Toward Automatic Reconstruction of 3D Environment with an Active Binocular Head

Chung-Yi Lin* Sheng-Wen Shih† Yi-Ping Hung‡

julian@iis.sinica.edu.tw swshih@ncnu.edu.tw hung@iis.sinica.edu.tw

* Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

† Department of Computer Science and Information Engineering, National Chi Nan University, Nantou, Taiwan.

‡ Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan.

Abstract

In this paper, we propose a new automatic approach to reconstructing a model for the 3D environment by use of an active binocular head. To efficiently store and access the depth estimates, we propose the use of the inverse polar octree which can transform both the unbounded estimate and the unbounded estimation error into a bounded 3D space with appropriate resolution. The depth estimates are computed by using the asymptotic Bayesian estimation method, which includes the use of Markov random fields. In order to apply this method, the active binocular head (the IIS head) has been calibrated with very high accuracy. The path of the local motion required by the asymptotic Bayesian method is determined online automatically to reduce the ambiguity of stereo matching. Some rules for checking the consistency between the new observation and the previous observations have been developed to properly update the inverse polar octree. Experimental results have shown that the proposed approach is very promising for automatic generation of 3D models which can be used for rendering a 3D scene in a virtual reality system.

Keywords: *Active Vision, Stereo Vision, 3D Reconstruction, Asymptotic Bayesian Estimation, 3D Data Integration, View Interpolation*

Conference Track : *Computer Vision.*

1 Introduction

In the last few years, virtual reality (VR) has found many applications in different areas, such as education, business and entertainment. Because of the rapid growth of VR applications, automatic generation of 3D models from images has attracted much attention recently. One popular approach to reconstructing 3D models from images is to use the stereo vision techniques. However, it is well known that stereo correspondence is a very difficult problem. Although some multiple-baseline stereo vision systems have been proposed to simplify the stereo correspondence problem [1, 2], the results of 3D reconstruction obtained by using automatic stereo matching algorithms [3] are still not reliable enough for practical use. Therefore, most 3D reconstruction systems used at the present time choose to solve the stereo correspondence problem either manually or semi-automatically [4, 5].

If one wants to render a 3D scene by computing the shading of an object given the position of light sources and surface properties of the object, the 3D model of the object has to be very accurate because the computed shading appearance is very sensitive to 3D noise. On the other hand, when using image-based techniques to render a 3D scene with texture extracted from real images, the quality of the rendered image is more tolerable to inaccurate 3D data. Knowing that image-based techniques do not require accurate 3D reconstruction and that the stereo correspondence problem is an ill-posed problem, we do not intend to reconstruct highly accurate 3D model of the scene. Instead, our goal is to reconstruct an approximate 3D model having some associated texture information such that this approximation model, together with the texture information, can be used to synthesize images which look similar to the real ones when observing from arbitrary viewpoints within a pre-specified viewing area. In other words, if the synthesized image looks different from the real image when observing from a new point of view, then our goal is to update the current scene model so that the model will be consistent with all the previous views the vision system has observed. Hopefully, the scene model will become more accurate as the vision system samples more viewpoints within the specified viewing area.

In this paper, we propose a new approach to reconstructing a model of the 3D environment automatically by using a well-calibrated active binocular head [6]. The reconstructed 3D

points and their gray level values are stored in a volumetric data structure, i.e., the inverse polar octree, which will be described in section 2. An active control scheme has been used to minimize the ambiguity in stereo matching. The 3D structure of the scene is estimated by using the asymptotic Bayesian estimation method [7]. Details of the reconstruction process is described in section 3. Some experimental results on reconstructing the 3D model of a complex scene are presented in section 4. Concluding remarks are given in section 5.

2 Inverse Polar Octree

Two internal representations are frequently used to describe a 3D scene, namely the 3D mesh-based representation and the voxel-based representation. Because 3D data will be accessed repeatedly for examining the consistency of the 3D estimates observed from different views during the reconstruction process which will be described in the next section, we need to find a representation that can provide efficient access of the 3D data stored. Since the complexity of accessing a voxel in an octree, which is $O(\log(N))$, is much lower than that of accessing a 3D point in a mesh-based representation, we have chosen the voxel-based octree to store the reconstructed 3D information. To use the voxel-based 3D representation, we have to first solve the problem of packing the 3D information contained in the infinite 3D space into the finite memory space in the computer. To deal with this problem, we first notice that the 3D measurement error of a stereo vision system is inversely proportional to the distance between the object and the stereo cameras[8]. This fact suggests that uniform quantization of the 3D data obtained by the stereo vision system should be inefficient. A better quantization scheme is to have the resolution of the volumetric representation inversely proportional to the object distance. However, non-uniform quantization will result in complicated octree representations. Our solution to this problem is to take an *inverse polar transform* before quantizing the estimated 3D data into voxels. This inverse polar transform can be an inverse cylindrical transform or an inverse spherical transform. As an example, the inverse spherical transform is described in the following :

1. Transform the 3D Cartesian coordinates, $(x,\ y,\ z)$, to a spherical coordinate system, $(\rho,\ \theta,\ \phi)$.

2. For the 3D spherical coordinates, $(\rho,\ \theta,\ \phi)$, compute its inverse polar coordinates, $(\frac{1}{\rho},\ \theta,\ \phi)$.

There are two major advantages of taking the inverse spherical transform. The first one is that, after the transformation, all the surrounding 3D objects farther than a minimum distance to the observer, say R_{min}, will be enclosed within a sphere with radius $\frac{1}{R_{min}}$. In other words, the infinite 3D world outside a sphere is now mapped into a finite sphere, as shown in Figure 1. The second advantage is that, after taking the inverse spherical transform, we can apply a uniform quantization because the estimation error is now bounded, which is explained below in more detail.

