Machine Vision and Applications (2003) 14: 94–102
Digital Object Identifier (DOI) 10.1007/s00138-002-0103-0

Machine Vision and Applications

Omnidirectional image-based modeling: three approaches to approximated plenoptic representations

Hiroshi Ishiguro[1], **Kim C. Ng**[2], **Richard Capella**[2], **Mohan M. Trivedi**[2]

[1] Department of Adaptive Machine Systems, Osaka University, Japan
[2] Department of Electrical and Computer Engineering, University of California, San Diego, USA

Abstract. In this paper we present a set of novel methods for image-based modeling using omnidirectional vision sensors. The basic idea is to directly and efficiently acquire *plenoptic representations* by using omnidirectional vision sensors. The three methods, in order of increasing complexity, are *direct memorization, discrete interpolation*, and *smooth interpolation*. Results of these methods are compared visually with ground-truth images taken from a standard camera walking along the same path. The experimental results demonstrate that our methods are successful at generating high-quality virtual images. In particular, the smooth interpolation technique approximates the plenoptic function most closely. A comparative analysis of the computational costs associated with the three methods is also presented.

Keywords: Video array – Real-time tracking – Intelligent room – Omnidirectional camera – Face detection

1 Introduction

Visual modeling, as discussed in this paper, deals with the development of a computer-based representation of the 3D volumetric as well as illumination-cum-reflectance properties of an environment from any desired vantage point. Efficient means for deriving such visual models are necessary for a wide range of applications in virtual/augmented reality, telepresence, or remote surveillance. Automatic means for deriving such models have great demand and potential. However, developing such an efficient algorithm can be difficult, especially if the coverage scene is large and with dynamic objects prevailing (not to mention the additional difficulties introduced if concurrent, multiple novel views are allowed). In this paper, we discuss a set of methods to address such a need.

The existing literature contains two basic types of visual modeling methods that commonly utilize multiple cameras/images. One is an extension of the multiple-camera stereo developed in computer vision, and the other is to approximate *plenoptic function* [1] with densely sampled raw images in computer graphics. The plenoptic function has been proposed as an ideal function representing complete visual information in a 3D space. The approaches developed in this paper represent a hybrid of these two methods based on the unique properties of *omnidirectional images* (ODIs). Our methods do not extract three dimensions for the whole scene; in fact, we extract only a single 3D point lying along the center viewing direction of our desired virtual view plane. The virtual view plane at that 3D point is used to affine transform our selected raw image into the novel view. The selected raw image is chosen based on its correlation of distance and viewing direction to our desired virtual view. The basic ideas are summarized as:

- *Omnidirectional vision sensors* (ODVSs), developed by us, directly approximate the plenoptic representation of the environment.
- 3D geometrical constraints[1] extracted by our modified multiple-camera stereo are used to interpolate novel views between omnidirectional images (where ODVSs do not exist).

In the remaining part of this section, we survey the related works and state our approach. The following sections explain our methods through three modeling steps using ODVSs.

Generally, 3D reconstruction by stereo is not sufficiently stable for practical use. On the other hand, recent progress in vision devices and the related computer interfaces enable us to use many cameras simultaneously. Multiple-camera stereo using many cameras compensates for the problem of matching between images and provides robust and stable range information.

Okutomi and Kanade [2] proposed *multiple-baseline stereo*. The stereo was applied to a virtual-reality system called *virtual dome* [3]. Fifty-six precisely calibrated cameras observe targets inwards from the surroundings and reconstruct 3D models in a small constrained area.

Boyd and colleagues [4] developed a 3D modeling system called *multiple-perspective interactive video*. Multiple-baseline stereo performs template matching on the image plane, while the method of Boyd et al. finds corresponding

Correspondence to: H. Ishiguro
(e-mail: ishiguro@sys.wakayama-u.ac.jp)

[1] This means range information. However, our methods do not directly refer to range information.

Fig. 1. Compact and low-cost ODVS

pixels based on a volume occupancy array that covers the target area. Recently, Seitz and colleagues [5] have improved their method and applied it to precise 3D modeling for small objects by using 16 densely arranged cameras.

The idea of Seitz et al. is closely related to the idea of plenoptic representation. Generally, it is impossible to obtain the complete plenoptic function by using cameras. Several approximation methods have been proposed. The *lumigraph* technique proposed by Gortler [6] and *light-field rendering* proposed by Levoy [7] approximate the 5D plenoptic function with a 4D function.

