PAPER

Stereo Matching between Three Images by Iterative Refinement in PVS

Makoto KIMURA[†], *Student Member*, Hideo SAITO[†], *Regular Member*, *and* Takeo KANADE[††], *Nonmember*

SUMMARY In the field of computer vision and computer graphics, Image-Based-Rendering (IBR) methods are often used to synthesize images from real scene The image synthesis by IBR requires dense correct matching points in the images However, IBR does not require 3D geometry reconstruction or camera calibration in Euclidean geometry On the other hand, 3D reconstructed model can easily point out the occlusion in images In this paper, we propose an approach to reconstruct 3D shape in a voxel space, which is named Projective Voxel Space (PVS) Since PVS is defined by projective geometry, it requires only weak calibration PVS is determined by rectifications of the epipolar lines in three images Three rectified images are orthogonal projected images of a scene in PVS, so processing about image projection is easy in PVS In both PVS and Euclidean geometry, a point in an image is on a projection from a point on a surface of the object in the scene Then the other image might have a correct matching point without occlusion, or no matching point because of occlusion This is a kind of restriction about searching matching points or surface of the object Taking advantage of simplicity of projection in PVS, the correlation values of points in images are computed, and the values are iteratively refined using the restriction described above Finally, the shapes of the objects in the scene are acquired in PVS The reconstructed shape in PVS does not have similarity to 3D shape in Euclidean geometry However, it denotes consistent matching points in three images, and also indicates the existence of occluded points Therefore, the reconstructed shape in PVS is sufficient for image synthesis by IBR

key words: *image processing, 3D models, computer vision, robot vision*

1. Introduction

For synthesizing images of the real scene from arbitrary viewpoint, one approach is to reconstruct 3D shape of the objects using multiple images, so that images can be generated from the 3D shapes and the texture data 3D geometry reconstruction is a common subject in the field of computer vision, and it requires camera calibration that has difficulty and complexity of acquisition One of the common problems in 3D geometry reconstruction is occlusion in the images 3D geometry can be reconstructed by camera calibration and matching points in the images, while no matching points can be detected for occluded regions Even excepting the problem of occlusion, the detection of matching points is still difficult Especially when the baseline of the cameras being wide, it is more difficult to obtain the correct matching points in each image [6], [7]

There are researches for 3D geometry reconstruction using more than two cameras to solve this problem [8], [10], [11] With processing the 3D geometry in world coordinate, which is generally described as Euclidean geometry, all data can be handled in a common coordinate system Thus, using many cameras make it possible to reconstruct accurate 3D geometry, as the occluded region in an image might be seen from another camera This method has been applied in Virtualized Reality [5], [11], which requires calibration of about 50 cameras This calibration for each camera is performed by checking the correspondence between 3D geometry in world coordinates and 2D geometry in image coordinates

Considering the purpose of image synthesis, Image-Based-Rendering methods can generate the same kind of images For example, morphing method [1], [9] doesn't require 3D shape, and it requires only dense matching points in each image

Recently, projective geometry is often used in the field of computer vision [1], [6], [7], [9], because projective geometry can be determined easier than Euclidean geometry While determination of Euclidean geometry requires a map of correspondences between the points in the image and Euclidean geometry of those points, determination of projective geometry requires only matching points in each image [12] Then, we call the traditional camera calibration "*strong calibration*" and calibration for projective geometry "*weak calibration*" Projective geometry makes it possible to determine the epipolar line for any point in the image, however it has no notion of a world coordinate [12] Thus, projective geometry is easy to calibrate, and denotes the projective relation between the cameras, but it doesn't determine common coordinates like Euclidean geometry

In this paper, we propose an approach to reconstruct a projective 3D voxel space from three images In this method, a 3D voxel space and three orthographically projected images are generated from three input images and weak calibration for the cameras This 3D

Manuscript received July 18, 2001

Manuscript revised February 26, 2002

[†]The authors are with the Department of Information and Computer Science, Keio University, Yokohama-shi, 223–8522 Japan

[††]The author is with the Robotics Institute, Carnegie Mellon University, USA

voxel space is not based on Euclidean 3D geometry, however we can handle this voxel space just like Euclidean voxel space In this 3D voxel space, consistent matching points in all three images can be solved, because matching points in a pair of images can automatically point out a projected point in the other image

2. Epipolar Geometry

Epipolar geometry is one form of projective geometry. Figure 1(a) shows the general camera model with two cameras A point in a scene, which is visible in an image, must exist on a back-projection line from the camera. Therefore, a point in a scene must be equal to the intersection of back-projection lines from each matching point in the images Epipolar plane is a plane going through the line, which connects focus points of the cameras. Epipolar line is a line on an image plane obtained by projecting the back-projection line of the other camera. As shown in Fig. 1(b), all epipolar lines go through a point called *epipole* Epipole is the projected point from the focus point of the other camera.

Fundamental matrix is one form of description about epipolar geometry between two images [3], [12] Actually, fundamental matrix is a 3×3 matrix, with seven degrees of freedom Therefore, fundamental matrix can be solved with only seven matching points in the images with non-linear method [2] More matching points make this solution more accurate. Once the fundamental matrix is solved, weak calibration for the camera pair is established The weak calibration is different from traditional calibration, which we call "*strong calibration*" While the strong calibration is identification of all camera parameters, weak calibration identifies only epipolar geometry of a camera pair

The obtained fundamental matrix transfers a point in an image to an epipolar line in another image. A point on the epipolar line is also transferred to an epipolar line in the original image. Thus, arbitrary corresponding epipolar lines in images can be acquired by fundamental matrix With the consideration of the description in Sect 2, matching points must exist in the corresponding epipolar lines Therefore, fundamental matrix can limit the searching area of matching points.

There are many methods to estimate fundamental matrix from matching points in the images In any method, the accuracy of the estimated fundamental matrix depends on the number and the accuracy of matching points

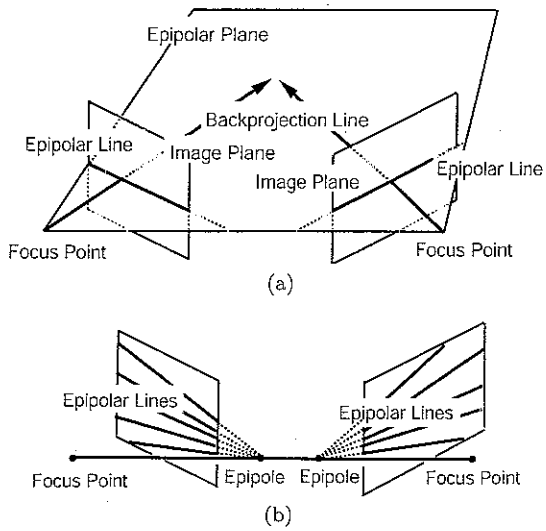


Fig. 1 Epipolar geometry: (a) Epipolar lines exist on the image planes in the camera model; (b) Epipole is the projected position of fucus point of the other camera

3. Definition of Projective Voxel Space (PVS)

In this section, a 3D voxel space based on projective geometry is determined.