Let ρ be the distance of an object point away from the observer. Since the 3D estimation error is proportional to the object distance, the 3D estimation error of the object point is approximately $k \cdot \rho$, where k is a constant determined by the configuration of the stereo cameras. That is, the estimate of the object distance may be $(1 + k)\rho$. Notice that the estimation error is unbounded because the estimation error will approach infinite as the object distance approaches infinite. However, after the inverse polar transformation, the object distance is now mapped to $\frac{1}{\rho(1+k)}$. Since the 3D estimation error is usually much smaller than ρ, i.e., $k \ll 1$, we have

$$\frac{1}{\rho(1+k)} \approx \frac{1}{\rho} - \frac{k}{\rho}. \tag{1}$$

Here, the second term of the right hand side of equation (1) is the transformed estimation error, which is now bounded by $\frac{R}{R_{min}}$ if $\rho > R_{min}$. If we choose the quantization unit to be $\frac{k}{R_{min}}$, then the estimation error will be less than the quantization error for all the object points outside the sphere of radius R_{min}. It is important to have a quantization error larger than the estimation error, because if the estimation error is larger than the quantization unit, then there will be many undesired "ghost" voxels that are caused by the estimation noise and are located around the real object position. The increase of resolution, after the quantization unit is smaller than $\frac{k}{R_{min}}$, will result in not only much larger memory occupation but also sparser scattering of 3D measurement data, which will make the data integration more difficult.

After taking the inverse polar transform, 3D data are stored in an octree data structure according to the three coordinate values, $\frac{1}{\rho}$, θ, and ϕ. Let $\Delta\theta$ and $\Delta\phi$ be the angular resolution of the octree. The octree is created in the spherical coordinate system to maintain the uniform angular resolution, as shown in Figure 2. That is, two 3D points with 3D coordinates $(\rho,\ \theta,\ \phi)$ and $(\rho,\ \theta + \Delta\theta,\ \phi + \Delta\phi)$ will be stored at $(\frac{1}{\rho},\ \theta,\ \phi)$ and $(\frac{1}{\rho},\ \theta + \Delta\theta,\ \phi + \Delta\phi)$, respectively, which shows that the angular resolution will not change after the inverse polar transformation and the quantization into an octree.

3 Automatic 3D Reconstruction

3.1 Visually-Inconsistent Regions

The schematic diagram of our active 3D reconstruction process is shown in Figure 3. We assume that a well-calibrated active binocular head equipped with an accurate position and orientation sensor, such as InterSense IS-900 CT, Fastrak, or Flock of Birds, is available for exploring and reconstructing the 3D environment. That is, the camera parameters of the stereo cameras on the binocular head are known at any time instant, based on the configuration and the kinematic model of the binocular head and the readings of the position and orientation sensors. Hence, we can adopt the asymptotic Bayesian estimation method which assumes the camera parameters are know for each camera position.

To save the reconstruction time, we do not apply the asymptotic Bayesian estimation to an image region unless it is necessary, or more precisely, unless it is a *visually-inconsistent region* (which is defined below). If a set of camera parameters are specified, we can synthesize an images according to the current world model stored in the inverse polar octree. Next, the synthesized image is subtracted from the observed image to obtain a difference image. The difference image is thresholded and then filtered by using the morphological opening to remove noise. The regions contained in the resulted binary image indicate where the depth information are either incorrect or not available, which implies that the depth information in these regions needs to be updated with the observed image. The above regions will be referred to as the visually-inconsistent regions because the observed image is visually inconsistent

with the synthesized image according to the inverse polar octree. At the very beginning, the inverse polar octree contains no valid data. Hence, the whole image is visually-inconsistent, and has to be processed to estimated 3D depth as described in the next subsection. Once some 3D depth estimates are stored in the inverse polar octree, only the visually-inconsistent regions have to be processed.

3.2 Depth Estimation

To estimate the 3D depth, we first partition the visually inconsistent regions into small square blocks, and then assume that each square block in the left image is projected from a 3D planar patch having a constant depth. The depth estimation method we used in this work is mainly the asymptotic Bayesian estimation method[7], which is described briefly below. Suppose that the depth, d, of a square patch, P, in the left image is to be estimated, and that the current estimate of the inverse covariance matrix of d is Φ, which is a one by one matrix (i.e., a scalar) in this case. Instead of directly determining the stereo correspondence using the left and right images, we first move the left camera locally and incrementally in order to compute a rough estimate of the depth of the surface patch (the way we determine the path of the local motion will be described in the next subsection). Now, suppose we obtain a new left image after a local motion. Since the binocular head is well calibrated, we have the relative geometric relation (i.e., the relative camera position and orientation) of the stereo image pair taken by the left camera at the initial and the new positions. Based on this geometric relation, we can compute, for each pixel in the initial image patch, the corresponding image point in the new image if a depth estimate, $\hat{d}$, is given. For convenience, let s be the center of an image patch in the initial reference image and let $u_n(s, \hat{d})$ denote its corresponding image point in the new image i.e., (the nth image), as showed in Figure 4. The depth of the image patch, d, can be refined by minimizing

$$J_n(d) = \frac{1}{2}(d - \hat{d}_{n-1})^t \Phi_{n-1}(d - \hat{d}_{n-1}) + \frac{1}{2}\sum_{s \in P} \left[I_n(u_n(s,d)) - I_1(s)\right]^2, \qquad (2)$$

where $I_1(s)$ and $I_n(u_n(s,d))$ are the intensity value of pixel s in image 1 and the intensity value of pixel $u_n(s,d)$ in image n, respectively, and Φ_{n-1} denotes the inverse covariance matrix of the estimated depth $\hat{d}_{n-1}$ given images 1, 2, ..., $n-1$. The inverse covariance

matrix can be updated by using the following equation:

$$\Phi_n = \Phi_{n-1} + \frac{\partial^2}{\partial d \partial d} \left\{ \frac{1}{2} \sum_{s \in P} \left[I_n(u_n(s,d)) - I_1(s) \right]^2 \right\} \Big|_{d=\hat{d_n}} . \quad (3)$$

For more details, please refer to [7].

The asymptotic Bayesian process for estimating the depth of an image patch P is summarized in the following.

1. Compute $\hat{d_n}$ by minimizing the error function in (2) using a gradient descent method.
2. Update Φ_n with equation (3).