Visual modeling with many densely sampled images does not require matching among them. However, a very large number of cameras/images is needed. This method is, unfortunately, impractical for applications that cover large areas, especially ones dominated by dynamic objects. The ideas of Boyd and Seitz compensate for the problem to some extent; however, they are still inefficient for covering a wide area. Narayanan's approach using multiple-baseline stereo is one of the practical solutions. However, their purpose is still to reconstruct *complete* 3D models of small 3D spaces and then reproject pixels to a novel view. The method is both computationally and memory intensive. For applications that require wide scene coverage more efficient methods are needed.

Our solution to this problem is to directly acquire the plenoptic representation by using ODVSs, each of which has a wide visual field, and interpolate between them with geometrical constraints given by our modified multiple-camera stereo.

2 ODVS and ODI

ODVS was first proposed by Rees [8] in the patent submitted to the United States government in 1970. Yagi [9], Hong [10], and Yamazawa [11] developed them again in 1990, 1991, and 1993, respectively. Recently, Nayar [12] has geometrically analyzed the complete class of single-lens, single-mirror catadioptric imaging systems and developed an ideal ODVS using a parabola mirror.

However, these researchers were mainly interested in developing the ODVSs as prototypes, and they used them for investigating properties of ODIs taken by them. Therefore, the developed ODVSs were not so compact and the costs were high. In order to develop practical vision systems, we have designed original low-cost and compact ODVSs. The ODVS has a hyperboloidal mirror and a special mechanism to acquire a clear ODI [13]. Figure 1 shows the developed compact

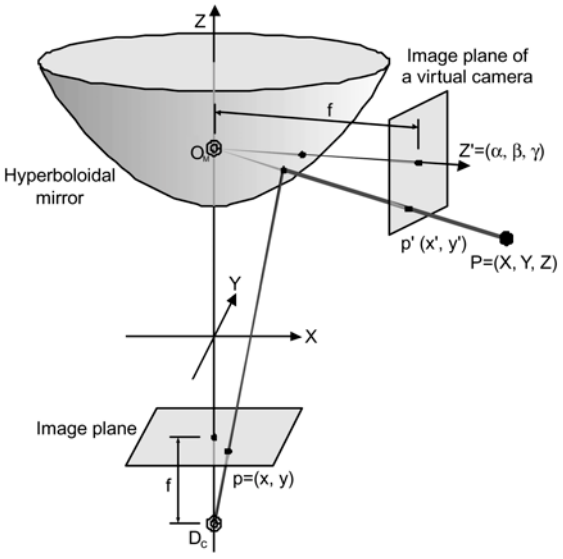

Fig. 2. Transformation into a perspective image

ODVS. The height is about 3.5 in. including a camera unit that provides an NTSC image signal. Figure 4 shows ODIs taken by the ODVS in an indoor environment.

The ODI $p = (x, y)$ is transformed to a perspective image $p' = (x', y')$ by the following equations:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{f}{kR + C} \begin{bmatrix} X \\ Y \end{bmatrix}, \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \cos\gamma & -\sin\alpha\cos\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & \cos\alpha & C \end{bmatrix} \begin{bmatrix} x' \\ -y' \\ f' \end{bmatrix}$$

$$R = \sqrt{X^2 + Y^2}\,, \quad k = (Z - C)/R\,, \quad C = \sqrt{a^2 + b^2}$$

$$R = (CK + b\sqrt{1 + k^2}/(b/a)^2 - k^2) \tag{1}$$

where a and b are parameters of the hyperboloidal surface $R^2/a^2 - Z^2/b^2 = -1$. Figure 2 shows the geometry. This computation does not take much time. To synthesize a perspective image of 200×200 pixels with a Pentium III 450 MHz, the computational time is about 39.6 ms. Furthermore, by replacing local computations with mapping tables it is possible to drastically reduce the computational cost C_{trans}.

3 Modeling method 1: direct memorization

This section and the following two sections describe three methods for visual modeling based on ODIs. Figure 1 shows a hallway that serves as the location of our experimental environment. Figure 5 shows the structure of the sensing environment from a top-down view as well as the desired smooth walk path for view synthesis. Gray circles indicate positions where the ODIs are taken. The interval between the sensors is 2 ft. Our final goal is to synthesize an image sequence seamlessly with a spatiotemporal context preserved along the indicated path from the input of 16 ODIs. Figure 6 shows the discrete paths (with arrowheads) by which our first two methods are used to approximate the smooth walk path. Figure 6a generates views at the sensors' camera center using direct memorization,