3.1 Rectification of Three Images

In this paper, we suppose to use three cameras The fundamental matrices between all pairs of three cameras are required at first. With the fundamental matrices, it is possible to get an epipolar line from an arbitrary point in the other image Then, corresponding epipolar lines can be obtained in arbitrary density.

In two cameras, epipolar lines are distributed as shown in Fig. 2. As mentioned in Sect 2, searching area of matching points can theoretically be restricted on the corresponding epipolar lines Then the rectification of those epipolar lines is reasonable for processing of searching matching points As described below, we apply the same kind of rectification in three cameras.

Theoretically, position of the epipole can be acquired by an eigenvalue problem of fundamental matrix. We utilize the direction of epipole in the images in the following process, however accurate position of epipole is not required. To simplify the processes, we acquire the position of epipole by several practical epipolar lines. As shown in Fig 3, the points for acquisition of epipolar lines for the rectification are ar-

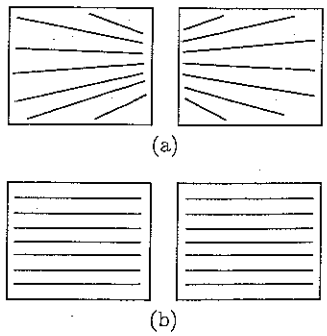


Fig. 2 Rectification between two images: (a) An example of distribution of epipolar lines between two images; (b) Rectified epipolar lines

bitrary set in the vertical direction of epipole. Thus, we can have epipolar lines in arbitrary density in the image plane, so that we can set arbitrary resampling rate in the rectification process. Since our final purpose is to find correct matching points in images, the density of epipolar lines affects the accuracy of outputs. It is difficult to find the suitable density automatically, because it strongly depends on the target scene and the settings of cameras. Therefore, we set the density manually based on our experience.

Applying this process in all pairs of three images, each image has two kinds of epipolar lines. Figure 4(a) shows the distribution of epipolar lines between each pair of three cameras. Considering the intersections of these two kinds of epipolar lines as resampled points of the images, three rectified images are generated. Figure 4(b) shows the rectified images from Fig 4(a). If A and B in Fig 4(b) are detected as the correct matching points by somehow, we can say a point C is the correct matching point for A and B. This is because we have epipolar line from A and B, and C is the intersection of two epipolar lines. Actually, this rule is not always acceptable becase of occlusion, and this subject is discussed in the following sections.

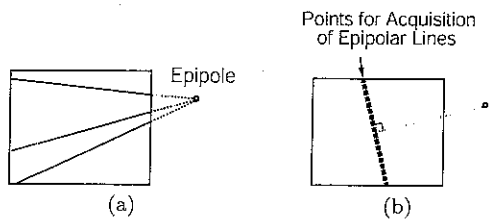


Fig. 3 Acquisition of epipolar lines: (a) Direction of epipole can be acquired by several epipolar lines; (b) Epipolar lines are set based on the location of epipole.

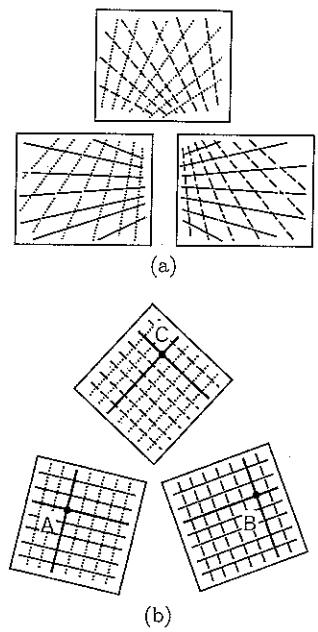


Fig. 4 Rectification between three images: (a) An example of distribution of epipolar lines between three images; (b) Rectified epipolar lines (A, B, C: an example of matching points)

3.2 Definition of Projective Voxel Space (PVS) by Rectified Images

In the previous section, three rectified images are generated. By setting the three rectified images like Fig. 5, a 3D voxel space can be determined. We call this voxel space Projective Voxel Space (PVS). The directions of spatial axes correspond to epipolar lines in each pair of images. In other words, the direction of each axis in PVS corresponds to the projection of each camera. As shown in Fig. 5, the relation between rectified images and PVS is complete orthographic projection. Such relation simplifies the geometrical transformation between PVS and the images. The detection of the existence of a voxel in PVS is equal to the detection of matching points in three rectified images.

PVS is a distorted space of Euclidean one. The concept of distortion between PVS and Euclidean space is shown in Fig. 6. The distortion between PVS and Euclidean space is unknown from weak calibration. Therefore, 3D shape in the real world, which is described in Euclidean geometry, can not be recovered from 3D shape in PVS. When the 3D shape of object is required (i.e. 3D measurement system), you have to reconstruct

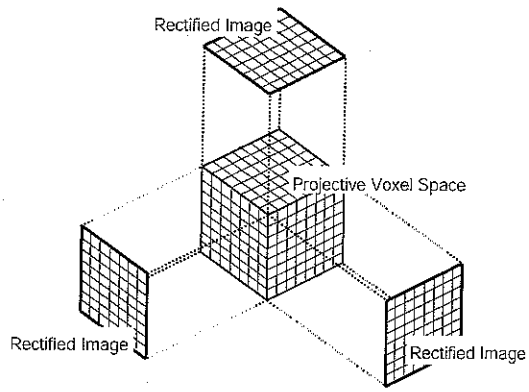


Fig. 5 The connection of the generated images and projective 3D voxel.

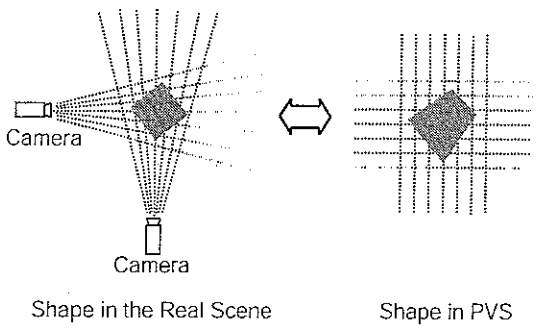


Fig. 6 The relation between real world and constructed PVS.

