Appearance sphere: background model for pan-tilt-zoom camera

Toshikazu Wada
Okayama University

Takashi Matsuyama
Kyoto University

Appearance Sphere:
Background Model for Pan-Tilt-Zoom Camera

Toshikazu WADA
Department of Information Technology,
Okayama University
Tsushima Naka, Okayama 700, JAPAN

Takashi MATSUYAMA
Department of Electronics and Communication,
Kyoto University
Sakyo, Kyoto 606-01, JAPAN

Abstract

Background subtraction is a simple and effective method to detect anomalous regions in images. In spite of the effectiveness, it cannot be used with an active(moving) camera head, because the background image varies with camera-parameter control. This paper presents a background subtraction method with pan-tilt-zoom control. The proposed method consists of an omnidirectional background model called appearance sphere *and* parallax free sensing. *Based on this model, precise background images can be generated and background subtraction can be performed for any combinations of pan-tilt-zoom parameters without restoring 3D scene information.*

1. Introduction

Background subtraction is a simple and effective method to detect anomalous image regions deviating from the known background image. For this method to be effective, however, the following conditions must be satisfied:

1. Background scene is stationary.
2. Camera parameters are fixed.

These conditions limit the utilities of this method.

In the real world, the stationarity assumption is violated by the following factors:

- Variations of objects: fluttering leaves and flags by the wind in outdoor scenes and the flicker of CRT displays in indoor scenes.
- Variations of the illuminations: sunlight fluctuation caused by the sun and cloud movements in outdoor scenes and the room-light variation in indoor scenes.

Background subtraction is affected by these variations. To cancel the variations of objects, we have to augment the background subtraction so that the known variations are regarded as normal. To cancel the variations of illuminations, background image must be renewed according to the input images.

As for the camera parameter fixation assumption, it disables active sensing. In most of active-vision systems, camera parameters are dynamically controlled. This causes the following types of image variations:

- Geometric variation caused by changing the camera location, view direction(pan, tilt) and zoom.
- Photometric variation caused by changing focus[1], iris and shutter speed.

[1]Geometric variation caused by changing the focus parameter can easily be calibrated and is negligible in the telecentric lens. Hence, here we simply describe that focus control causes the photometric variation.

In this paper, we address the problem of background subtraction with camera-parameter control. To realize a background subtraction with camera-parameter control, we have to generate precise background images for any combinations of camera parameters.

In general, complete 3D scene information (i.e., depth, optical properties of objects.....etc.) or whole scene-appearance information (i.e., observed images for all possible parameter combinations) is necessary to generate precise images for any combinations of camera parameters. But, those algorithms restoring 3D scene information from images do not work for general scenes. Also, whole scene appearance information cannot be observed and stored, because it consists of a large number of images.

Many works related to generating images of arbitrary camera parameters has been proposed in Robot Vision[1], Computer Graphics and Virtual Reality[2]~[7], but there is no method to generate precise images enough for background subtraction.

In this paper, it is shown that if the observed images do not involve motion parallax, image variations caused by changing pan-tilt-zoom parameters can exactly be simulated from a limited number of images without restoring 3D scene information. In our method, precise background images for arbitrary pan-tilt-zoom parameters are generated from an omnidirectional image model called *appearance sphere* which consists of a limited number of images taken by *parallax free sensing*.

Our method enables not only the background subtraction but also many other vision tasks with camera-parameter controls: egomotion analysis, target tracking, omnidirectional stereo, ... etc. . Hence, it is considered as a basic technology in Active Vision.

Detailed method and the experimental results are described in the following sections.

2. Appearance sphere

In practical lens systems, those rays from objects which form a projected image always pass through the front and rear *nodal points*[2]. For the approximation of the optical projection by the central projection, it is assumed that the front nodal point is the projection center and the images are projected to the imaginary screen in front of the lens as shown in Figure 1. Hereafter, we simply call imaginary screen "screen".