Fig. 3. ODIs taken in a hallway

Fig. 4. Omnidirectional image

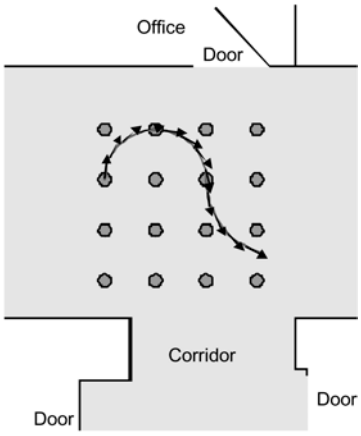

Fig. 5. Positions of ODIs and a desired smooth walk path for view synthesis

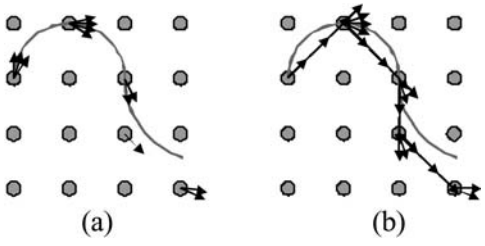

Fig. 6. Discrete paths for view synthesis to approximate the desired smooth walk path

while Fig. 6b allows discrete interpolation between sensors. Method 3 will be the one that provides the closest approximation to the smooth walk path. Figure 3 shows the ODIs that were taken in the hallway, and Fig. 7 shows an image sequence taken by a standard digital camera along the same path in real environments. By comparing with the figure we can evaluate the quality of image sequences synthesized by our developed methods.

The simplest idea of visual modeling using ODIs is to snap ODIs in the environment and transform the ODIs into perspective images. By using the ODIs we can synthesize an image sequence that simulates an observer's view walking through the hallway, as shown in Fig. 8. We select ODIs that have the closest proximity and aligned viewing angle with respect to the intended walk path (Fig. 6a) and then transform the selected ODIs into rectilinear perspective images in the same directions that correspond to the motion directions indicated by the arrows.

The quality of the synthesized images depends on the resolution of the CCD used for the ODVS. In this experimentation, we have used a CCD with a resolution of 410,000 pixels and an image capture board that can take an image of 640×480 pixels. Each of the synthesized color images has 179×126 pixels.

The memory cost is proportional to the number of ODIs. However, the coverage is larger than previous multiple-camera systems using standard rectilinear images with the same number of cameras. Although the required number of ODIs depends on the complexity of the environment and the required quality of synthesized images, we have currently *chosen* to cover the space of $10 \times 10\,\mathrm{ft}^2$ with 16 ODIs. The computational cost is mainly spent on the transformation from ODIs into perspective images. Although the time spent depends on the resolution of the synthesized perspective image, this type of view synthesis is done in real time in our current implementation.

The quality of the synthesized image and the computational cost are comparable to popular virtual-reality systems using Quick Time VR. Furthermore, this method is superior in its flexible control of the viewing direction.

The visual modeling method proposed here shares some common aspects with previous approaches representing visual information based on *plenoptic* functions. The ODI itself can be considered an ideal plenoptic representation of the environment at a 3D point. However, the view sequence generated in this manner is equivalent to jumping from one camera's center to another camera's center during the translation pe-