3D shape in Euclidean geometry, not in PVS However, PVS is enough for synthesis of the images from new viewpoints Image synthesis doesn't require 3D shape itself, because images can be synthesized from consistent and correct matching points in all images [1], [9] Because PVS provides the way to handle three images at one time, PVS is effective to detect consistent matching points in three images

4. Iterative Refinement in PVS for Stereo Matching

In this section, a stereo matching method between three images is described First, Zitnick's method [13], which reduces the ambiguity of matching between two images, is mentioned After that, it describes how we extend the method to PVS Since the existence of matching points in images is equal to the existence of a voxel in PVS, all voxels are initially evaluated by correlation between the images Then, all voxels are refined by an iterative method with checking the around voxels

4 1 Iterative Refinement of Two Cameras Stereo Matching

Zitnick et al proposed an iterative method [13] to reduce the ambiguity of matching points between two images Their method is based on two general assumptions of stereo matching [4] The first assumption is that a single unique match exists for each pixel in images The second assumption is that disparity values are generally continuous They called their method "cooperative stereo algorithm," which uses disparity space to utilize those two assumptions Figure 7 shows the concept of Zitnick's cooperative stereo algorithm First of all, they define lattice points in a 2D plane, which is so called disparity space, so that a matching point in images can be equal to the existence of lattice point in this plane Then, they defined a parameter, which stands for certainty of the matching We call this parameter "likelihood" in this paper Likelihoods for all combinations of matching points are initialized by correlation values Then, they took into account two above-mentioned assumptions

Using the first assumption about uniqueness, an area called "inhibition area" in the disparity space is defined for each combination of matching points In Fig 7(a), the inhibition area for the black point is shown as the light gray area The inhibition area is equal to the projection lines from the black point to the cameras If a point in disparity space denotes correct matching points, there should be no more matching points in its inhibition area

When a checking point in the disparity space has singular high likelihood in its inhibition area, it reliably denote correct matching points As long as the variance of likelihoods in the inhibition area is small,

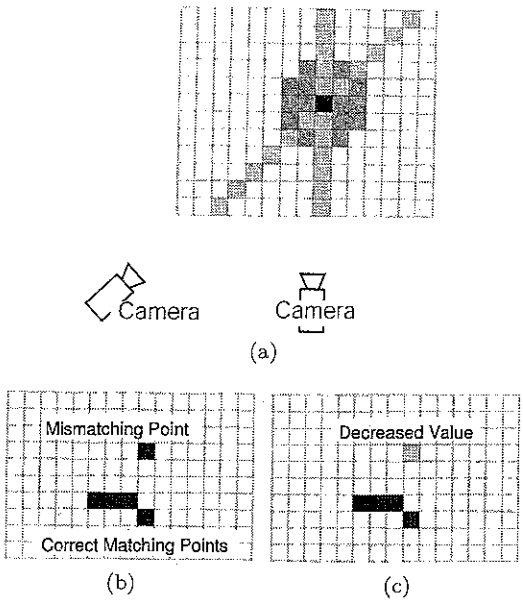


Fig 7 Concept of Zitnick's cooperative stereo algorithm: (a) The black point is checking point, the light gray area is inhibition area for the black point, and the dark gray area is smoothing area for the black point ; (b) When many points in the disparity space, that are drawn as black points in this figure, have high correlations drawn as black points, it is difficult to detect correct matching points ; (c) Zitnick's cooperative algorithm will decrease the value in the non-unique or non-continuous point, because there are two other high values in its inhibition area and there are no other high values in its smoothing area However, the unique and continuous matching points will remain relative high values

it is difficult to discern whether that point denotes correct matching points or not Then, they proposed an iterative calculation to enlarge the variance in the inhibition area for any point in the whole disparity space In short, their approach is realized with division of the likelihood value by the sum of the squared likelihoods in its inhibition area

Using the second assumption about continuity, they applied smoothing to the likelihood values with reference of the neighborhoods in disparity space, which is shown as a dark gray area in Fig 7(a)

In practice, the division and smoothing are applied at one time With applying these processes iteratively, the values in whole disparity space will cooperatively changes into satisfying the both of the two above-mentioned assumptions In Fig 7(b), the mismatching point has many high values in its inhibition area, and also has no high values in its smoothing area On the other hand, the correct matching points have only a little high values in its inhibition area, and also have many high values in its smoothing area Then, the mismatching point glows relatively lower, and the correct matching points glow relatively higher Therefore, only correct matching points will remain high values after the iteration of enough times

4.2 Initialization of Voxel Values

As described in Sect. 3, PVS is a voxel space determined by the projective geometry between three images. Conversely, the relation between three images and voxel space can be described as follows: there is a scene in this 3D voxel space, and the three rectified images are orthographic projected images of this scene. In this way, the detection of matching points in each pair of the images is equal to the detection of the surface of the objects in PVS.

Like most of the stereo matching methods, our method for searching matching points is based on correlation. We calculate normalized correlation values for each voxel. Normalized correlation can be calculated as C in equation (1). $I_k(k = 1, 2)$ is the intensity of the image, and $\overline{I_k}$ is the average of intensities of pixels in the window for correlation, which is sized $(2m+1) \times (2m+1)$. I'_k is the subtraction between I_k and $\overline{I_k}$ in the window. $\sigma(I_k)$ is the standard deviation of the intensities in the window.

$$C = \frac{\sum_{i=-m}^{m} \sum_{j=-m}^{m} I'_1(u_1, v_1) \times I'_2(u_2, v_2)}{(2m+1)^2 \sqrt{\sigma^2(I_1) \times \sigma^2(I_2)}} \quad (1)$$

$$\overline{I_k(u_k, v_k)} = \sum_{i=-m}^{m} \sum_{j=-m}^{m} \frac{I_k(u_k + i, v_k + j)}{(2m+1)^2} \quad (2)$$

$$I'_k(u_k, v_k) = I_1(u_k + i, v_k + j) - \overline{I_k(u_k, v_k)} \quad (3)$$

Since a voxel corresponds to the matching points, it seems to be reasonable to adopt the average of all correlation values between three pairs of the images. However, we have to be careful about occlusion. When a voxel is in occlusion, the occluded voxel corresponds to matching points in only two images. In such case, the other voxel must exist between that voxel and one of the cameras. This subject is shown in Fig. 8. Thus, the occluded voxel should not be evaluated with the average of three correlation values. It should be evaluated with the correlation value between the correct matching points. Generally, we can expect the correlation value between correct matching points is higher than the correlation value between mismatching points. Then, we adopt the maximum of three correlation values as a voxel value. Because we don't know a voxel is in occlusion or not in practice, this adoption rule is used for all voxels. Applying this operation to all voxels, each voxel has some value.

$$V = Max\ of\{C_{12}, C_{23}, C_{31}\} \quad (4)$$

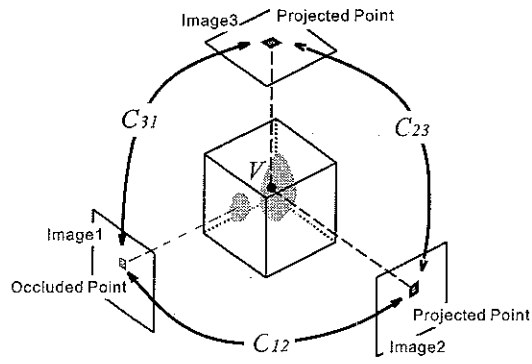


Fig. 8 Three correlations between three images. V is the checking voxel, and C_{ij} is the correlation between image i and image j. In this figure, C_{12} and C_{31} are invalid because of occlusion, and C_{23} is valid for evaluating the checking voxel V.

4.3 Iterative Refinement of Voxel Values

In this paper, we extend the Zitnick's algorithm to be suitable for three input images. It is a kind of extension from 2D to 3D PVS.