If the images on different screens are observed with a fixed front nodal point, they have no parallax. We

[2]Principal points[8] and nodal points are equivalent if the refractive indices of front and rear media are the same.

1015-4651/96 $5.00 © 1996 IEEE
Proceedings of ICPR '96

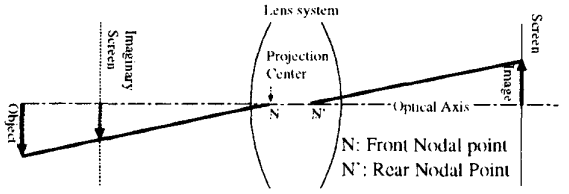


Figure 1. Nodal points and central projection.

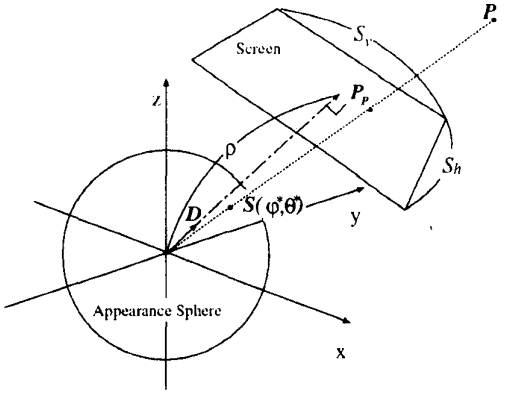


Figure 2. Projections to a screen and appearance sphere

call such images *parallax free images*. Sensing method to obtain parallax free images is described in Section 3.

Parallax free images can be projected to a virtual screen by the central projection. If the projection center and the front nodal point are coincide, the projected images are merged without any inconsistencies and form a seamless image on the screen.

For the omnidirectional background image description, the virtual screen must be a closed surface. Appearance sphere is the projected image on a spherical screen[3]. Re-projecting the image on the sphere to arbitrary planar screens, we can generate background images of arbitrary directions and zoom as described below:

A planar screen having a unit normal vector $D = (k, l, m)$ is represented as

$$kx + ly + mz = \rho. \quad (1)$$

where ρ represents the distance from the origin. D and ρ correspond to the view direction and the zoom, respectively.

To approximate the optical projection from 3D point to the screen, we use the central projection whose center (front nodal point) is on the coordinate origin. By this projection, a 3D point P is projected to P_p on the screen:

$$P_p = \frac{\rho}{D^T P} P. \quad (2)$$

Here we suppose a virtual spherical screen:

$$S(\varphi, \theta) = r(\cos\varphi\cos\theta, \cos\varphi\sin\theta, \sin\varphi). \quad (3)$$

where r represents the radius of the sphere, and $-\pi/2 \leq \varphi \leq \pi/2$, $0 \leq \theta < 2\pi$.

By the central projection whose center is located on the origin, a 3D point $P = (x, y, z)$ is projected to a point

[3]The shape of the virtual screen do not have to be a sphere, but a star-shaped surface.

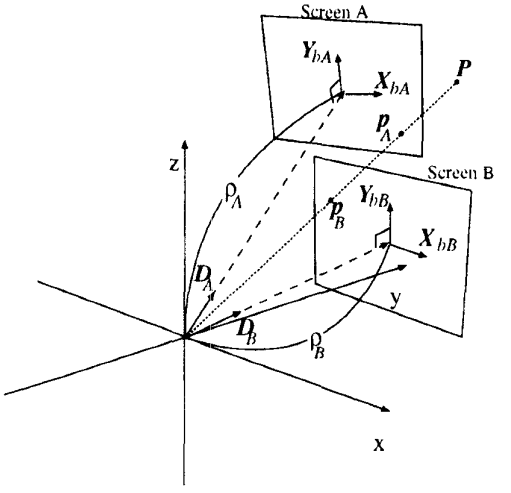


Figure 3. Projection to different screens

$S(\varphi^*, \theta^*)$ on the spherical screen:

$$S(\varphi^*, \theta^*) = \frac{r}{||P||} P, \quad (4)$$

where $\varphi^* = \cos^{-1}\frac{z}{\sqrt{x^2+y^2+z^2}}$ and $\theta^* = \cos^{-1}\frac{x}{\sqrt{x^2+y^2}}$.