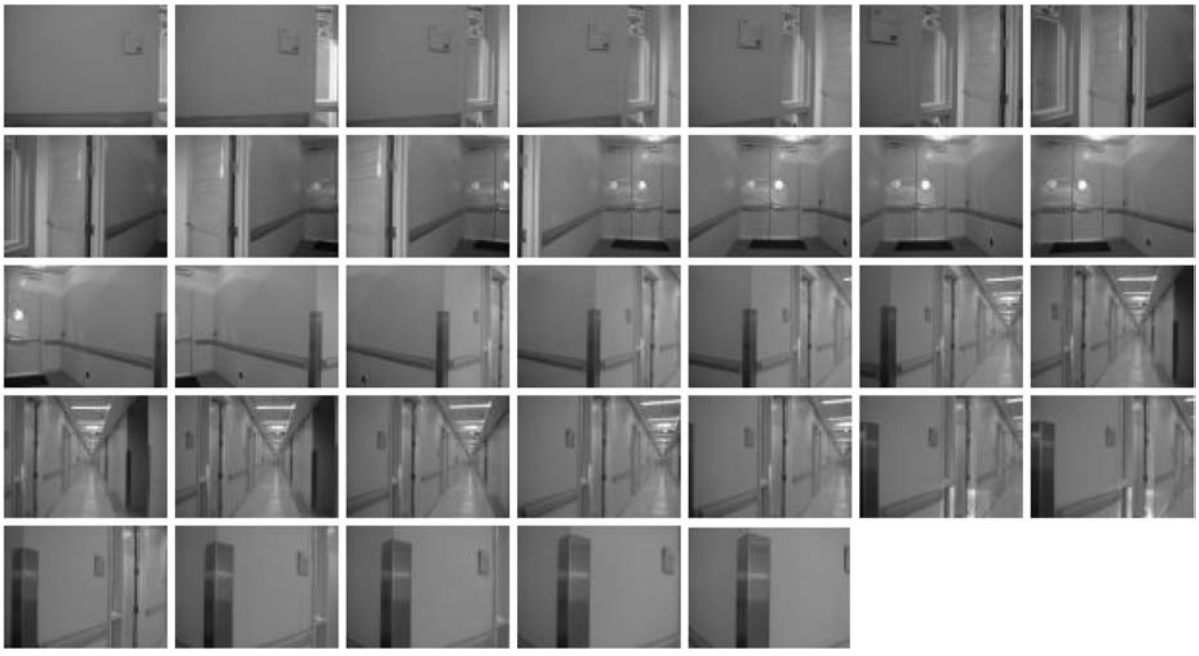

Fig. 7. Image sequence taken by a standard camera along the same desired virtual path

Fig. 8. Synthesized image sequence by the direct method

riod of the walk-through. The spatiotemporal context is lost with the jumps, especially if two cameras are very far apart. The loss of the spatiotemporal context causes visual confusion to the observer and disorientation in the environment results. The following two sections propose more sophisticated methods of acquiring better plenoptic representations for smoother walk-through.

4 Modeling method 2: discrete interpolation

The method described in the previous section needs to densely record ODIs for representing detailed models of the environment since it directly utilizes the ODIs. In order to save the number of ODIs, a mean that interpolates between ODIs is required. A simple idea that requires no expensive memory and computational costs is to estimate changes of visual appearance along lines that connect two arbitrary ODIs. As a parameter to represent the changes, we have employed a zooming ratio between a pair of ODIs that is estimated by a simple template matching.

Figure 11 shows the zooming stereo using a pair of ODIs. By using a simple template-matching method, we can estimate the distance to the objects and the ratio S_b/S_a that represents a means of changing the size of the perspective images to continuously interpolate between the ODIs. Here we perform the template matching with the largest template covering the area of S_b for robust matching. If we cannot perform stable template matching, in other words, if the correlation values do not have a single peak, that means the distance between the pair of ODIs is too long. That is, the zooming stereo suggests