According to our experience and analysis, 50 millimeters of incremental local motion can reduce the depth uncertainty to some extent such that the search region for stereo correspondence is less than 10 pixels in our setup, i.e.,

$$\left| u_n(s, \hat{d_n}) - u_n(s, d_{true}) \right| \leq 5, \quad (4)$$

where d_{true} is the true value of the depth. Therefore, once the effective motion baseline created by the incremental local motion is greater than 50 millimeters, our system will use the image taken by the right camera as the new input image of the asymptotic Bayesian estimation process (i.e., a big jump) and perform an exhaustive search for the minima of (2) in the search region $\left[u_{right}(s, \hat{d_n}) - 5, u_{right}(s, \hat{d_n}) + 5 \right]$, followed by a gradient descent search to further refine the depth estimate. After the depth estimates of all the patches in the visually-inconsistent regions are computed with the above process, Markov random fields can be used to smooth the depth map while preserving the depth discontinuity [9]

3.3 Path Planning for Local Motion

Having a well-calibrated active binocular head, we are able to control the cameras to move along a path which can minimize the ambiguity of stereo matching. Our path planning method is based on the following observation. When performing stereo matching, the stereo correspondence can be determined more easily and reliably if the edge orientation is perpendicular to the epipolar line, as shown in Figure 5. On the other hand, if the edge orientation

is parallel to the epipolar line, then finding stereo correspondence is an ill-posed problem. To eliminate the ambiguity in stereo matching, the local motion is selected to form epipolar lines which are perpendicular to most edges having highly uncertain depth estimates. The following procedure describes the way we determine the local motion:

1. Perform Sobel edge detection on the new input image and record the orientations θ_j of each edge pixel j.

2. For each edge pixel j, get its inverse variance value (i.e., the value of the one by one inverse covariance matrix, Φ_j, determined in the asymptotic Bayesian process). Notice that a larger value of Φ_j indicates that the depth estimate of pixel j is more reliable because $\frac{1}{\Phi_j}$ is the variance of the depth estimate of pixel j.

3. Compute the average edge orientation weighted by its variance value as follows:

$$\theta = \frac{\sum_{j:\Phi_j>0}(\frac{\theta_j}{\Phi_j})}{\sum_{j:\Phi_j>0}(\frac{1}{\Phi_j})}. \tag{5}$$

   Notice that in equation (5), edge orientations corresponding to depth estimates of higher uncertainty will be weighted more heavily.

4. Compute the horizontal and vertical motion components, H_{move} and V_{move}, of the camera:

$$H_{move} = \Delta_H \cos(\theta + \frac{\pi}{2}), \tag{6}$$

   and

$$V_{move} = \Delta_V \sin(\theta + \frac{\pi}{2}), \tag{7}$$

   where Δ_H and Δ_V are two predetermined constants specifying the step size of each movement.

3.4 Consistency Check for a new 3D Observation

By using the asymptotic Bayesian estimation method described in section 3.2, a large amount of 3D points can be determined. However, since the stereo correspondence may contain some false matching, consistency check before integrating new depth estimates into the existing

octree is necessary. For the consistency check, image intensity (or more generally, color information) observed from different viewpoints should be stored for each 3D data contained in the octree. In the consistency check, two questions are asked. Is the new depth estimate of a 3D point "consistent" with the previous observations? If we accept a new 3D depth estimate that is "consistent" with the previous observations, how many voxels in the octree should be removed to maintain the coherence? Our answers to the above two questions will be briefly described below.

Suppose a new depth estimate, whose 3D coordinates are p_{3D}, is obtained at a new point of observation. Let $p_{2D}(n)$ be the image location of p_{3D} on the reference image of the new point of observation, and $p_{2D}(i)$ be the projected 2D image location of p_{3D} on the reference image of the ith point of observation. We say that the new observation p_{3D} is *compatible* with the ith observation if p_{3D} is not occluded by the ith observation (or the other way around) and its color (or graylevel) is compatible with that of the ith observation, i.e.,

$$|I_n(p_{2D}(n)) - I_i(p_{2D}(i))| < \tau_C,$$

where τ_C is a given threshold value. If two third of the previous observations are compatible with the new observation, p_{3D}, then p_{3D} is said to be *largely consistent* with the previous observations and is used to update the inverse polar octree. If the new observation is not largely consistent with the previous observations, it is discarded. For a new observation p_{3D} that is largely consistent with the previous observations, we should also consider whether some old data should be removed from the octree if they are not compatible with the new observation, p_{3D}.

4 Experiments

In our experiment, a well-calibrated binocular head is used to acquire stereo image sequences. The binocular head is mounted on an X-Y table which is used for simulating a mobile robot platform. In this section, we show how a complex scene in our laboratory is reconstructed with the active binocular head. 20 viewpoints for 3D measurement were chosen in advance and the reference images acquired by the left camera are shown in Figure 7. The relative

position of the objects and these 20 viewpoints are shown in Figure 8, where Object 1 is the bookshelf in the background, Object 2 is a textured cardboard and Object 3 is the box. At each viewpoint, a sequence of local movements are performed to estimate the depth value of visually-inconsistent regions by using the asymptotic Bayesian method. Figure 9 shows a typical image sequence acquired along a path of local motion which is determined by the method described in section 3.3. Notice that in Figure 9(c) the inverse variance value increased from left to right as more and more images were acquired and processed. Also, the computed local motion drove the camera to move both vertically and horizontally to reduce the ambiguity of stereo matching. Since the vertical camera motion could not be generated by the X-Y translation table, we moved the tilt joint to generate an equivalent vertical camera motion, which was possible because the lens center of the camera was located a distance off the rotation axis of the tilt joint.

The 3D reconstruction obtained by using the images taken at viewpoints $A1$–$A4$ is shown in Figure 10(a). Notice that when observing from viewpoints $A1$–$A4$, part of the bookshelf, e.g., region R marked in this figure, will be occluded by Object 3. Figure 10(d) shows an image synthesized at a virtual viewpoint, V, which is located between viewpoints A and C, will contain several "holes" (black image regions) because the 3D information of region R are still not valid. When the images observed from viewpoints $A1$–$C4$ were used update the inverse polar octree, the 3D structure become more complete, as shown in Figure 10(b), and the synthesized image based on the latest updated 3D structure looks better (much of the "holes" has been patched), as shown in Figure 10(e). However, after all the images observed at these 20 viewpoints were used to update the inverse polar octree, the 3D structure become more complete, as shown in Figure 10(c), and the synthesized image based on the updated 3D structure looks much better (most of the "holes" has been patched), as shown in Figure 10(f). The overall reconstruction progress using the images from A-1 to E-4 is shown in Figure 11 that the virtual camera is set at a virtual viewpoint U to overlook the scene.