In the same way of Zitnick's method, we can set inhibition areas in PVS as shown in Fig. 9. Here, occlusion must be considered again. Because we assume three input images, even occluded point in an image may be visible in the other two images, as shown in Fig. 8. In such case, the checking voxel denotes correct matching points, however another voxel exists in one of its inhibition areas. On the other hand, the hiding voxel in this case also has another voxel in one of its inhibition areas.

Thus, the first assumption — a single unique match exists — is acceptable only in two projection lines in this case. This means we should excuse the existence of another voxel in one of the three axis directions. Therefore, two of the inhibition areas in Fig. 9 are chosen by the sum of likelihoods in each of them. This selection of inhibition area is formulated as a nonlinear function $M()$ in later mention. With taking into account this subject of the first assumption of uniqueness, we perform the refinement of voxel values by Eq. (5).

$$V_{n+1}(x, y, z) = \left(\frac{V_n(x, y, z)}{M(S_x(x, y, z), S_y(x, y, z), S_z(x, y, z))} \right)^{\alpha} \quad (5)$$

$S_x(X, Y, Z)$ is a sum total of squared values of all voxels in the projection line from voxel (x, y, z) to the

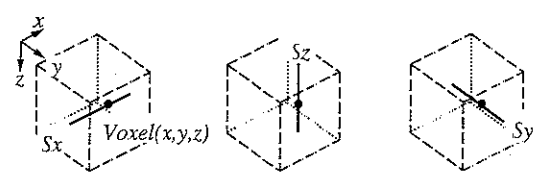


Fig. 9 Three parts of inhibition area for a voxel in PVS. Two of them are chosen as inhibition area, because of the consideration for occlusion.

direction of x axis. In other words, $S_x(X,Y,Z)$ is a sum total of squared values in one of the inhibition areas in Fig. 9. $S_y(X,Y,Z)$ and $S_z(X,Y,Z)$ are the same kind of values in the other inhibition areas. $M(a,b,c)$ is sum of two smaller values of a, b, and c. $V_n(x,y,z)$ is a voxel value in the geometry (x,y,z) at iteration n. α is a constant, which denotes the strength of convergence. When α is too small, the variance of all voxel values changes only a little, so that the voxel values can not converge. When α is too large, the correct matching points are instantly lost. We set the value $\alpha = 3$ based on our experience.

First, we check the sum total of squared values on each inhibition area, and throw away the maximum one in three values, which is indicated as $M()$. This is because we excuse the existence of another voxel in one of the inhibition areas, as mentioned above. Then, we have two values left. Next, we decimate the voxel value by sum of the two values, and the refined value is the third power to the decimated value. In the meaning of absolute value, all voxels are revised to lower values by applying this operation. However, the variance of the voxel values is reviled to lager.

Using the second assumption about the continuity of disparity, we also apply a smoothing method to the voxel values in the iteration. The smoothing mask defined by the assumption of disparity countinuity has a spherical distribution in PVS, which is also applied in Zitnick's method [13]. In this paper, we apply the modified smoothing mask as shown in Fig. 10, instead of a spherical mask. This mask implies an assumption of uniform disparity, which is more restricted assumption than continuity assumption, but we can obtain better voxel values in the "uniform disparity" neighborhoods. PVS includes three kinds of disparities, because it handles matching points in three pairs of the images at one time. As shown in Fig. 10(a), uniform disparity area in a pair of the images stands for a line on a cross section plane in PVS. The directing arrows in Fig. 10(a) show the directions of the continuities of voxels. Considering these directions, we determine a plane with thickness of 1 voxel in PVS, which is shown in Fig. 10(b). Because this plane is parallel to all directing arrows in Fig. 10(a), applying smoothing process in this plane is reasonable. We also consider the distance from the voxel to be smoothed. Thus, a smoothing mask is determined to be a circle on the plane in PVS, hich is shown in Fig. 10(c). The values of all voxels in the smoothing mask are $\frac{1}{W}$ (W is the area size of the smoothing mask). To perform the smoothing process, this smoothing mask is three-dimensionally convoluted to all voxels in PVS.

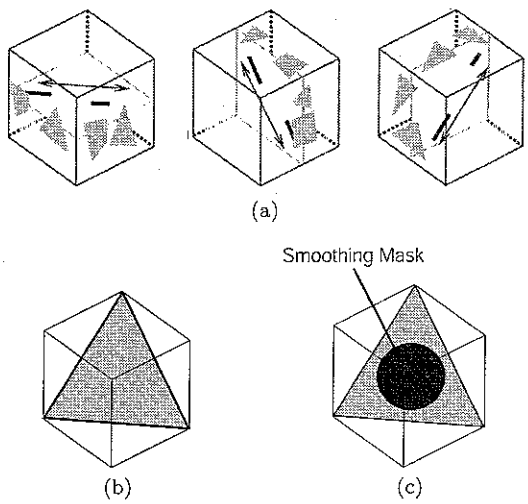


Fig. 10 Definition of smoothing mask shape: (a) There are three kinds of disparities in PVS, because three pair of images are handled in it. Therefore, there are three kinds of "uniform disparity" in PVS. In this paper, we assume that each object in the scene has uniform disparity in each pair of images; (b) a "uniform disparity" plane, which is parallel to three directions in (a); (c) the smoothing mask has a shape of circle in the plane of (b).

In this way, the practical refinement process is performed by the smoothing method and Eq. (1). By applying this refinement iteratively, all voxel values affect to others. Finally, continuous and unique likelihoods are generated in PVS. In the generated likelihoods, the uniqueness are considered with the possibility of occlusion.

5. Intermediate View Synthesis Based on the PVS

In the previous sections, voxel values are refined in PVS, so that the shape of the object can be detected as the local maximum voxels. The obtained shape itself is much different from the shape in Euclidean geometry. However, the obtained 3D shape contains not only visible matching points between the input images, but also the projected position of occluded points on the objects.

The image synthesis is performed by a simple 2D morphing method, which calculates a set of weighted mean geometry of the matching points and weighted mean color of the matching points. Using the projected positions of occluded points, even the geometry of occluded points can be synthesized by the 2D morphing. Then, occluded points can appear or disappear by moving the virtual viewpoint, and it makes the synthesized images plausibly.

6. Experiments

To show the effectiveness of our proposed method, we applied it to some real images. The parameters (number of matching points for weak calibration, resampling resolution of rectification, size of correlation window, and number of iteration) were set manually based on our experiences, because it depends on the target scene.

and distribution of cameras

Before the actual experiments, we evaluated several methods for estimating fundamental matrices The evaluated methods are a linear method with data normalization, a method using least median of squares, and a nonlinear method based on gradient-weighted epipolar errors. In our experiments, the camera's position, angle, and zooming are static, so that images for weak calibration can be easily taken For example, matching points can be easily acquired from scenes of lighting LED in dark room, then sufficient numbers of matching points can be acquired in high accuracy Generally, the accuracy of estimated fundamental matrix strongly depends on the number and the accuracy of given matching points As described later, we had huge number (over 300) of matching points to estimate fundamental matrices, and the mismatching points were canceled manually Then, the errors of the estimated fundamental matrices, which can be calculated as the distance between the calibration points and the epipolar lines, were less than a pixel width in any evaluated method This accuracy is enough for our purpose, because obtained fundamental matrices are used only for determining corresponding epipolar lines in our method We actually adopted a method using least median of squares [3], [12].

