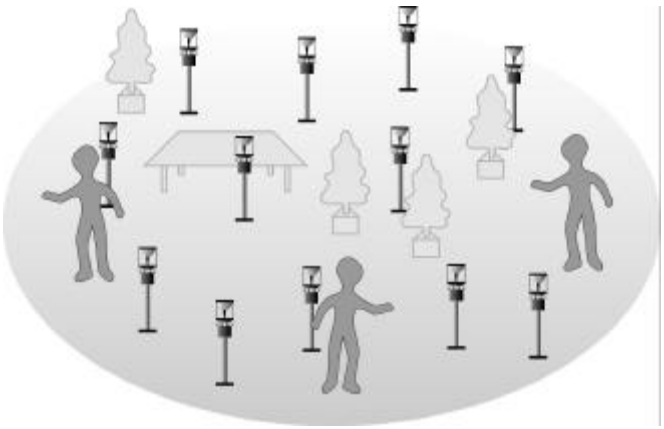

Figure 1: Ubiquitous vision system

There are other types of multiple camera system from the computer graphics community. Their works are dealing with reconstructing visual information by approximating *plenoptic* functions [Adelson, 1991]. *Lumigraph* [Levoy, 1996] and *light-field rendering* [Gortler, 1996] are the two most prominent methods in the approximation. These methods are generally simpler than the methods derived from computer vision community. They do not require matching between images; however, a very large number of cameras or image samples are needed by placing the cameras or by acquiring images very densely in close proximity.

All of these multiple camera systems are 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 large scene area, where dynamic objects prevail.

Recently, systems monitoring wide areas on practical purposes have been proposed especially for visual surveillance [VSAM, 1999]. The work reported in this paper has similar purpose as theirs, but our approach is slightly different. Those works using multiple cameras to cover wide scene areas seldom take advantage of data redundancy from multiple cameras as we do. The large overlapping viewing area from multiple cameras provide these data redundancy.

The concept of ubiquitous vision

Ubiquitous vision system utilizes redundant visual information for robust monitoring tasks in large scene area. Several vision sensors observe a common area and provide redundant information. This redundant observation contains rich information for robust vision functions. The ubiquitous vision concept follows the one of distributed vision proposed by Ishiguro [Ishiguro, 1987] with more generality and practicality. Key specifications of the ubiquitous vision system are summarized as follows:

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

Based on the above specifications, we have developed a vision system using multiple ODVSs. In the following sections, we first briefly introduce an original design of the ODVSs. Then a real-time human tracking system using the ODVSs is described as one of the functions for surveillance purpose. Further, we show that the system has capability of synthesizing novel virtual views at arbitrary viewpoints.

2 Development of Compact ODVSs

The ODVS has been first proposed by Rees [Rees, 1970] in the patent submitted to US government in 1970. Then, Yagi [Yagi, 1990], Hong [Hong, 1991] and Yamazawa [Yamazawa, 1993] developed again in 1990, 1991 and 1993, respectively. Recently, Nayar [Nayar, 1997] has geometrically analyzed the complete class of single-lens single-mirror catadioptric imaging systems and developed an ideal ODVS using a parabola mirror.

Figure 2: Developed compact ODVS

In these previous works, researchers developed ODVSs as prototypes and investigated properties of *Omnidirectional Images* (ODIs) taken by ODVSs. Therefore, the developed ODVSs were not so compact and their costs were high. In order to develop practical vision systems, we have proposed original ideas and developed *low-cost and compact ODVSs*. Figure 2 shows the developed compact ODVSs [Ishiguro, 98].

3 Real-Time Human Activity Tracker

3.1 Multiple ODV stereo

Passive 3-D extraction approaches are most suitable for this application domain [Marapane, 1994; Dalmia, 1996]. The passive stereo system, which we are employing, comprises of many ODVSs. One of the merits of using ODVS vision system is their ability of self-identification and self-localization precisely [Ishiguro, 1999]. Based on the precise measurement of the camera positions, we perform multiple camera stereo. This multiple camera stereo is an extension of trinocular stereo [Kitamura, 1990] to many cameras. The basic process of our multiple camera stereo is as follows:

1. Detect moving regions on the images by background subtraction.
2. For all combination of two moving regions on different ODVSs, estimate the distance to the region and apply a circle as a human model (the larger circles in Figure 3).
3. Check the overlapping of the circles and determine person's locations.