5 Concluding Remarks

We have presented a new approach to reconstruct the 3D environment automatically with an active binocular head. Active vision has been advocated by many researchers such as Bajcsy, Aloimonos, and Ahuja about a decade ago. For example, Alominos has shown that many computer vision problems which are ill-posed, nonlinear and unstable for a passive observer become well-posed, linear and stable for an active observer[10]. Abbott and Ahuja have studied the 3D reconstruction problem using an active stereo vision system[11]. However, most active stereo vision system has been applied to target tracking and no much progress on 3D reconstruction using active stereo has ever been made after Abbott and Ahuja's work mainly because calibrating an active binocular head is much more difficult than calibrating a fixed cameras. We have spent many years on calibrating our binocular head and have achieved very accurate calibration results[6]. Based on our well calibrated binocular head, we have developed an active stereo vision algorithm which can estimate the 3D depth automatically, plan and maneuver a sequence of local movements to reduce the ambiguity in stereo matching, and integrate 3D data obtained in different points of observation. Real experiments have been performed to verify the algorithm proposed in this paper. The experimental results shows that the proposed algorithm is promising. Currently, we are still improving the reliability of this algorithm. Also, we are developing a path planning module which can be used to lead the binocular head to a new point of observation to collect more complete 3D information. Once the binocular head is mounted onto a mobile robot platform, our path planning module can be integrated into the local motion module to guide the mobile robot to move around and acquire information automatically.

Acknowledgements

This work was supported in part by National Science Council of Taiwan, ROC, under Grants NSC 2745-E-001-002.

References

[1] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 353–363, 1993.

[2] S. Kang and R. Szeliski, "3-d scene data recovery using omnidirectional multibaseline stereo," in *Proceeding of IEEE Conference on Computer Vision and Patttern Recogniton, San Francisco, California*, pp. 364–370, 1996.

[3] U. R. Dhond and J. Aggarwal, "Structure from stereo—a review," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 6, pp. 1489–1510, 1989.

[4] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach," in *Proceeding of ACM SIGGRAPH'96*, pp. 11–20, 1996.

[5] Y.-P. Hung, K.-C. Hung, C.-S. Chen, and C.-S. Fuh, "Multi-pass hierarchical stereo matching for generation of digital terrain models from aerial images," *to appear in Machine Vision and Applications*, vol. 11, no. 1, 1998.

[6] S.-W. Shih, Y.-P. Hung, and W.-S. Lin, "Calibration of an active binocular head," *to appear in IEEE Transactions on Systems, Man and Cybernetics*, vol. 28, no. 3, 1998.

[7] Y.-P. Hung, D. B. Cooper, and B. Cernuschi-Frias, "Asymptotic bayesian surface estimation using an image sequence," *International Journal of Computer Vision*, vol. 6, no. 2, pp. 105–132, 1991.

[8] S. D. Blostein and T. S. Huang, "Error analysis in stereo determination of 3-d point positions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 6, pp. 752–765, 1987.

[9] D. B. Cooper, J. Subrahmonia, Y.-P. Hung, and B. Cernuschi-Frias, *The Use of Markov Random Fields in Estimating and Recognizing Objects in 3D Space Markov Random Fields: Theory and Applications, edited by Rama Chellapa and Anil Jain.* Academic Press, 1993.

[10] J. Aloimonos and A. Badyopadhyay, "Active vision," in *Proceesings of the First International Conference on Computer Vision*, pp. 35–54, 1987.

[11] N. Ahuja and L. Abbott, "Active stereo: Integrating disparity, vergence, focus, aperture, and calibration for structure estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1007–1029, 1993.

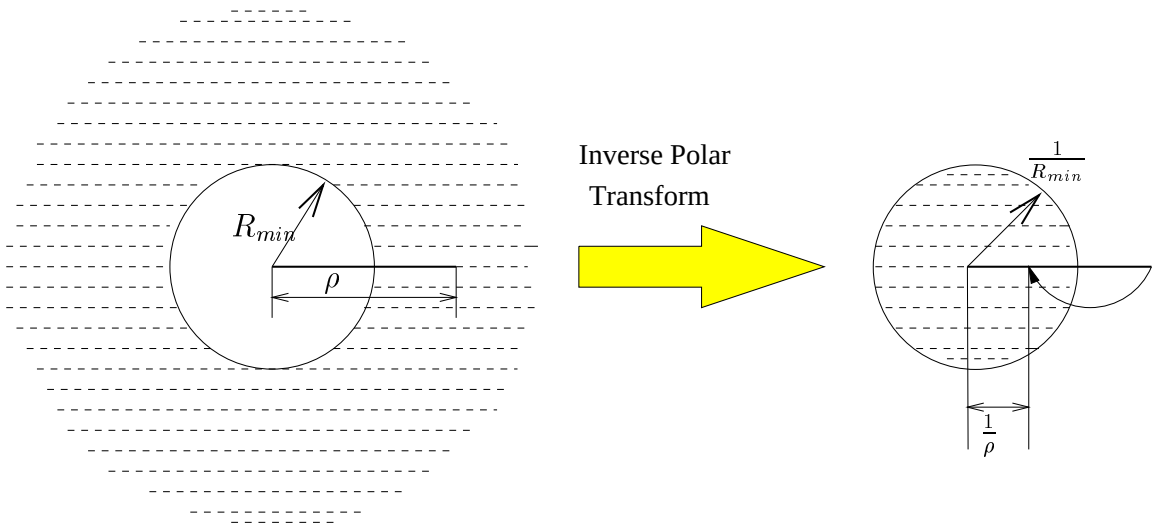

Figure 1: Inverse polar transform.