A 3D point P is projected to P_p on the planar screen by Equation (2), which can be re-projected to the following point on the spherical screen by Equation (4):

$$\frac{r}{||P_p||} P_p = \frac{r}{||P||} P. \quad (5)$$

This means that as long as the front nodal point and the spherical projection center are coincide, re-projected point on the sphere is computed independent of the intermediate screens (Figure 2). Hence, we can construct a unified omnidirectional background model from a limited number of images.

As for the image generation, a point P on the sphere can be re-projected to a point P_p on the planar screen, where P_p is represented by Equation (2). By substituting P in Equation (2) by the projected point on the sphere $(r/||P||)\,P$, we obtain

$$\frac{\rho}{D^T \frac{r}{||P||} P} \frac{r}{||P||} P = \frac{\rho}{D^T P} P. \quad (6)$$

This means that the projected points on a planar screen from a 3D point and a point on the sphere are coincide. Hence, the images on any screens can be generated from the appearance sphere.

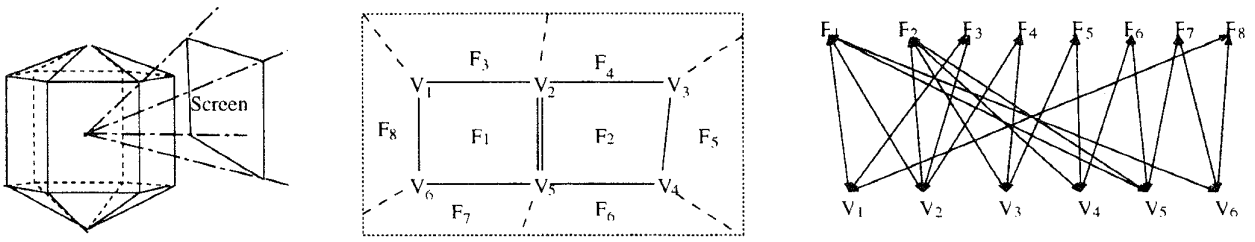
The radius of the appearance sphere don't have to be constant. That is, its shape can be generalized as

$$S(\varphi, \theta) = r(\varphi, \theta)(\cos\varphi\cos\theta, \cos\varphi\sin\theta, \sin\varphi), \quad (7)$$

where $r(\varphi, \theta) > 0$. In this case, since the projected points on a screen from a 3D point P and a point on the generalized appearance sphere are coincide, images on any screens can be generated based on this model.

3. Parallax free sensing

Parallax free sensing is realized by locating the front nodal point on the pan and tilt axes. Under this configuration, the front nodal point does not move by the camera rotation. The zoom control may shift the front nodal point from the rotational center. In this case, we can also adjust the front nodal point to the rotational center by sliding the camera body. A method of the calibration is described in Section 5.1.



(a) Polyhedral model (b) Polyhedral model projected on a screen (c) Logical data structure

Figure 4. Data structure of the polyhedral model

The screen motion with fixed front nodal point will not cause motion parallax. Under this configuration, images projected to different screens are mutually transformed by a pair of 2D geometric transformations independent of the scene depth as described below:

A 3D point $P = (x_p, y_p, z_p)$ on a screen has a 2D screen-coordinate value $p = (x_s, y_s)$:

$$p = \begin{pmatrix} X_b^T \\ Y_b^T \end{pmatrix} P \tag{8}$$

where X_b and Y_b are the 3D orthogonal basis vectors of the screen-coordinate system satisfying $X_b^T Y_b = Y_b^T X_b = 0$, $X_b^T D = Y_b^T D = 0$ and $||X_b|| = ||Y_b|| = 1$. These vectors are represented by

$$X_b = \frac{1}{\sqrt{1-m^2}} \begin{pmatrix} l \\ -k \\ 0 \end{pmatrix}, \quad Y_b = \frac{m}{\sqrt{1-m^2}} \begin{pmatrix} -k \\ -l \\ \frac{1-m^2}{m} \end{pmatrix}. \tag{9}$$

Inversely, a 2D point $p = (x_s, y_s)$ in the screen-coordinate system has a 3D world-coordinate value $P = (x_p, y_p, z_p)$:

$$P = \begin{pmatrix} X_b & Y_b \end{pmatrix} p + \rho D. \tag{10}$$

When a 3D point P is projected to p_A on screen A and p_B on screen B (Figure 3), p_B is represented by the following equation from Equation (2),(8) and (10):

$$p_B = \frac{\rho_B \begin{pmatrix} X_{bB}^T \\ Y_{bB}^T \end{pmatrix} \left\{ \begin{pmatrix} X_{bA} & Y_{bA} \end{pmatrix} p_A + \rho_A D_A \right\}}{D_B^T \left\{ \begin{pmatrix} X_{bA} & Y_{bA} \end{pmatrix} p_A + \rho_A D_A \right\}}, \tag{11}$$

where screen A and B have unit normal vectors D_A and D_B, distances ρ_A and ρ_B, and basis vectors (X_{bA}, Y_{bA}) and (X_{bB}, Y_{bB}) respectively.

From Equation (11), it is confirmed that the relationship between the projected points from a 3D point P on different screens can be represented by 2D transformation independent of the 3D point P.

4. Data structure and algorithm

In the case of analog appearance sphere, observed images are projected onto a sphere. But, in practice, since digital images consist of pixels, pixels on the sphere should be uniformly distributed and have the same topology as that of square grid so that the interpolation can be performed on it. Unfortunately, we cannot define such sampling points on the sphere.

To construct the generalized model without interpolation, we employ a polyhedral appearance model[4] whose facets are the images taken by the parallax free sensing (Figure 4(a)).

The polyhedral appearance model consists of vertices and facets. These objects must have the following attributes:

[4] Polyhedral model is included in the generalized appearance sphere in Equation (7).

vertex: 3D location (x, y, z).

facet: Normal vector $\boldsymbol{D}$, distance ρ and the image I.

The visible vertices projected inside the image frame can easily be extracted from the model (Figure 4(b)). To determine the visible facets corresponding to the visible vertices, each vertex should have links to neighboring facets. Also, to determine the visible area of each facet, each facet should have the links to the vertices on its corners. Hence, the set of vertices and facets form a bidirectional bipartite graph as shown in Figure 4(c).

Background images are generated by the following algorithm:

1. Extract the visible vertices.
2. For each extracted vertex, extract neighboring facets.
3. For each extracted facet, compute the visible area and compute the pixel values in the area by Equation (11).

5. Experimental results

Here we show experimental results of rotational center calibration, background image generation and anomalous region detection using SONY 3-CCD camera DXC-325 with Canon zoom lens VCL-810BX.

5.1. Calibration of the rotational axis

The rotational axes location can be calibrated to be just on the front nodal point based on the following properties (Figure 5(a)):

- If a beam light passes the front nodal point of the camera, any points on the beam are projected to the same point on the screen.
- If the front nodal point is just on the rotational axes, a beam light passing the front nodal point always passes the point while rotating the camera.

For the calibration, we use a linear stage mounted on a pan stage and a laser beam oscillator (Figure 6), and the calibration is done by the following manner (Figure 7):

Step1 Set a laser beam so that it passes the front nodal point of the camera.

A bright spot appears on a translucent screen by the laser beam as shown in Figure 5. If the beam passes the front nodal point, two spots on translucent screens of different depth are projected to the same point on the image plane.