For stable background subtraction against lighting conditions, several methods can be employed. In this system, we have implemented a simple method for real-time processing. We have taken 10 frames for the background image and estimated noise range for each pixel. Based on the estimated noise, pixels that represent the moving regions are detected.

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 where the target is known. However, appearance of a human body frequently deforms on the image by changes of the 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.

When over three circles overlap each other (the number should be determined according to the number of ODVSs and the size of the environment), we decide a person exists in the center of gravity of the circles. The

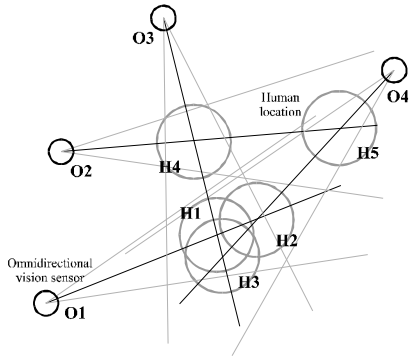

Figure 3: N-ocular stereo with ODVSs

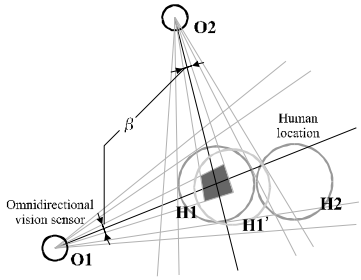

Figure 4: Error compensation

overlapping can be checked by making a graph and by analyzing the structure. For the graph, the ODVSs are refereed as nodes and an arc exists between a pair of ODVSs used for determining the position of a circle. If a complete graph exists, we assure the circles overlap each other.

3.2 Treatment of deformable human bodies

The difficulty of this method is to approximate a human body with a circle. The deforming human body sometime brings serious errors. The followings cause the errors:

1. Observation of the human body from different viewpoints.
2. Motions of components of a human body, such as hands.
3. Low precision of a binocular stereo for targets locating along the baseline [Ishiguro, 1991].
4. Low precision of a binocular stereo for distant targets.

If the system has sufficient number of ODVSs, the redundant observations solve these problems. However, it is difficult to use a very large number of ODVSs in some situations. A method that can deal with the problem by using a given number of ODVSs is needed.

The idea we have employed is to shift the positions of the circles as shown in Figure 4. Suppose there are two circles, H1 and H2, as results of stereo matching using pairs of ODVSs. If the circles do not overlap each other the person cannot be detected since the system needs three circles overlapping each other at least. In this case, we swivel the viewing directions to the target in the range of ß , and move the circle H1 in the gray area. We assume 30% errors for the human detection on the image and setß as 1.6 × [projection size of the circle]. If the circle H1 overlaps with H2, we assure the circles overlap each other. The system applies this process for all circles.

4 Dynamic View Synthesis

Another important function that supports the concept of ubiquitous vision is to provide 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 seeing from the walking person's perspective.

4.1 View synthesis for observing walking people

For synthesizing views that observe a walking person, we can utilize the walking trajectories provided by the real-time human tracking algorithm along with the information of the ODVSs locations. It allows us to easily select the best ODVS for observing the person. The omnidirectional sensor is selected based on the distance to a walking person and the motion direction of the person (see Figure 5).

In the case where there are obstacles between the person and the ODVS, the method cannot be applied. For dealing with the obstacle, we can select the best ODVS in a same way as discussed below. However, in this case, we do not need to acquire the range information.