6.1 Scene without Occlusion

We applied the proposed method to an object of simple shape without occlusion, which is shown in Fig 11. Input images are rectified as shown in Fig 12. Silhouettes of the object in the images are obtained using subtraction of background images and manual modification All voxels are filled with normalized correlations using windows sized 9×9 Figure 13(a) shows the normalized correlation values in a cross section plane in PVS. These values mean certainty of matching points between the scan lines in Fig 12(a) and (c). We can see that it is difficult to obtain the correct matching points from raw correlation values. Figure 13(b) shows the refined voxel values by the proposed operation. In the refined voxel values, the ambiguity of the matching points is much reduced As shown in Fig. 13(c), the matching points are detected as the local maximum voxel In Fig. 14(a), an image from the virtual

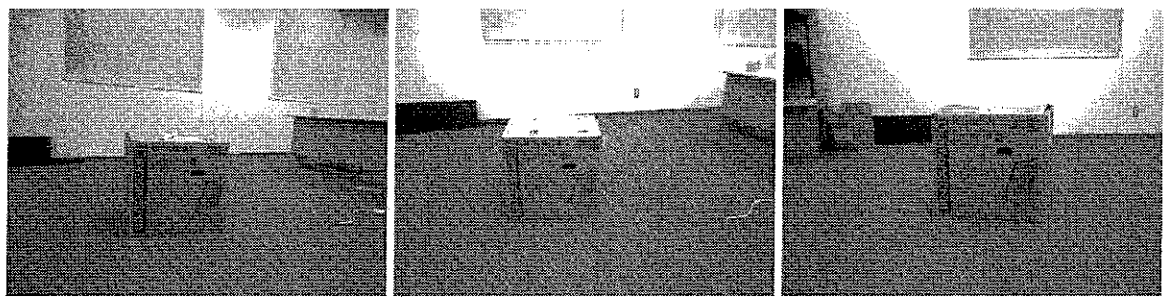


Fig. 11 Input images: (a) left image; (b) top image; (c) right image

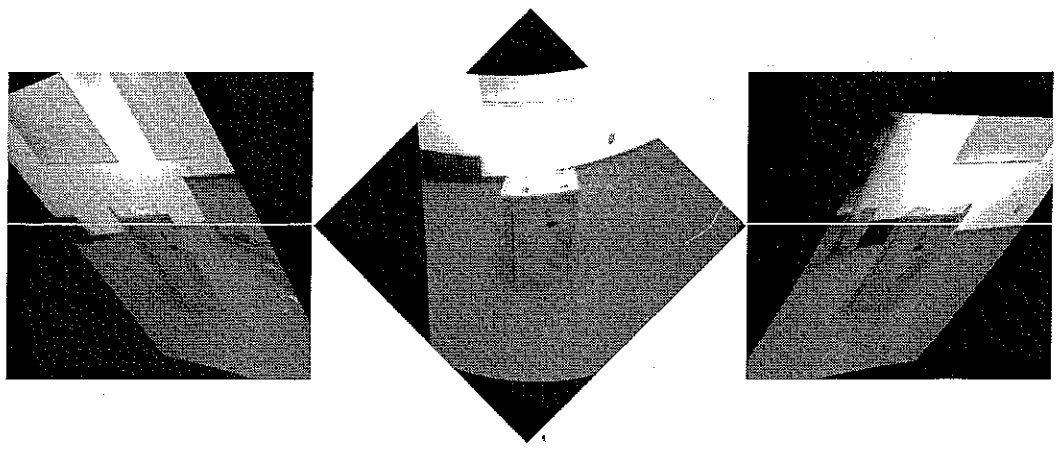


Fig. 12 Rectified images and example of epipolar lines between each pair of images: (a) left image; (b) top image; (c) right image.

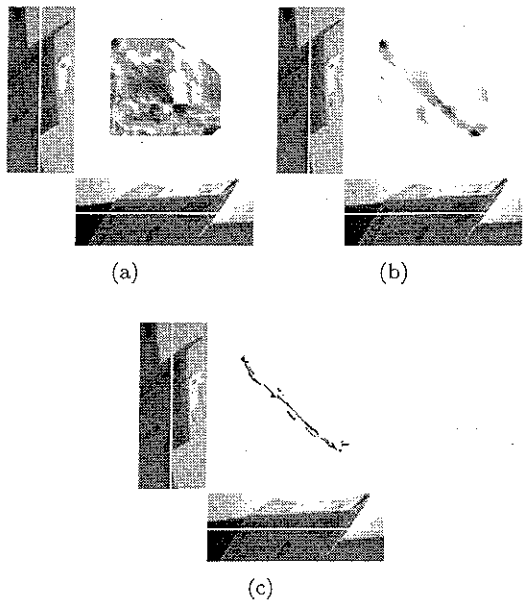


Fig. 13 Likelihood values in a cross section of PVS: (a) adopted maximum values from three correlations; (b) refined likelihood value from (a); (c) detected matching points from (b).

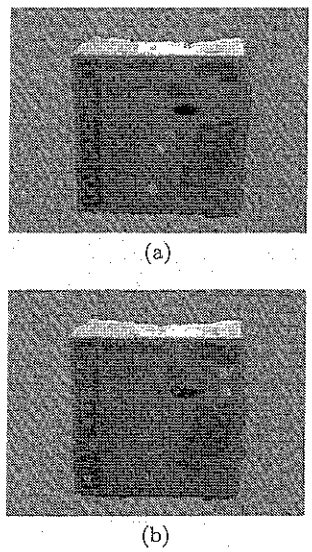


Fig. 14 Synthesized images (the virtual viewpoint is between left and right camera): (a) based on refined voxel values; (b) based on non-refined correlations.

viewpoint is synthesized. For comparison, Fig 14(b) shows the synthesized image based on non-refined correlations. While we can easily read some characters on the box in Fig. 14(a), the texture on the box is much collapsed and blurred in Fig. 14(b), because of the mismatching points.

6.2 Scene Including Occlusion

For the experiment of Figs. 15–17, the fundamental matrices were estimated with about 500 matching points in each image. All baselines of each pair of cameras were about 0.8m. A 256×256×256 voxel space is constructed from 320×240 sized three images. The correlation values were calculated with windows sized 11×11. The number of iteration times was two. Figure 15 shows input images and example of epipolar lines on them. Figures 16 shows rectified images.

In this case, there are may occlusions: for example, the right foot of the right person is invisible in the right image, and the left thigh of the left person is invisible in the left image at the same time. Figure 17 shows depth maps based on the obtained matching points in PVS. The disparity maps for the occluded areas are well obtained, because proposed method tacitly uses only valid matching points in three pairs of the images. Figure 18 shows a synthesized image with virtual viewpoint at the center of three cameras. The above mentioned occlusions are plausibly synthesized in the image.