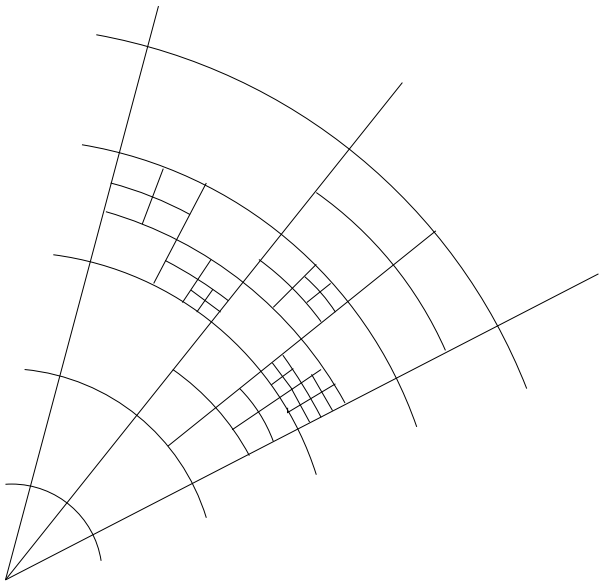

Figure 2: An illustration of the inverse polar octree.

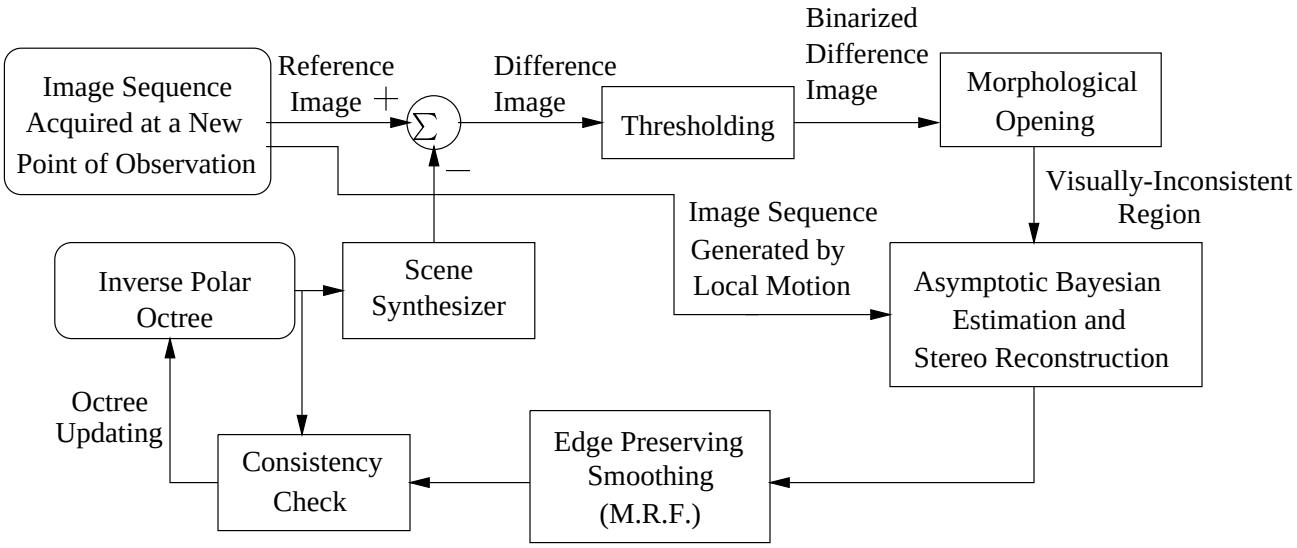

Figure 3: The schematic diagram of the automatic 3D reconstruction process.

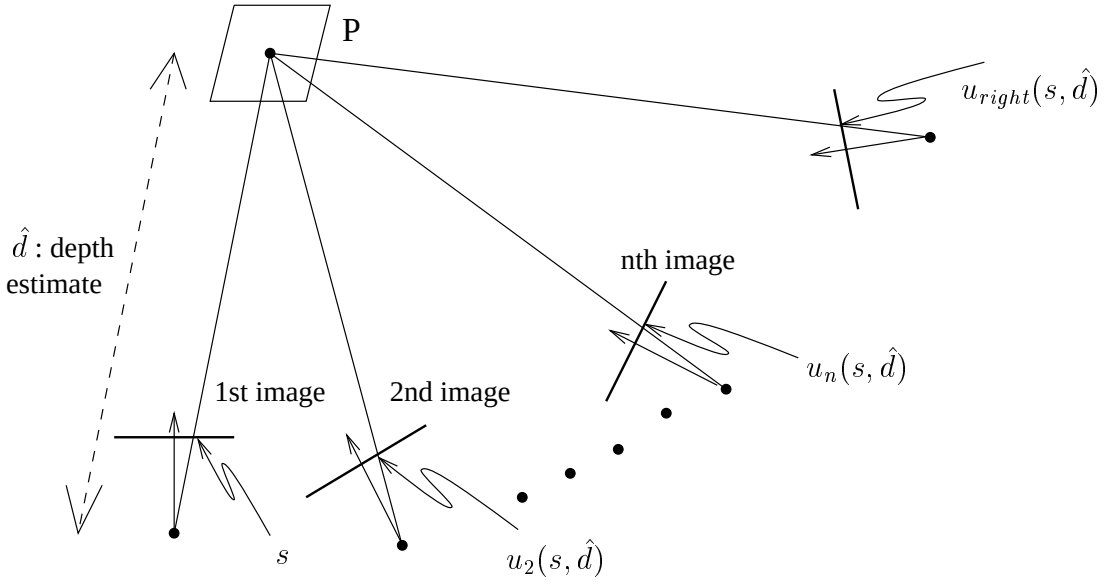

Figure 4: The stereo correspondence of pixel s computed by using the estimated depth $\hat{d}$ and the relative camera geometric parameters.