4.2 Walking person's view generation

In the case of synthesizing the view of a walking

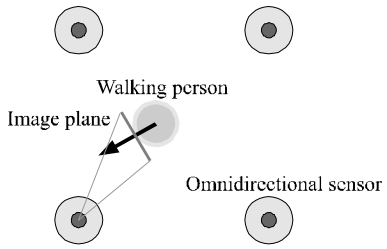

Figure 5: Virtual-view generation

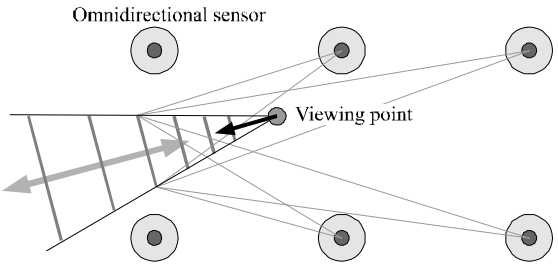

Figure 6: Walking person's view generation

person's perspective, we need to estimate ranges to objects that the person is watching. That is, a general method for view synthesis at an arbitrary viewpoint is needed. For synthesizing walking person's views, we have used area-based stereo instead of the model-based stereo used for human tracking. The general idea is same as the multiple baseline stereo proposed by Kanade and his colleagues [Kanade, 1990], but we have modified it for our ubiquitous vision system. Our originalities are listed in the modifications listed below:

- A *range-space search* method for finding corresponding points that accepts arbitrary camera positions and wide baseline configuration.
- Robust template matching using all possible combinations of ODVSs' pairs.
- Utilization of large templates for robust matching.

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 search is to search for the best match in the 3-D space and the assumed 3-D points are mapped into the image coordinates for template matching. The mapping algorithm can be found in the patent of these ODVSs. Figure 6 shows the range-space search that is performed between multiple ODVSs in the stereo matching process. Here, the search intervals are determined based on the pixel resolution of the nearest ODVS to preserve maximum range resolution during the search. When the range resolution is preserved, feature point 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 than the one at the outer radius.

Further, for finding the best match, we have modified the computation of template matching. By minimizing the following error *E*, we can find the best match:

$$E = min\left(\sum_{w}\left(p_i - p_j\right)\Big/ n(i,j)\right)$$

where *i* and *j* are IDs of the ODVSs. $min(f(i,j))$ means to take the minimum value among $min(f(i,j))$, $i,j = 1 \ldots N$ (N is the number of ODVSs used for matching). $n(i,j)$ is a nomalization factor determined according to the number of pixels contributing to the matching. This method dynamically selects ODVSs that provides the best match based on the matching error *E* and avoids the occlusion problem that easily happens especially in the case where the ODVSs are placed at arbitrary locations.

Our purpose is not to acquire precise range information in this work, but to robustly provide smooth