Figure 19 shows the yet another input images with another camera distributions. All baselines of each pair of cameras were about 100 cm. For these images, the fundamental matrices were solved with about 300 matching points in preprocessing. A 256×256×256 voxel space is constructed from 640×486 sized three images. The rectified images are shown in Fig. 20. The correlation values were calculated with windows sized 11×11. The number of iteration times was three.

Figure 21 shows some examples of voxel values in a cross section plane in PVS. This cross section denotes matching between epipolar lines shown in Fig. 20. Figures 21(a) and (b) are shown just for the reference data, and not used in the proposed method. Figure 21(a) shows raw correlation values between left and right images. Figure 21(b) shows the average of three correlation values between three pair of images. Since the matching points are detected as local maximum voxels, it is difficult to obtain the correct matching points from Figs. 21(a) and (b). Figure 21(c) shows the effectiveness of the adoption of the valid correlation value, which is described in Sect. 4.2. In this figure, many mismatching points have reduced values. Figure 21(d) shows the voxel values after the iterative refinements. After iterative refinements, the correct matching points have singular high values. Thus, matching points are finally detected as shown in Fig. 21(d).

For evaluation of the estimated matching points, we detected matching points on the epipolar line by manual, which is shown in Fig. 21(e). The matching results of the proposed method includes 6.6 pixel width in the average error, while the non-refined voxels in Fig. 21(c) gives 17.2 pixel width in the average error. Thus, much mismatching is reduced in the iterative refinement. These errors are mostly caused by the luck of texture on objects in the scene. However, the error value does not always affect the quality of synthesized images directly. When the object has flat texture and

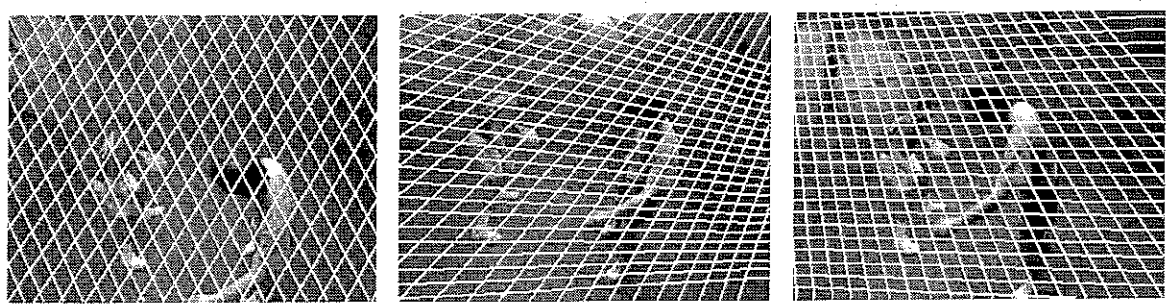
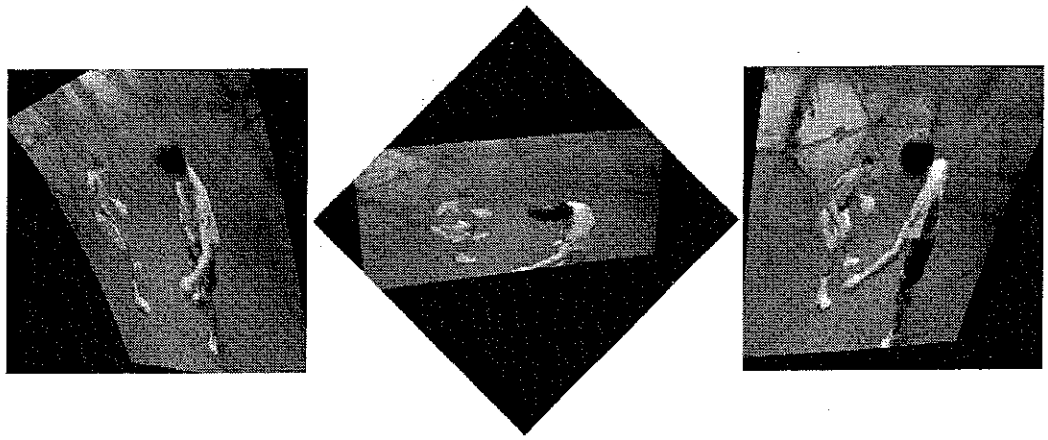


Fig. 15 Input images and example of epipolar lines on them



(a) (b) (c)

Fig 16 Rectified images generated from the input images based on epipolar geometry

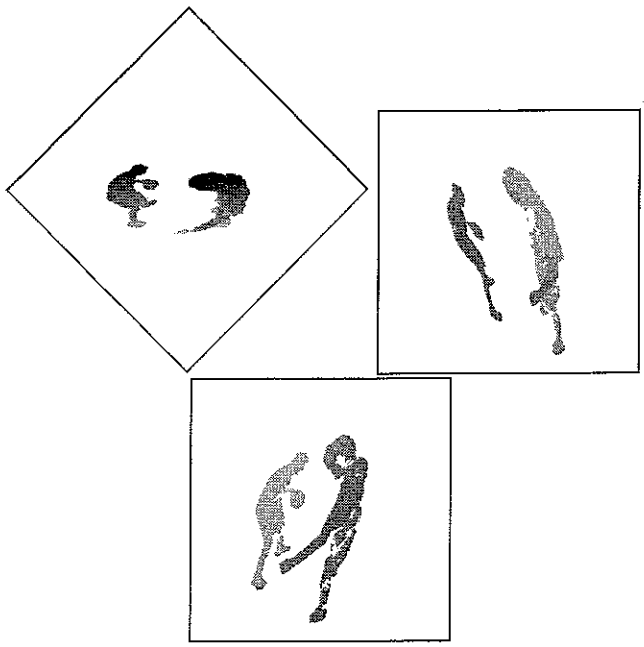


Fig 17 Estimated depth maps in PVS

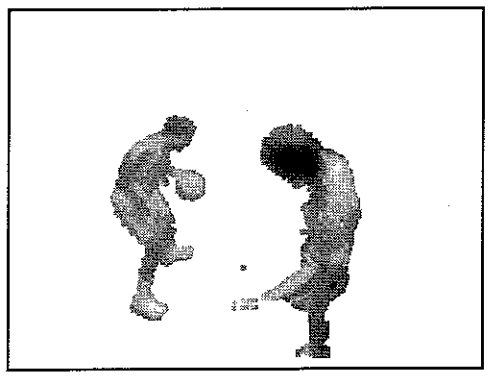
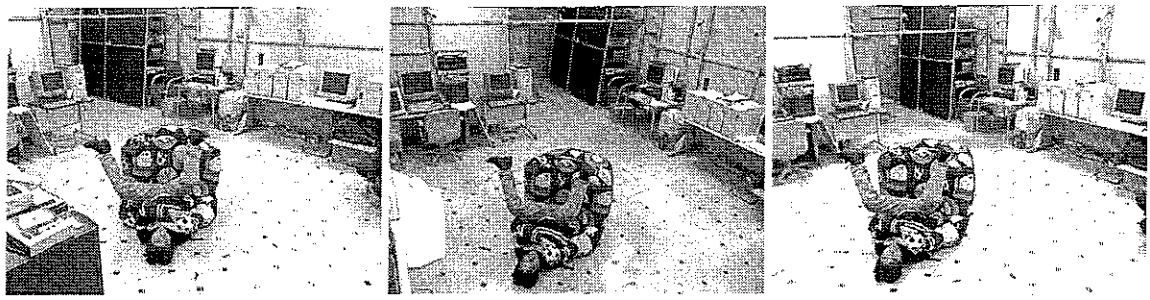