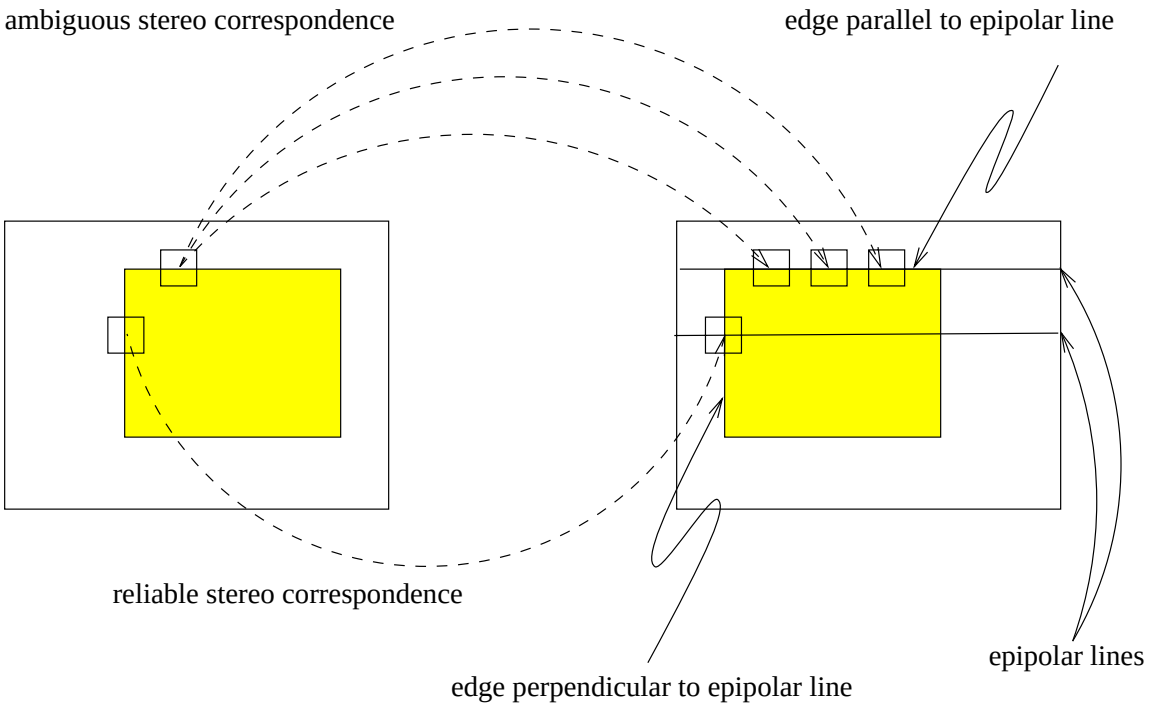

Figure 5: Edge parallel to the epipolar line will cause ambiguity in stereo matching.

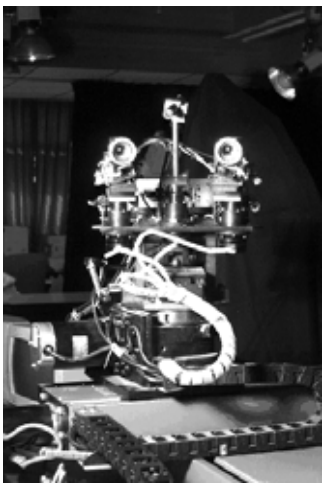

Figure 6: The active binocular head (the IIS head) used in the experiment.