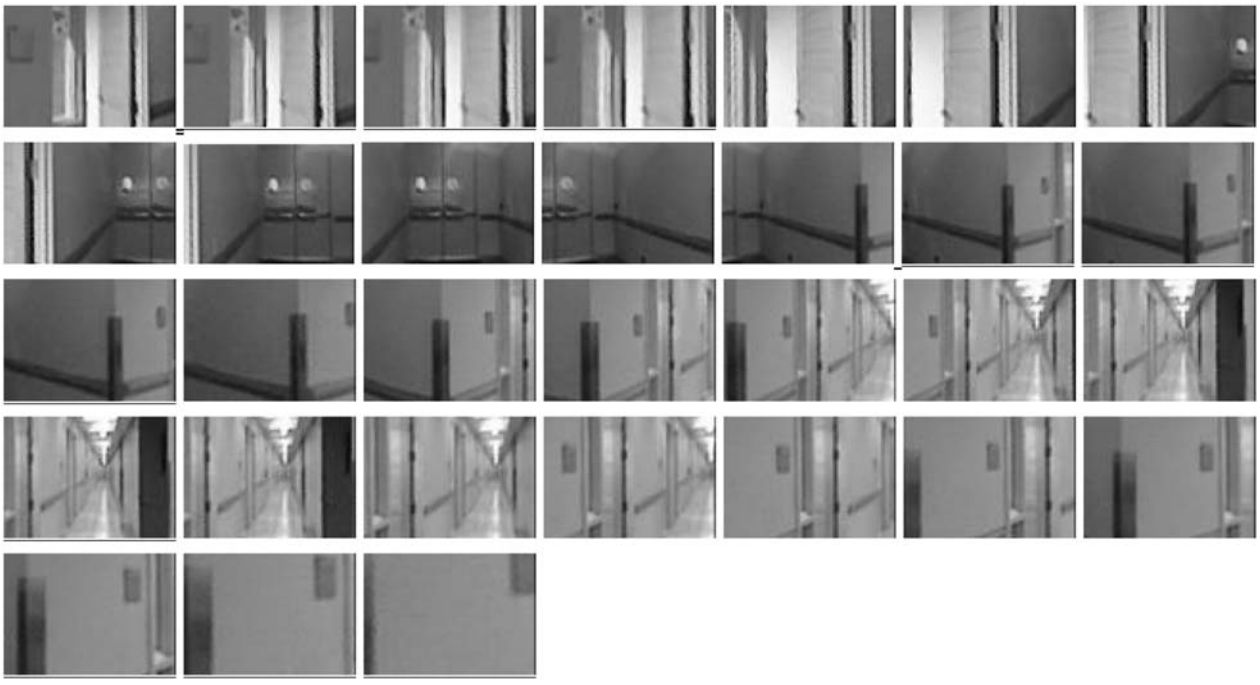

Fig. 9. Synthesized image sequence by the method using zooming stereo

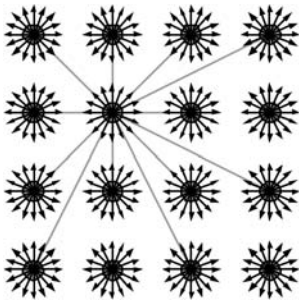

Fig. 10. Plenoptic representation using zooming ratios

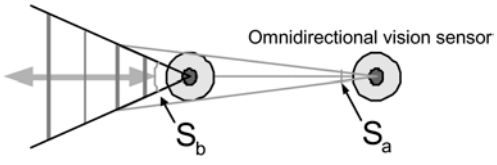

Fig. 11. Zooming stereo

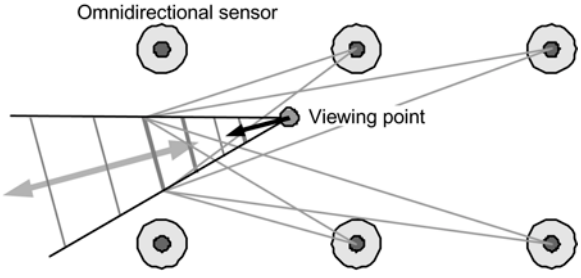

Fig. 12. Multiple-camera stereo

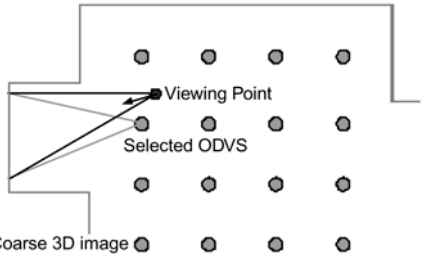

Fig. 13. Plenoptic representation based on coarse structure of the environment

that we take ODIs again for acquiring a better model of the environment.

The computational cost of the zooming stereo depends on the distance D between the ODVSs, the azimuth resolution r of the ODI (viewing field for one pixel), and the size of the synthesized view. It can be roughly estimated as

$$C_{zooming} = C_{\text{correlation}}.S_b/(2 \cdot r)$$

where the computationnal cost $C_{\text{correlation}}$ of correlation is the cost for the template matching between two images and consists of multiplications. Suppose $S_b = 40^\circ$ and $r = 0.2^\circ$. For real-time processing, $C_{\text{correlation}}$ should be less than 0.33 ms. With a 450 MHz Pentium III processor and an image size of 200×200 pixels, the total computational time is about 4.8 ms. If we use a special processor for the correlation computation, the computational speed becomes faster than the video rate.

When there are obstacles between a pair of ODIs, this method cannot be applied. However, in other cases, this method continuously generates views between a pair of ODIs. Furthermore, if several ODIs are located along a line as shown in Fig. 10, the ratio can be applied to all pairs of ODIs along the line.

With the zooming ratios we can acquire a better plenoptic representation than in method 1. Figure 10 depicts the approxi-

mated plenoptic representation with the zooming ratios. In the figure, the arrows indicate zooming ratios assigned to ODIs. The memory cost for this representation is close to the previous one since the additional information is several integers for each ODI; however, it provides a much better virtual view (see Fig. 9). In the figure, the images indicated by the wavy underlines show the smoother interpolation obtained by using the zooming ratios, which method 1 is unable to provide. Nevertheless, the method here still does not provide us with the freedom of arbitrary walk-through. In other words, the walk-through is very much constrained in a linear fashion on a straight line that connects two ODVSs, as shown in Fig. 5b. For an arbitrary, smooth walk-through we introduce the third method – smooth interpolation using multiple-camera stereo.

5 Modeling method 3: smooth interpolation

An ideal interpolation is to synthesize images at arbitrary viewpoints without any constraints. Such smooth interpolation between ODIs must estimate the local environment structure and select the best ODVSs for the virtual view synthesis based on it. We have employed a multiple-camera stereo method and modified it for realizing robust range estimation.