Figure 7: System overview

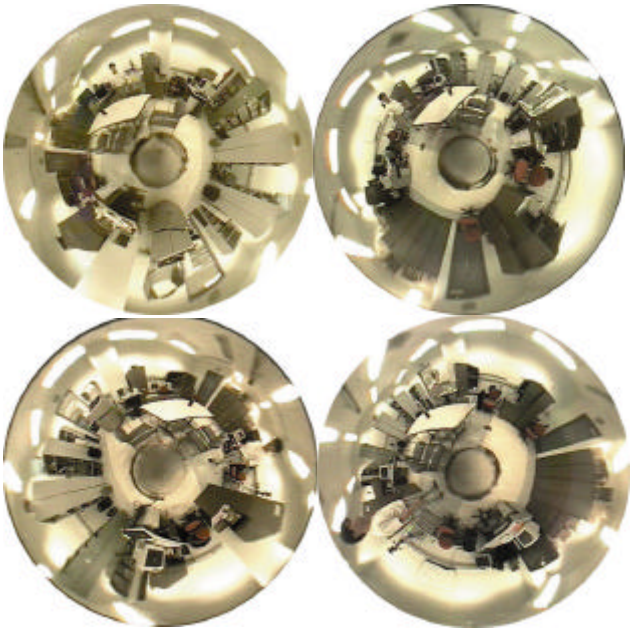

Figure 8: ODIs taken by ODVSs

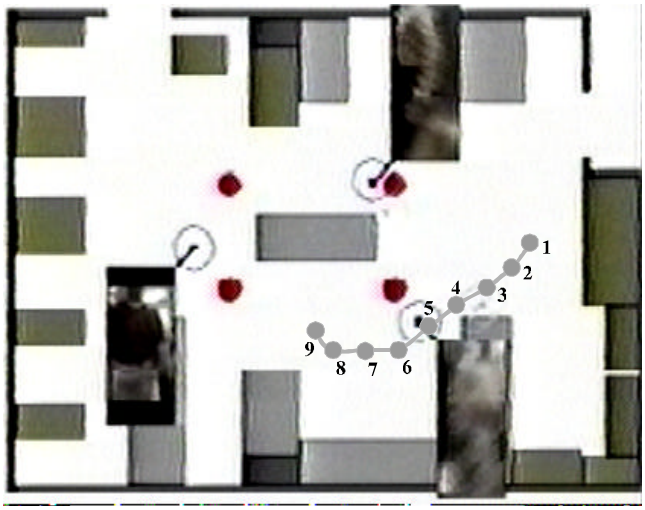

Figure 9: Screen shot of the developed system

image sequences in real time for monitoring dynamic events. Large templates having the same resolution as the desired virtual image is used for robust matching. This range-space search accommodates the window distortion that larger baseline stereo inherits, which simple rectangular template in disparity spacecannot trivially solve.

5 The ubiquitous vision system

5.1 System configuration

We have developed the ubiquitous vision system by using the proposed techniques of real-time human tracking and view synthesis. The system uses four ODVSs in the laboratory as shown in Figure 7. The height of the ODVS is 1.3m from the floor. The images (Figure 8) taken by the ODVSs are sent to a quadrant image unit and its output (standard video signal) is sent to a computer. The computer has Pentium 200MHz CPU, 128Mbyte memory and a standard image capture board (Matrox).

In Figure 8, a person is seen by the four ODVSs at the same time instant. The layout of the images is in the same order as the sensor layout shown in both Figure 7 and Figure 8. Figure 8 shows an example 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 real-time. In the figure, four gray circles around the table located at the center indicate the ODVS positions. Three circles and the small images attached to them are people locations and the views observing the walking person, respectively. The system covers the room of 7×9m with four ODVSs. 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 with number indicated from 1 to 9 is the walking profile of a walking person. The locations with number are for the following discussions.

5.2 Real-time tracking of walking people

The method proposed here robustly tracks walking person in real time with four ODVSs. Figure 10 shows a set of trajectories of a walking person. The system was continuously acquiring video streams for about 3 minutes. Resulting tracks indicate high localization accuracies, 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.

The system 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. However, the experimental results show that the system correctly tracked multiple people (up to three people) 90% of the time. The 10% errors mainly occurred when more than three people were in the scene.

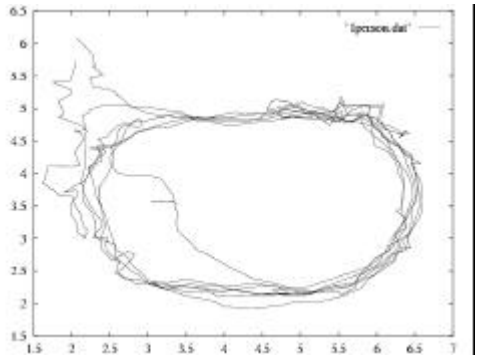

Figure 10: Real-time trajectory detection of a walking person


Figure 11: A ODVS observes a walking person at three instances


Figure 12: Virtual image sequence (from left to right) for observing a walking person

An important point is that the system can obtain the good performance without referring to the visual feature of the targets, although utilization of features should decrease incorrect matches. Additionally, increasing number of ODVSs employed will increase the robustness of tracking further. Also it will allow more people being tracked in the scene simultaneously. We are planning to implement a system consisting of sixteen ODVSs in the future.