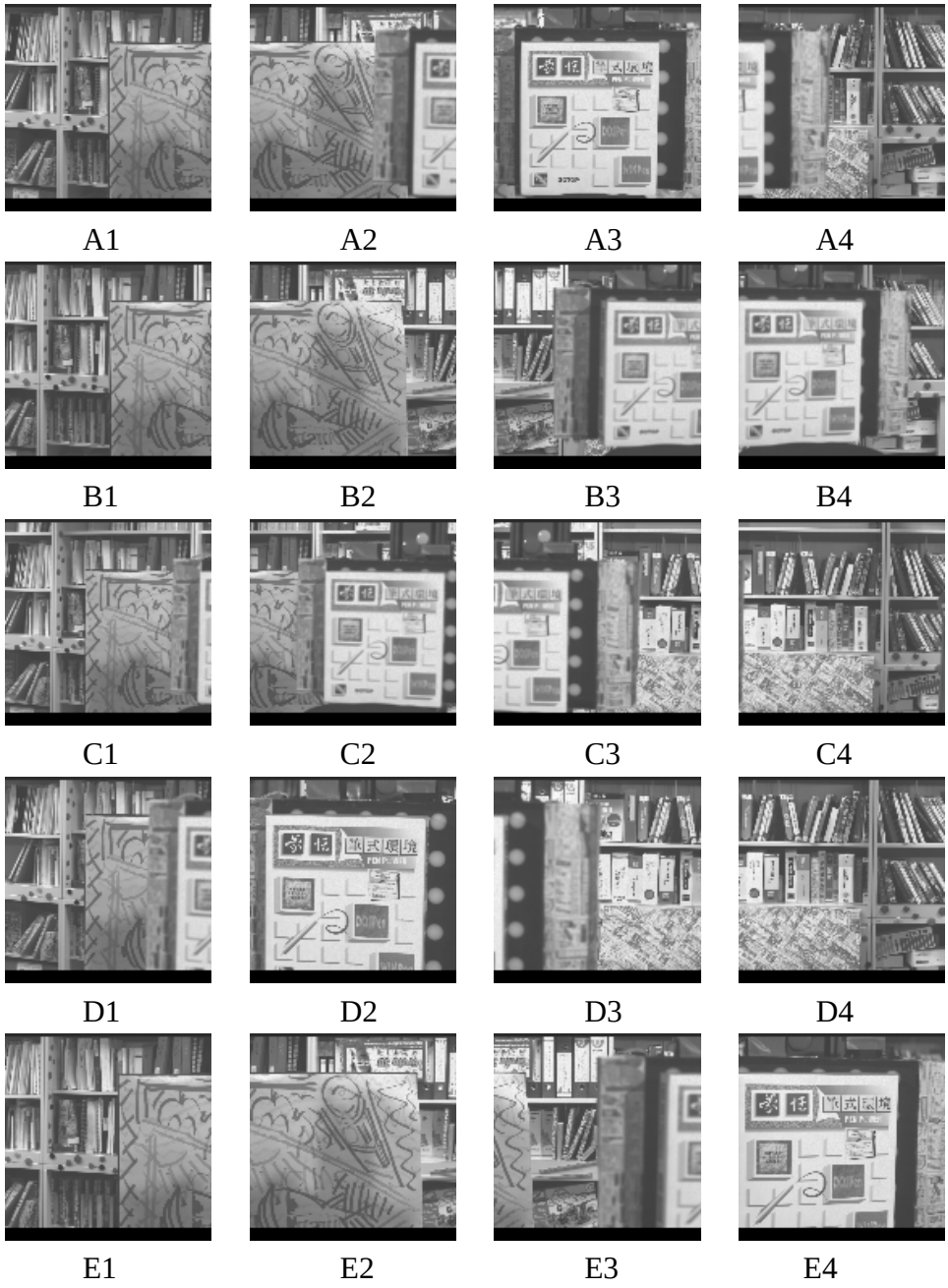

Figure 7: 20 reference images acquired by the left camera at the twelve points of view shown in Figure 8.

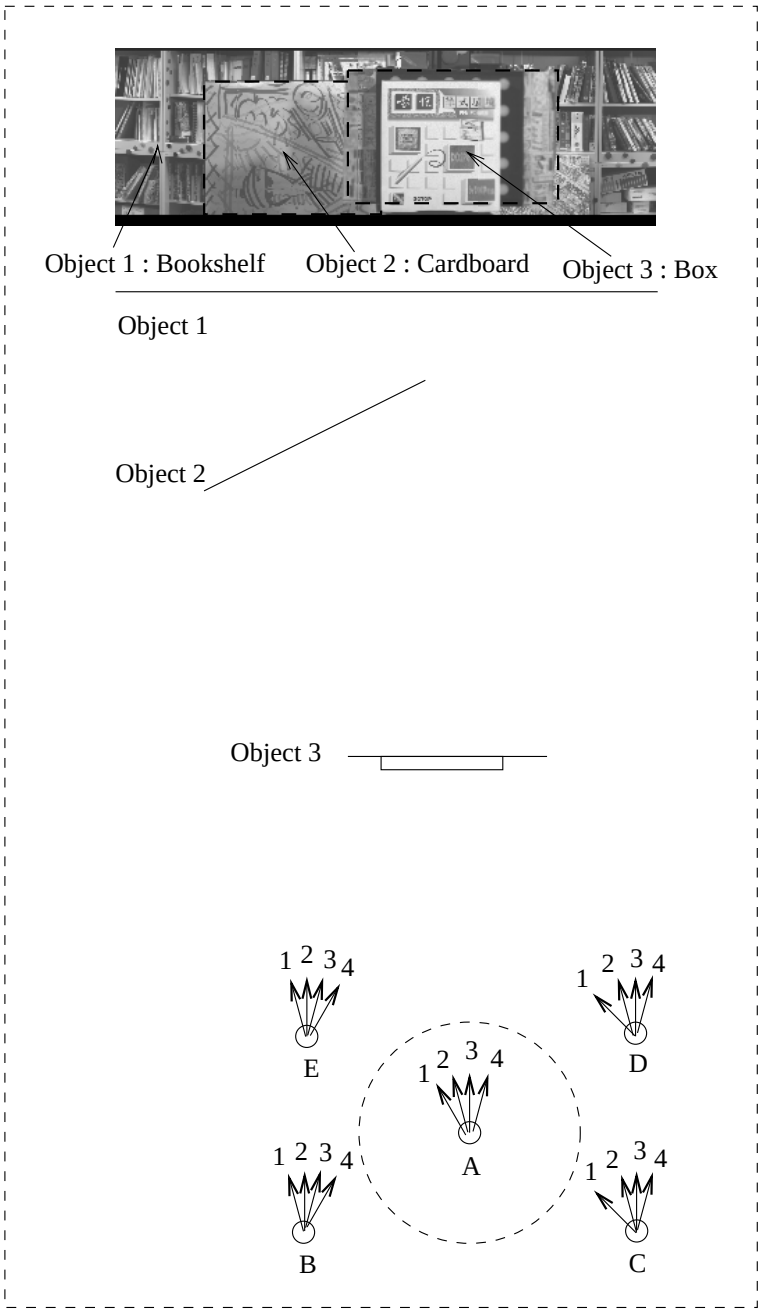

Figure 8: The distribution of the 20 points of view used to acquire the images shown in Figure 7.

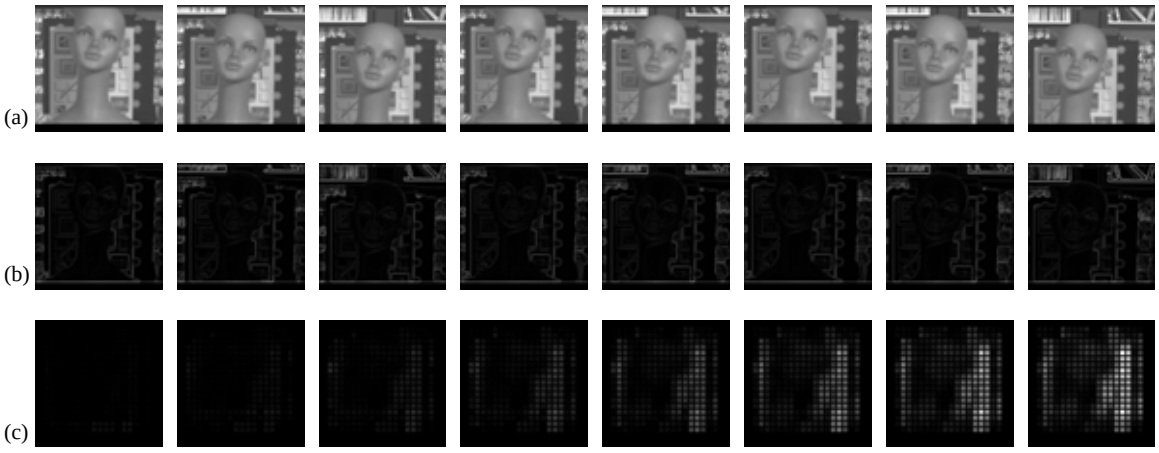

Figure 9: (a) Typical images acquired in a sequence of local motion whose path is determined on line. (b) Sobel edge map for each image in (a). (c) The inverse variance values Φ for each image in (a).