The basic idea is similar to multiple-baseline stereo [2] by exploiting the advantages of having redundant information, but we have modified it for general camera arrangements so as to allow an arbitrary search path. Our original findings are listed below:

- *Range-space search* method for finding corresponding points that accept arbitrary camera positions and arbitrary search path
- Camera selection based on robust statistics
- Utilization of large templates for robust template 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 performing the search in the disparity space, especially if the cameras are not closely positioned and their field of view is large (such as that of the ODI). In range-space search, positions of templates for finding corresponding points on camera images are given as projections of 3D points along a virtual 3D search line onto the camera images. Figure 12 shows a range-space search performed among multiple ODVSs. Here the search intervals along the virtual 3D line are determined based on pixel resolution of the nearest ODVS so as to preserve the maximum range resolution. Note that pixels in the ODI cover the 3D space with different resolutions.

Furthermore, for finding the best match we have modified the computation of template matching for multiple-camera stereo. Suppose three ODVSs observe a common object and another ODVS observes an obstacle located in front of the ODVS. If we perform standard template matching among those four ODVSs, the result will suffer from significant errors. Our idea is to minimize the error by choosing the ones best suited among these ODVSs employing the technique in robust statistics. The template size of $n \times m$ on ODIs can be represented as points in an $n \times m$-dimensional parameter space. In this parameter space, we perform outlier detection using standard deviation [14] and reject ODVSs at the extreme values. This method dynamically selects ODVSs that provide the best match and avoids the occlusion problem that easily occurs, especially when the ODVSs are placed at arbitrary locations.

Our purpose is not to acquire precise range information in this work but to robustly provide smooth image sequences of virtual views. Large templates having the same resolution as the desired virtual image are used for robust matching. This range-space search accommodates the window distortion that a larger baseline stereo inherits, which the simple rectangular template in the disparity space cannot easily solve.

Figure 15 shows a part of the acquired range data by the multiple-camera stereo for the virtual-view synthesis along the desired path indicated in Fig. 4. The range data are not accurate, especially in the upper left side of the figure, since there is no texture on the wall. However, this does not seriously influence the view synthesis. Based on the coarse range data, we have selected the camera that best simulates the virtual view at a virtual observation point. Here the template used for the template matching is directly referred to as the virtual view. Figure 13 shows the plenoptic representation based on the extracted coarse structure of the environment. Figure 14 shows the virtual image sequence synthesized along the smooth path. By comparing this with Fig. 9, we can find that the image sequence is smoother, especially in the images indicated by the wavy underlines that do not exist in the image sequence of Fig. 9. Furthermore, we have compared it with the image sequence taken by a standard camera along the same path in the real environment (Fig. 7). By comparing their image sequences, we find they closely resemble each other.

The computational cost is, however, relatively high. If we use ODIs for the multiple-camera stereo, the computational cost is about $M \times C_{\mathrm{correlation}}$. Furthermore, it takes a cost $C_{\mathrm{statistic}}$ for the statistical analysis. For real-time computation it requires parallel computers.

More practical implementation is to give the range information or compute it in advance of the view synthesis. By performing the multiple-camera stereo prior to reconstructing the environmental structure, the system can synthesize the virtual image sequence relatively efficiently.

6 Concluding remarks

Another important issue for the modeling methods discussed in this paper is camera localization. To date several methods have been proposed. If the application does not require precise camera positions, the method proposed by Ishiguro and colleagues [15] is the most flexible one. The method decomposes an ODI into phase and magnitude components that represent the direction of the reference azimuth axis against the environment structure and visual uniqueness of the observation point, respectively. By referring to both the phase and magnitude components, we can estimate topologically correct spatial positions of the observation points. The second method is to utilize a mobile platform for taking ODIs at various observation points. In this case, the camera position is given with the precision of the mobile platform control. Use of the mobile platform is convenient for taking a large number of ODIs