Fig. 18 Synthesized image (the virtual viewpoint is in the center of three input cameras)

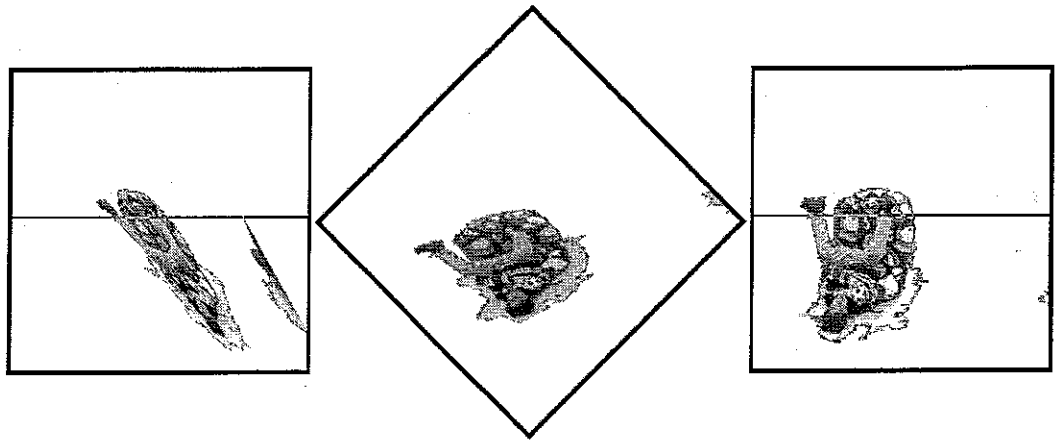
the error of mismatching points is enough little, the quality of synthesized images will not be so much damaged

Figure 22 shows synthesized images based on the acquired matching points The viewpoint is moving through baselines of three cameras



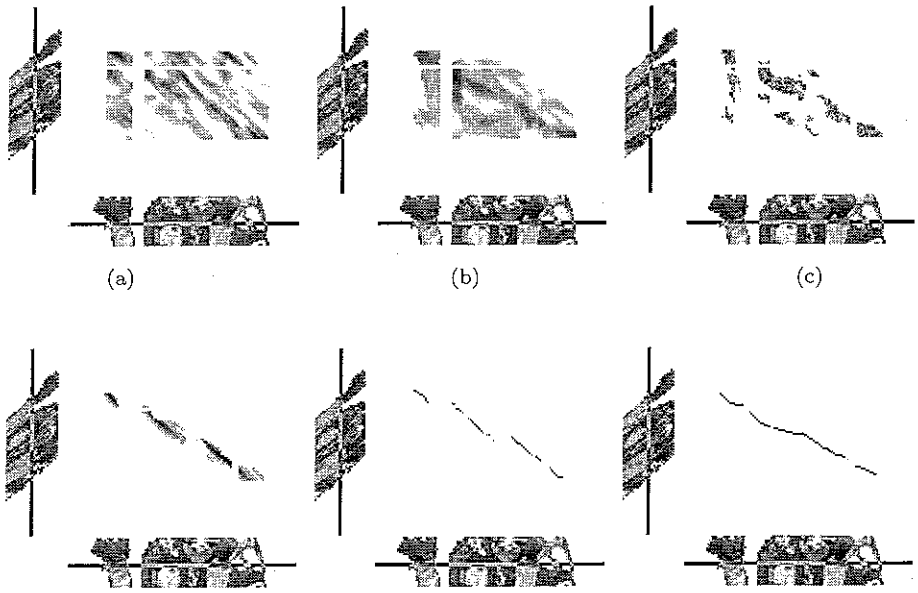
(a) (b) (c)

Fig. 19 Input images: (a) left image; (b) top image; (c) right image



(a) (b) (c)

Fig 20 Rectified silhouette images with an example of epipolar line between two images : (a) left image; (b) top image; (c) right image



(a) (b) (c)

(d) (e) (f)

Fig. 21 Likelihood values and matching points: (a) correlation values between two images; (b) average of correlation values between three pairs of images; (c) adopted values from three correlations; (d) refined likelihood values from (c); (e) estimated matching points based on (d); (f) correspondence pointed by human