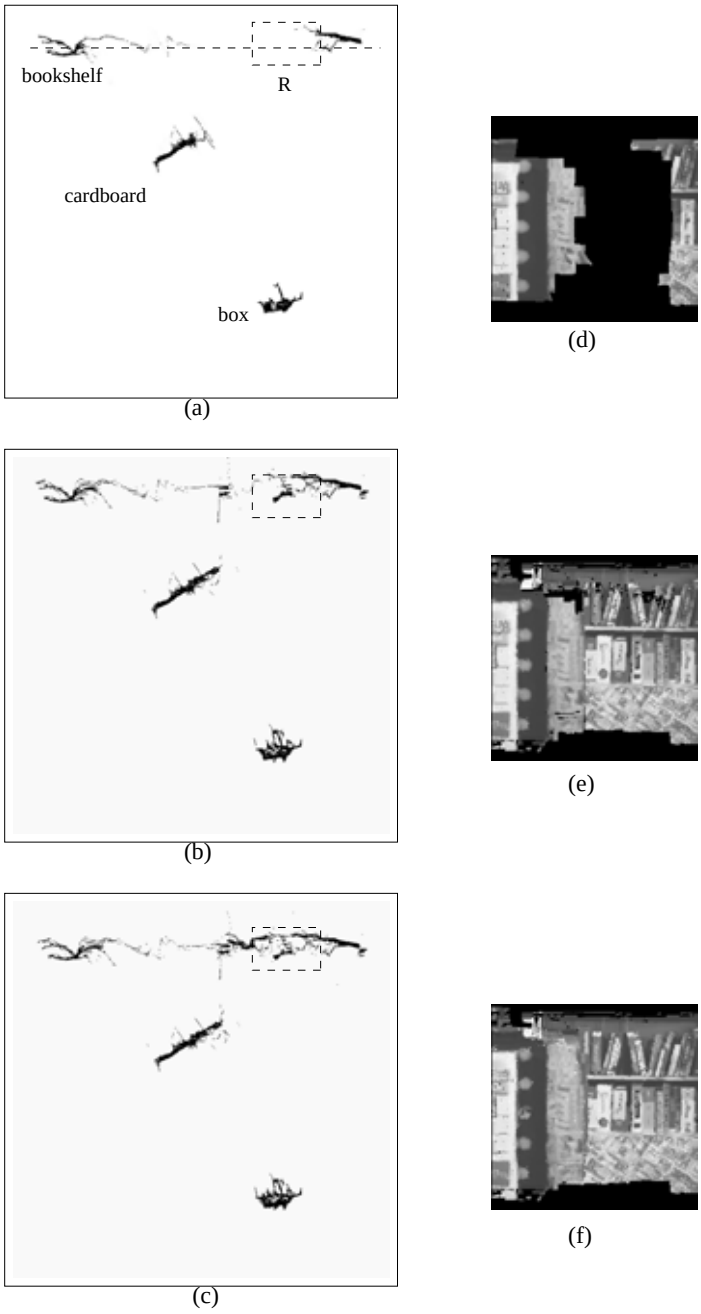

Figure 10: (a) The 3D data contained in the octree reconstructed at viewpoints $A1$–$A4$ were projected to a plane parallel to the ground. (b) The 3D data contained in the octree reconstructed at viewpoints $A1$–$C4$. (c) The 3D data contained in the octree reconstructed at viewpoints $A1$–$E4$. (d) An image synthesized at a virtual viewpoint, V, located between viewpoints A and C by using the octree data shown in (a). (e) An image synthesized at a virtual viewpoint, V, by using the octree data shown in (b). (f) An image synthesized at a virtual viewpoint, V, by using the octree data shown in (c).

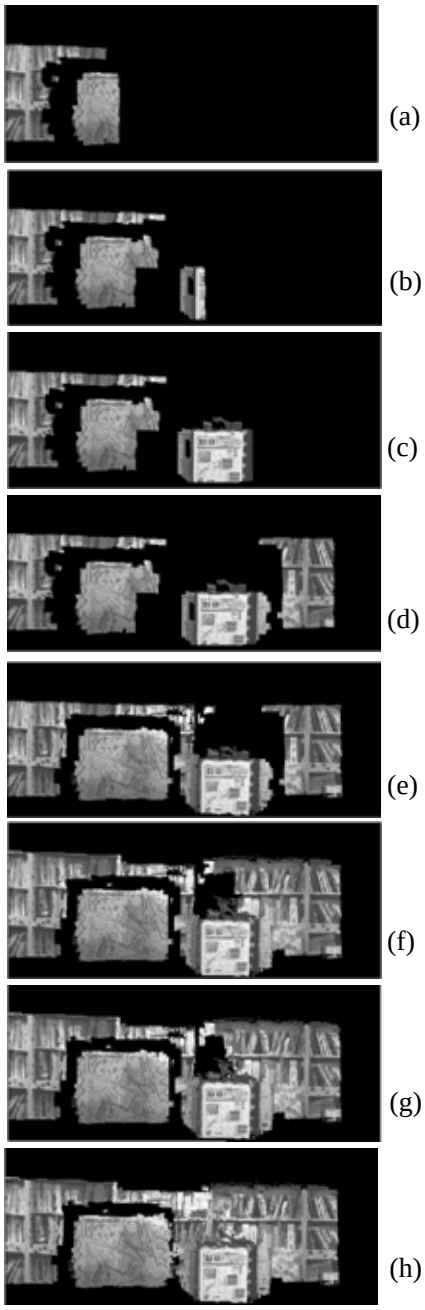

Figure 11: (a) The scene is synthesized at the virtual viewpoint, U, located to overlook the scene by using IPO reconstructed at viewpoints $A1$. (b), (c), (d), (e), (f), (g) and (h) The scene is synthesized at U, by using IPO reconstructed at viewpoints $A1$–$A2$, $A1$–$A3$, $A1$–$A4$, $A1$–$B4$, $A1$–$C4$, $A1$–$D4$ and $A1$–$E4$ respecively.