Step2 Rotate the stage to the left side, and measure the distance between beam spots in the observed images sliding the linear stage. The same measurement is performed for the same angle to the right side.

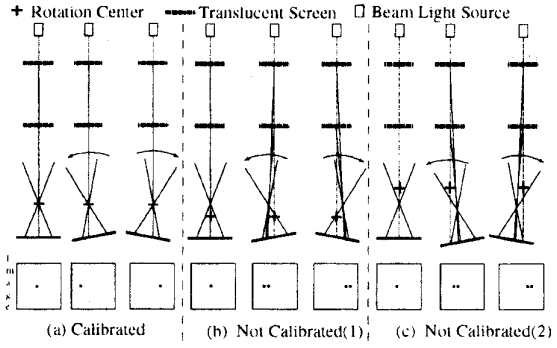


Figure 5. Calibration scheme using a beam light

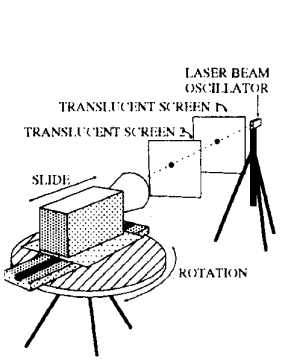


Figure 6. Settings for calibration

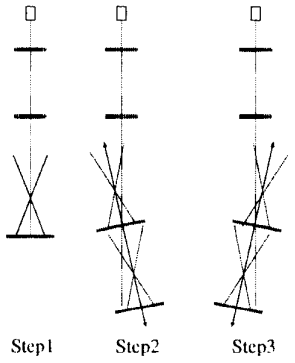


Figure 7. Calibration of rotational axis

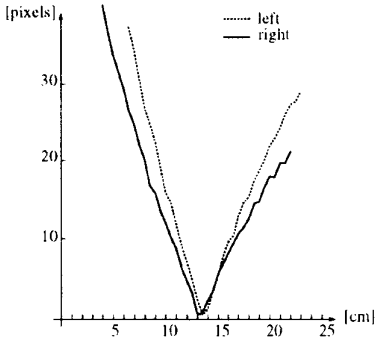


Figure 8. Distance between projected beam spots.

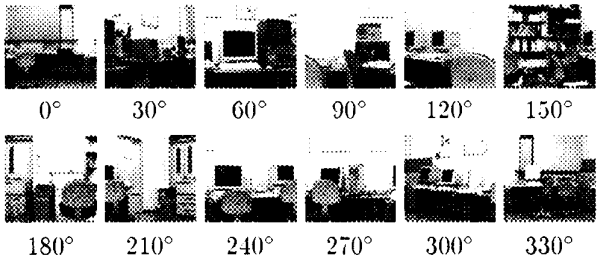


Figure 9. Observed images.

Step3 The stage location which minimizes the sum of the distances is optimal.

Figure 8 shows the graph of this calibration result for rotational angle 12°. The horizontal and vertical axes represent the linear stage location and the distance between the projected spots respectively. From this figure, the optimal stage location is determined as 13.5[cm]. With this stage location, horizontal view angle of the camera is determined by measuring the maximum rotation angle which keeps an object in the frame. The horizontal view angle measures 34°.

5.2. Image generation and region detection

To construct the polyhedral appearance sphere, 12 images are observed by panning the camera clockwise from a fixed view point with rotational angle interval about 30° (Figure 9). Each image size is 512 × 480.

The 1st order radial lens distortion coefficient of the observed image is computed by Tsai's algorithm[9], and the image distortion of the observed images are restored based on this coefficient. The crossing line of the observed images are determined by the template matching and the images are pasted into a polyhedral cylinder (Figure 10).