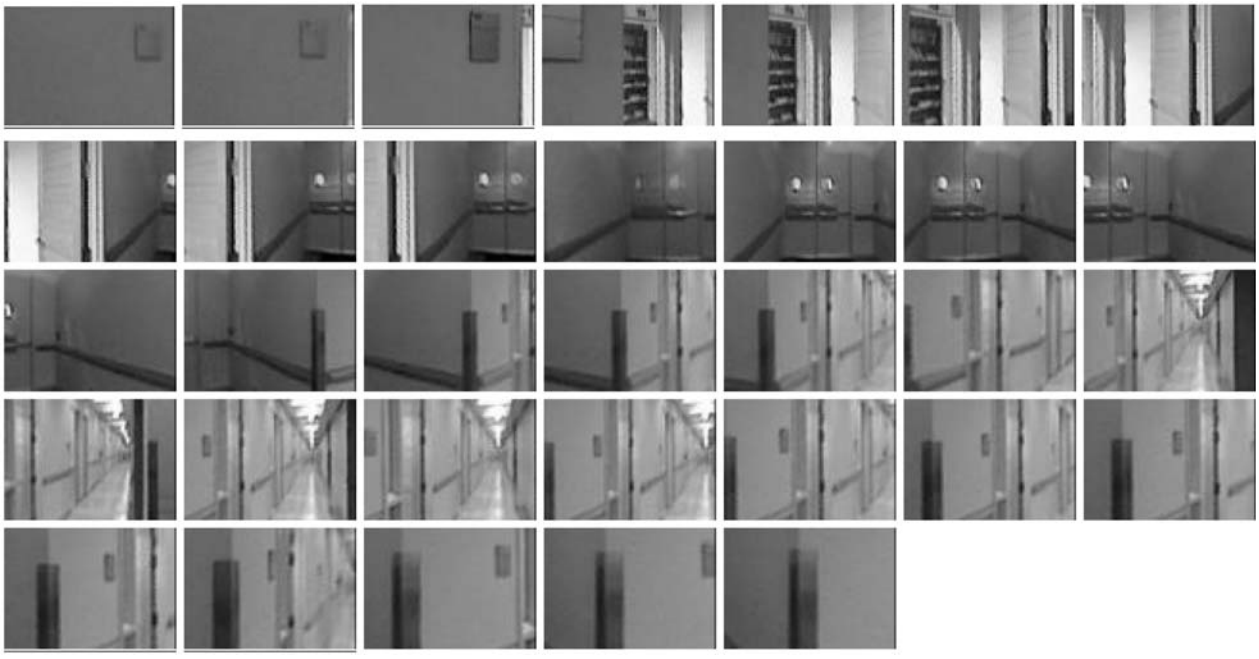

Fig. 14. Synthesized image sequence using multiple ODIs

Table 1. Comparison of three proposed methods for image-based modeling

	Computational cost	Memory cost	Applications
Direct method	C_{trans}	N*ODI	Virtual view synthesis for simple environments
			Dynamic monitoring using multiple ODVSs
Discrete interpolation	C_{trans} +$C_{\mathrm{correlation}}$	N*ODI+ N*40∼80Byte	Virtual view synthesis with a smaller number of ODVSs
			Real-time virtual view synthesis using multiple ODVSs
Smooth interpolation complex environments	C_{trans} +M*$C_{\mathrm{correlation}}$	N*ODI+ (3D coarse structure)	Virtual view synthesis for
	+$C_{statistic}$		Real time virtual view synthesis using parallel computers for the multiple-camera stereo

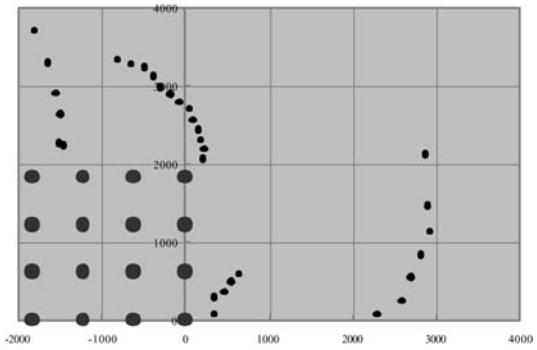

Fig. 15. Range data acquired by multiple-camera stereo

in a wide area; however, it requires accurate internal motion sensors. Generally, external sensor systems provide relatively better accuracy for the localization. The third method is to use such external sensors.

Our idea for camera localization is to use multiple ODVSs. ODVSs are also useful for locating moving objects since the identification between the sensors and calibration of the sensor positions is easy and robust [16]. The system using multiple ODVSs can perform multiple-camera stereo for localizing the ODVS used for taking ODIs.

The visual modeling methods using ODIs do not require much memory and computational cost. As discussed in this paper, only 16 ODIs cover a space of approximately $10 \times 10\ ft^2$ and generate virtual views taken at arbitrary viewpoints. The resolution of synthesized images is not very high; however, the real-time synthesis compensates for the demerits. When we watch a real-time video stream, we can obtain enough information even if the resolution is poor. The computational and memory costs of the proposed three methods are summarized in Table 1.