5.3 View synthesis

Figure 11 are the ODIs of the closest camera to the trajectory showing a walking person from points 7 to 9 as indicated in Figure 8. Figure 12 shows the virtual views for observing a walking person from points 1—9. The virtual camera is following the walking person one-meter away from his detected position. The viewing direction is 30 degrees deviated from the walking heading direction. 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 grainy look since the views are "digitally zoomed" in from the original images to create the view from the virtual camera location. Figure 13 shows 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 as well as the viewing direction aligns the best with it. It is easy to observe that the tripod is getting larger as the person walks closer to it as well as the person is turning to the right-hand side.

6 Conclusion

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

Acknowledgements

Mr. Osamu Nishihara (Accowle Co. Ltd., Japan) helped to design the ODVSs and made the prototype. Mr. Ryusuke Sagawa (Tokyo University, Japan) and Mr. Takashi Oya (Canon Inc., Japan) developed earlier version of human tracking system using multiple ODVSs. Prof. Toru Ishida (Kyoto University, Japan) supported the development of the first prototype. The authors thank them for their collaborations. The authors would also like to thank Sony Electronics Corporation, Compaq Computers and the California Digital Media Innovation Initiative (DiMI) for their support.

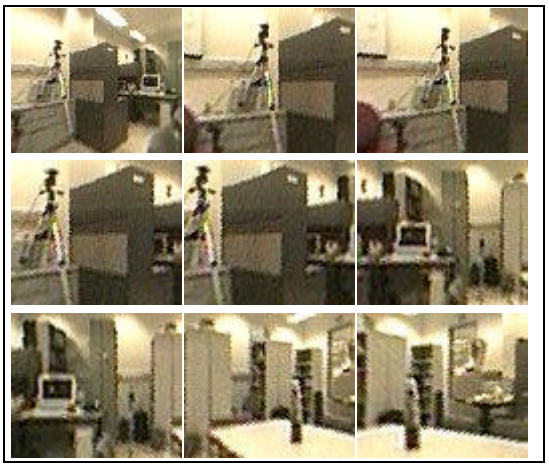

Figure 13: Virtual image sequence (from left to right) of an observer's view

References

[Adelson, 1991] 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.

[Dalmia, 1996] 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. p.43-55.

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

[Hong, 1991] J. Hong, and others, Image-based homing, Proc. Int. Conf. Robotics and Automation, 1991.

[Ishiguro, 1991] H. Ishiguro, M. Yamamoto and S. Tsuji, Omni-directional Stereo, IEEE Trans. PAMI, Vol. 14, No. 2, pp. 257-262, 1992.

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

[Ishiguro, 1998] H. Ishiguro, Development of low-cost and compact omnidirectional vision sensors and their applications, Proc. Int. Conf. Information systems, analysis and synthesis, pp. 433-439, 1998.

[Ishiguro, 1999] H. Ishiguro, M. Barth and T. Sogo, Distributed Vision Systems with Omnidirectional Vision Sensors, submitted to Int. Joint Conf. Artificial Intelligence, 1999.

[Kitamura, 1990] Y. Kitamura and M. Yachida, Three-dimensional data acquisition by trinocular vision, Advanced Robotics, Vol.4, No.1, pp. 29-42, 1990.

[Levoy, 1996] M. Levoy and P. Hanrahan, Light field rendering, Proc. SIGGRAPH, pp. 31-42, 1996.

[Marapane, 1994] S. B. Marapane and M. M. Trivedi, Multi-primitive hierachical (MPH) stereo analysis, IEEE Transactions on PAMI, Vol. 16, No. 3, 1994.

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

[Nayar, 1997] S. K. Nayar and S. Baker, Catadioptiric image formation, Proc. Image Understanding Workshop, pp. 1431-1437, 1997.

[Ng, 1998] 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.

[Rees, 1970] D. W. Rees, Panoramic television viewing system, United States Patent, No. 3, 505, 465, Apr. 1970.

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

[VSAM, 1999] http://www.cs.cmu.edu/~vsam/

[Yagi, 1990] Y. Yagi and S. Kawato, Panoramic scene analysis with conic projection, Proc. IROS, 1990.

[Yamazawa, 1993] K. Yamazawa, Y. Yagi and M. Yachida, Omnidirectional imaging with hyperboloidal projection, Proc. Int. Conf. Robots and Systems, 1993.