5.2.1 Calibration of model parameters

In this process, diameter of the model, image center and rotation angles of observed images are calibrated so that sum of the rotation angle is equal to 360° and the total matching score is maximized. In this experiment, the diameter is determined as 834.416 pixels. The difference from the predetermined diameter 837.338 is less than 3 pixels. This means the preciseness of the rotational center calibration.

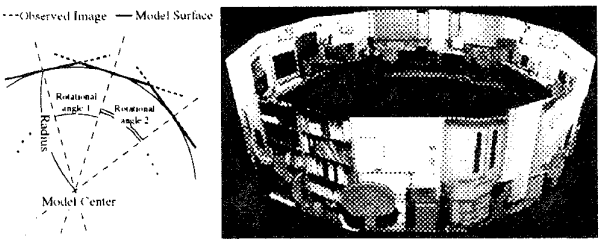
5.2.2 Calibration of observed image

Figure 11(a) shows an observed image at approximately 105°, (b) is the generated image at 105°, and (c) is the difference between (a) and (b). This result shows that the pan angle 105° is not accurate as the model parameter.

To determine the correct angle, mean errors of the pixel values between the observed and the generated images are computed by changing the angle parameter. Figure 11 (d) shows the graph of the mean error. From this graph, the optimal angle is determined as 104.64°. Generated image at this angle is shown in (e), and the difference between (a) and (e) is shown in (f). The mean error and the standard deviation of pixel value are 2.9 and 6.28 respectively. From this result, we can notice that the accurate angle of the observed image is required to generate accurate image. While the accurate angle can be found by minimizing the error, this procedure is computationally expensive. Fortunately, since a small angle of rotation can be approximated by the translation, it is enough to find the optimal horizontal translation minimizing the error.

5.2.3 Anomalous region detection

Figure 12 shows the region detection result. Figure 12 (a) is the observed image at approximately 70° with an unknown object, (b) is the generated image at at this angle, and by translating the generated image by 5 pixels left side, the region detection result (c) is obtained. The translation is determined by minimizing the difference between the generated and observed image. In this experiment, to neglect the shadowed regions by the object, regions are detected by thresholding the cosine of RGB-color vector angles between the images for each pixel. The threshold levelis 0.98.



(a) (b)

Figure 10. Polyhedral appearance model

6. Discussions

From the experimental results, it is noticed that many types of calibrations are necessary for our task. This distinguishes computer-vision tasks from the virtual-reality and imaging tasks[2] $\sim$ [7].

These calibration is classified into three types: 1) physical calibration (rotational axis, view angle), 2) model calibration (diameter,image center..... etc.), and 3) observed image calibration (rotation angle). The most severe calibration is 3), because in most of the cases, the distance to the screen is much longer than the pixel size, and hence a small angular error produces a crucial geometric error of the generated image. But, a small angular error can be approximated by translation, such error can be easily recovered.

The angle resolution $\delta(x)$ corresponding to the pixel resolution at x on the screen is represented as:

$$\delta(x) = \left| \tan^{-1}\left(\frac{x+0.5}{R}\right) - \tan^{-1}\left(\frac{x-0.5}{R}\right) \right|. \quad (12)$$

where R represents the distance to the screen. The angular error ω which produces 1-pixel image distortion satisfies the following equation:

$$R \times \left| tan\left(\frac{\theta}{2}\right) - \tan\left(\frac{\theta}{2} - \omega\right) - 2\tan\left(\frac{\omega}{2}\right) \right| = 1, \quad (13)$$

where θ represents the view angle of the camera.

In our experiment with fixed zoom, $\delta(0) = 0.069°$ at the image center, $\delta(256) = 0.062°$ at the image frame, and $\omega = \pm 0.38°$ which corresponds to ± 5.574-pixel translation. Rotational stages having small angle step less than 0.06° are commonly used. Also, if the angular error less than 0.38° presents, the optimal translation can be found without generating many images

7. Conclusions

We proposed a background subtraction method for pan-tilt-zoom cameras. This method consists of parallax free sensing and omnidirectional background model called appearance sphere. The most distinguishing property of our method is that no 3D scene information is necessary, i.e., the result is not affected by the geometric and photometric properties of the scene.