Based on these methods we can consider practical applications. First, the methods change virtual-reality systems on Web pages. Previous virtual-reality systems using real images needed to download a very large number of images, and they generally do not allow us to change the viewing direction. With VRML, 3D data with one color per pixel demand tremendous network bandwidth. Our methods do not have such serious demerits. Second, we can consider combining both ODIs and rectilinear images taken by a standard camera for applications that require high-resolution images. The low-resolution ODI's virtual image is smoothly "dissolved" into the standard rectilinear image for detailed observation. Let us consider developing a virtual museum. View syntheses using ODIs realize real-time walk-through, and high-resolution images allow us to appreciate works of art. Other applications may be found in such fields as entertainment, remote surveillance, and collaborative environments. Challenging issues in realizing such applications in dynamic environments with multiple concurrent viewers were discussed in Ng et al. [17, 18].

References

1. Adelson EH, Bergen EH (1991) The plenoptic function and the elements of early vision. In: Landy M, Movshon JA (eds) Computational models of visual processing. MIT Press, Cambridge, MA
2. Okutomi M, Kanade T (1993) A multiple baseline stereo. IEEE Transactions on pattern analysis and machine intelligence, 15(4):353–363
3. Narayanan PJ, Rander PW, Kanade T (1998) Constructing virtual worlds using dense stereo, Proceedings of ICCV, Bombay, India, pp 3–10
4. Boyd J, Hunter E, Kelly P, Tai L, Phillips C, Jain R (1998) MPI-Video infrastructure for dynamic environments. Proceedings of IEEE international conference on multimedia computing and systems, Austin, TX, p 249
5. Seitz SM, Kutulakos KN (1998) Plenoptic image editing. Proceedings of ICCV, Bombay, India, pp 17–24
6. Gortler SJ, Grzeszczuk R, Szeliski R, Cohen MF (1996) The lumigraph. Proceedings of SIGGRAPH, New Orleans, USA, pp 43–54
7. Levoy M, Hanrahan P (1996) Light field rendering, Proceedings of SIGGRAPH, New Orleans, USA, pp 31–42
8. Rees DW (1970) Panoramic television viewing system. US Patent No. 3,505,465
9. Yagi Y, Kawato S (1990) Panoramic scene analysis with conic projection. Proceedings of IROS, Tsuchiura, Japan
10. Hong J et al (1991) Image-based homing. Proceedings of ICRA, Sacramento, USA, pp 620-625
11. Yamazawa K, Yagi Y, Yachida M (1993) Omnidirectional imaging with hyperboloidal projection. Proceedings of IROS, Yokohama, Japan
12. Nayar SK, Baker S (1997) Catadioptiric image formation. Proceedings of the image understanding workshop, New Orleans, USA, pp 1431–1437
13. Ishiguro H (1998) Development of low-cost and compact omnidirectional vision sensors and their applications, Proceedings of the international conference on information systems, analysis and synthesis, Orland, USA, pp 433–439
14. Davies OL, Goldsmith PL (1972) Statistical methods in research and production. Imperial Chemical Industries Limited
15. Ishiguro H, Tsuji S (1996) Image-based memory of environment. Proceedings IROS, Osaka, Japan, pp 634–639
16. Kato K, Ishiguro H, Barth M (1999) Identifying and localizing robots in a mult-robot system environment, Proceeding in IROS, Busan, Korea, pp 966–972
17. Ng KC, Trivedi MM, Ishiguro H (1998) 3D ranging and virtual view generation using omni-view cameras. Proceedings of the multimedia systems and applications, Boston, USA, SPIE, vol. 3528,
18. Ng K, Ishiguro H, Trivedi M, Sogo T (1999) Monitoring dynamically changing environments by ubiquitous vision system. IEEE International workshop on visual surveillance, Fort Collins, USA

Hiroshi Ishiguro received the Doctor of Engineering degree from Osaka University, Japan in 1991. In that same year, he started working as a research assistant at the Department of Electrical Engineering and Computer Science, Yamanashi University, Japan. In 1992 he joined the Department of Systems Engineering, Osaka University, Japan as a research assistant. In 1994, he was an associate professor in the Department of Information Science, Kyoto University, Japan and began research on distributed vision using omnidirectional cameras. From 1998 to 1999, he was a visiting scholar in the Department of Electrical and Computer Engineering at the University of California at San Diego. In 1999, he developed interactive humanoid robots, Robovie in ATR, Japan. In 2001, he was a professor in the Department of Computer and Communication Sciences, Wakayama University, Japan. He is currently a professor in the Department of Adaptive Machine Systems, Osaka University, Japan and a visiting group leader in ATR Intelligent Robotics and Communication Laboratories, Japan.