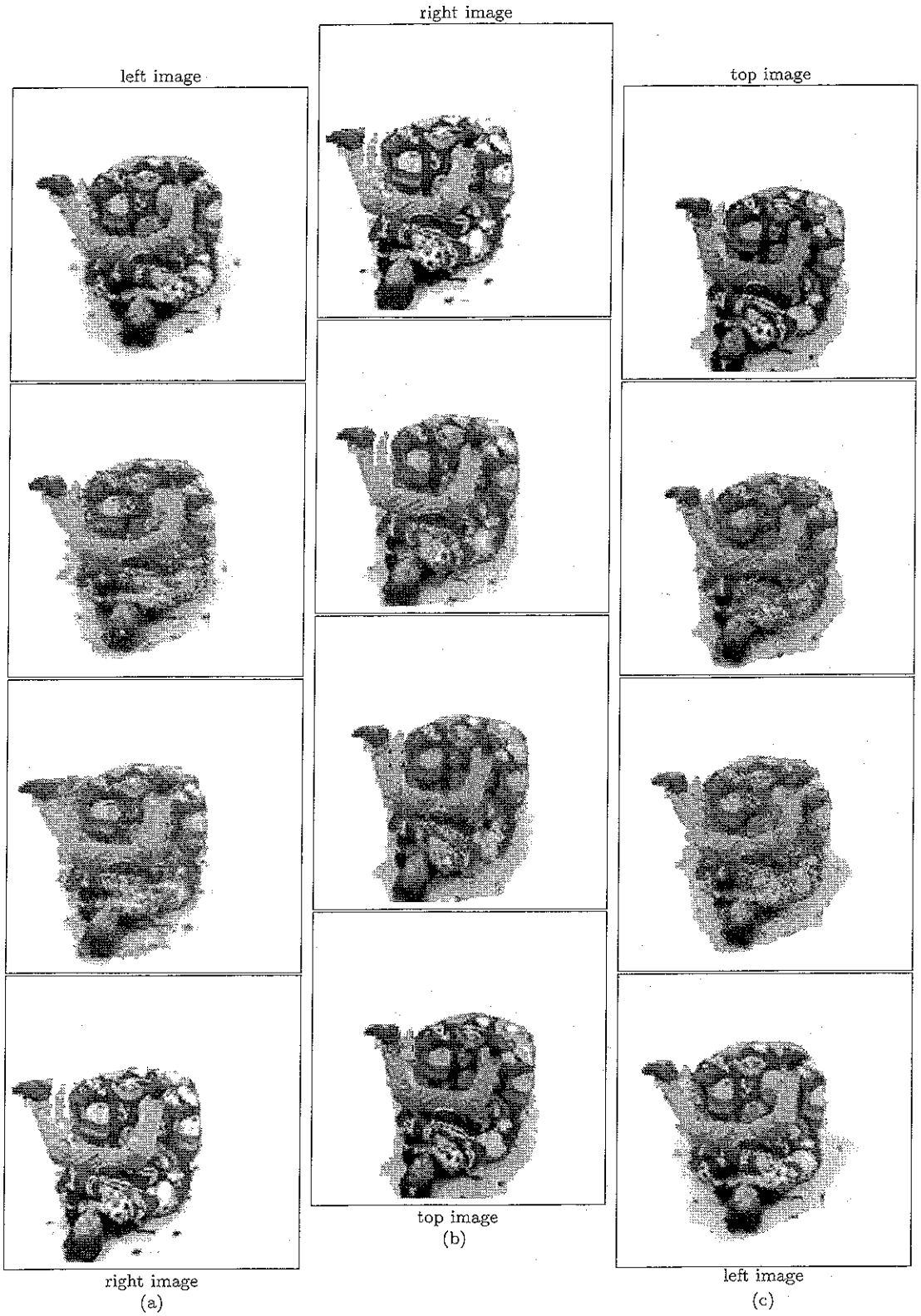


Fig. 22 Synthesized images: The viewpoint is moving (a) from left image (top) to right image (bottom); (b) from right image (top) to top image (bottom); (c) from top image (top) to left image (bottom)

7. Conclusion

This paper focused on 3D voxel concept in epipolar geometry. PVS is constructed with three images and fundamental matrices of all image pairs. PVS is easy to determine, because it requires only weak calibration, while Euclidean geometry requires strong calibration. The rectified images can be handled as orthographic projected images of PVS. In this paper, correspondence between three images is estimated by normalized correlation and iterative refinements.

The basic idea of iterative refinements in this paper is based on Zitnick's method [13]. It uses general assumptions of stereo matching [4], that a single unique match exists for each pixel in images, and disparity values are generally continuous. In this paper, such assumptions are utilized with the considerations for occlusion. Then, unique and continuous matching points are acquired in three images.

Because the proposed method uses normalized correlation as the initial value of iterative refinements, it still requires texture on the objects in input images. We think it's possible to decrease the dependence for texture by using some kinds of restriction in 3D space. Solving such problem is our future work.

References

[1] S. Avidan and A. Shashua, "Novel view synthesis by cascading trilinear tensors," IEEE Trans. Visualization and Computer Graphics, vol 4, no 4, pp 293–306, 1998

[2] R. Hartley, "In defense of the 8-point algorithm," IEEE ICCV95, pp 1065–1070, 1995.

[3] S. Laveau and O. Faugeras, "3-D scene representation as a collection of images and fundamental matrices," Research Report 2205, INRIA Sophia-Antipolis, France, Feb. 1994.

[4] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, ISBN 0–716–71567–8, W.H. Freeman and Company, 1982.

[5] P.J. Narayanan, P.W. Rander, and T. Kanade, "Constructing virtual worlds using dense stereo," Proc. 6th IEEE Int. Conf. on Computer Vision (ICCV '98), Bombay, India, pp 3–10, Jan. 1998.

[6] P. Pritchett and A. Zisserman, "Wide baseline stereo matching," Proc. 6th IEEE Int. Conf. on Computer Vision (ICCV '98), Bombay, India, pp 754–760, Jan. 1998.

[7] P. Pritchett and A. Zisserman, "Matching and reconstruction from widely separated views," Proc. 3D Structure from Multiple Images of Large-Scale Environments European Workshop (SMILE '98), p-viii+346, pp 78–92, 1998.

[8] H. Saito, S. Baba, M. Kimura, S. Vedula, and T. Kanade, "Appearance-based virtual view generation of temporally-varying events from multi-camera images in the 3D room," Computer Science Technical Report, CMU-CS-99-127, April 1999.

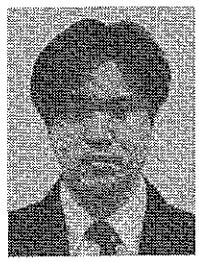
[9] S.M. Seitz and C.R. Dyer, "View morphing," Proc. SIGGRAPH '96, pp.21–30, 1996.

[10] S.M. Seitz and C.R. Dyer, "Photorealistic scene reconstruction by voxel coloring," Proc. Computer Vision and Pattern Recognition Conf., pp.1067–1073, 1997.

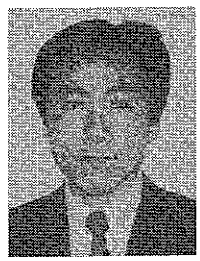
[11] S. Vedula, P. Rander, H. Saito, and T. Kanade, "Modeling, combining, and rendering dynamic real-world events from image sequences," Proc. 4th Int. Conf. on Virtual Systems and Multimedia, Gifu, Japan, Nov. 1998.

[12] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review," Research Report 2927, INRIA Sophia-Antipolis, France, July 1996.

[13] C.L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," IEEE Trans. Pattern Anal. & Mach. Intell., vol 22, no 7, pp 675–684, 2000.



Makoto Kimura received B.E., M.E., and Ph.D. degrees in Electrical Engineering from Keio University in 1995, 1997, and 2002 respectively. Since 1998 until 1999 he had been a visiting scholar of The Robotics Institute, CMU, USA. He has been engaging in the research areas of computer vision and image processing.



Hideo Saito received B.E., M.E., and Ph.D. degrees in Electrical Engineering from Keio University in 1987, 1989, and 1992 respectively. He had been working for Department of Electrical Engineering, Keio University as an Instructor since 1992, and now is Associate Professor. Since 1997 until 1999 he had been a visiting researcher of The Robotics Institute, CMU, USA. He has been engaging in the research areas of computer vision and image processing. He is a member of IPSJ, SICE, and IEEE.

Takeo Kanade is U.A. Helen Whitaker University Professor of Computer Science and Robotics at Carnegie Mellon University. He received his Doctoral degree in Electrical Engineering from Kyoto University, Japan, in 1974. After holding a faculty position at Department of Information Science, Kyoto University, he joined Carnegie Mellon University in 1980, where he was the Director of the Robotics Institute from 1992 to 2001. He has worked in multiple areas of robotics: computer vision, multi-media, manipulators, autonomous mobile robots, and sensors. He has written more than 250 technical papers and reports in these areas, as well as more than 15 patents. He has been the principal investigator of a dozen major vision and robotics projects at Carnegie Mellon. He has been elected to the National Academy of Engineering. He is a Fellow of the IEEE, the ACM, and American Association of Artificial Intelligence (AAAI), and the former founding editor of International Journal of Computer Vision.