The essential idea of the image generation itself happens to be equivalent to that of [3] $\sim$ [6] in the field of Computer Graphics and Virtual Reality. But, to generate precise images enough for Active Vision tasks analyzing real-world scene, parallax free sensing and some other calibration techniques described in this paper are necessary. That is, only the total technology of our method enables real-image analysis.

Since our method enables not only the background subtraction but also many other vision tasks with sensor-parameter control, this method is considered as a platform of Active Vision. Such tasks, egomotion analysis, target tracking, omnidirectional stereo, ... etc., can be realized based on our method. This will be done in the future works.

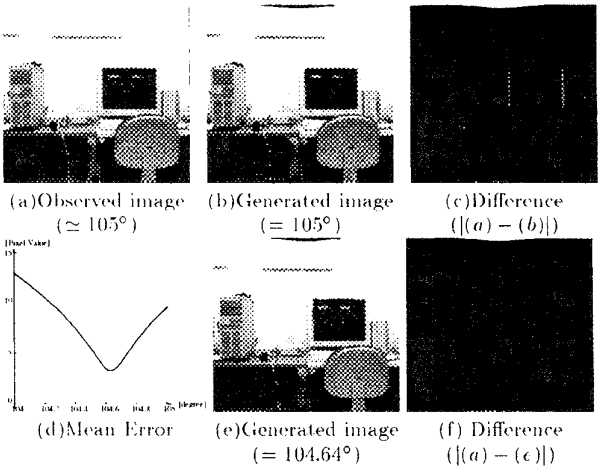
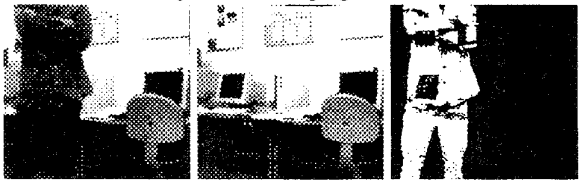


Figure 11. Image generation



(a) (b) (c)

Figure 12. Region detection result, (a):Observed Image, (b): Generated image , (c): Detected region.

Acknowledgment

Authors appreciate assistance of experiments by Mr. Norimichi Ukida and Mr. Katsuyuki Miyoshi in Okayama University.

References

[1] Yamazawa K., Yagi Y. and Yachida M., "Obstacle Detection with Omnidirectional Image Sensor HyperOmni Vision",**ICRA**, pp.1062 - 1067, Nagoya (1995)

[2] Lippman A., "Movie-Maps: An Application of the Optical Videodisc to Computer Graphics", **SIGGRAPH'80**, (1980)

[3] Blinn, J.F. and Newell, "Texture and Reflection in Computer Generated Images", **Comm. of the ACM**, 19(10), pp. 542-547, (1976)

[4] Hall R., "Hybrid Techniques for Rapid Image Synthesis" in Whitted T. and Cook R. eds. "Image Rendering Tricks", Course Notes 16 for **SIGGRAPH'86**, (1986)

[5] Greene N., "Environment Mapping and Other Applications of World Projections", **CGA**, 6(11), pp. 21-29, 1986

[6] Chen S.E., "QuickTime VR - An Image-Based Approach to Virtual Environment Navigation", **SIGGRAPH'95**,pp. 29-38, (1995)

[7] Kontani N., Sugiura H. and Fujino J., "Super-wide-angle imaging by synthesizing images", **J. of TV Soc.** Japan, vol. 48, No. 10, pp. 1189-1195, (1994)

[8] Kingslake R., "Optical System Design", Achademic Press, New York, NY, 1983

[9] Tsai, R.Y., "An efficient and accurate camera calibration technique fro 3D machine vision", **CVPR**, pp.364-374, 1986.