MVA'98 IAPR Workshop on Machine Vision Applications, Nov. 17-19, 1998, Makuhari, Chiba, Japan

8—16 Acquiring 3D Model of Object by Motion-Stereo

Caihua Wang* Katsuhiko Sakaue†
Machine Understanding Division, ETL
1-1-4 Umezono, Tsukuba-shi, Ibaraki 305 Japan

Abstract

We propose a passive method that acquires the 3D model of an object from a sequence of stereo images in which the object is moved to show all the views of it to a stationary stereo camera. First, we select a few of views of the object in the sequence, and recover the 3D shape of each view using motion-stereo frame, that is, two temporally continued frames of stereo images around the selected view. On the assumption that the background is static and the object is rigid, we can evaluate the reliability of the obtained shapes and thus can select the optimal motion-stereo frame for a view from which the shape of the object can be obtained most reliably, by examining how well the obtained matchings satisfy either rigidity condition or stationary condition. Finally, the partial shapes obtained at the optimal motion-stereo frame of each view are integrated to generate the 3D model of the object. Experimental results on a real object show the effectiveness of the method.

1 Introduction

Acquiring 3D model of an object is an important task in many applications of computer vision. In the passed decades, enormous research efforts were made in this field and there have been various methods proposed to solve the problem. Generally, most of those methods can be classified to two classes of active approach and passive approach. The active approaches, such as range finder with laser strips or other coded light-projecting, can obtain accurate 3D shapes of the object, but they always require special illumination and that the object exists at a close distance, so usually they can only be used in well controlled environment. On the other hand, the passive approaches, such as stereo, shape from shading and shape from motion, can be applied in usual environment, but they always can only achieve relatively lower accuracy and reliability of 3D shapes.

In recent years, due to the increasing demands on acquiring 3D model of an usual object in an usual environment, the passive approaches become attracting, and there are some new passive methods proposed. Shape from counter[1,2] attempted to acquire 3D model of an object by recovery a sequence of surface properties at the extremal edges when the object is rotating. This method requires that the rotation axe can be known accurately and the object is convex. Another kind of passive approaches is to improve traditional passive methods by using multiple input information of traditional passive approaches. As examples, multi-baseline stereo[3,4] used more than two stereo images with different baseline to improve the accuracy and reliability of stereo matchings, and stereo motion[5] utilized the consistency in stereo matching and optical flow to eliminate the unreliable stereo matchings and optical flow computed from a sequence of stereo images taken by a moving stereo camera. However, the objective of those methods is to recover the 3D structure of the scene, rather than to build the 3D model of a specific object.

In this paper, we propose a passive approach to acquire 3D model of an object from a sequence of motion stereo images taken from the object which is moved to show a full view of it to a stationary stereo camera. In the proposed approach, we first select a few of views of the object in the sequence, and recover the 3D shape at each view using two temporally continued frames of stereo images, called motion stereo frames. The views of the object are selected interactively by giving approximately some time points in the sequence. Then the optimal motion stereo frame around each time point are determined by examining how well the correspondences in the four images can be segmented into rigid motion region and static region, assuming that the background is static and the object is rigid. The 3D shape of each view is recovered by a method called motion stereo which utilizes both stereo information and motion information contained from the motion stereo frames. To generate 3D model from the 3D shapes of the selected views, we first estimate initially the relative pose parameters of the 3D shapes using a few initial matching points which can be obtained by tracing, and then refine them using texture matching. In our experiment, initial matching points are given manually. The 3D shapes of the selected views are merged by transforming them to a common coordinate system. When the views are selected such that each surface of the object appears in more than one views, we can extract the overlapped

*JST domestic research fellow, c-wang@etl.go.jp
†E-mail: sakaue@etl.go.jp

surfaces to eliminate some irrelevant parts of shape such as hands which touch to but does not belong to the object. The overlapped surfaces are filtered by a median filter to generate an unique and thin surface. Experimental results on a real object show the effectiveness of the method.

2 Motion Stereo

Motion stereo[6] is a method to obtain the 3D shape of the object from two frames of stereo images where the second frame is taken after the object moved slightly. Using such a motion stereo frame, we can utilize both stereo information and motion information to find more reliable correspondences in the images than just using only one. Here we give a brief description of the method. Figure 1 illustrates the motion stereo model.

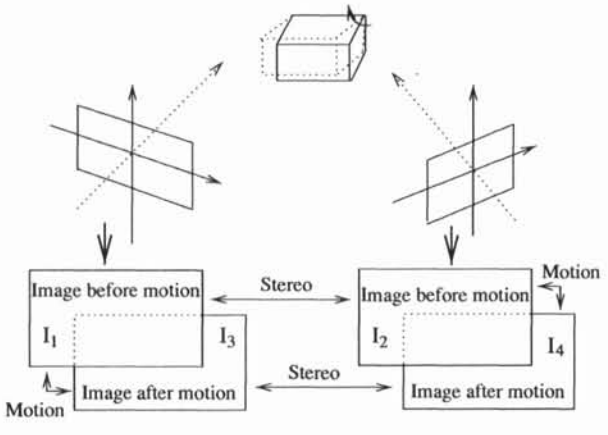


Figure 1: Motion stereo model

The stereo cameras are calibrated and the stereo images are transformed to standard stereo camera setting previously. From stereo image pairs of (I_1, I_2) and (I_3, I_4), we compute two sets of stereo matchings by finding matchings along epipolar lines, using a sub-pixel DP matching technique which takes matching continuity into account[7]. From image pairs of (I_1, I_3) and (I_2, I_4), two sets of optical flow are computed with temporal-spatial gradient technique[8]. From the assumption that the object is rigid and the background is stationary, the following conditions which the stereo matchings and optical flow should obey can be derived.

1. Consistency of stereo matchings and optical flow: for each quadruple of points which are corresponded in the four images, we have

$$v_l = v_r \quad \text{and} \quad u_r - u_l = d_2 - d_1. \qquad (1)$$

where (u_l, v_l) and (u_r, v_r) are the vector of optical flow in the left images and the right images, and d_1 and d_2 are the disparity in the stereo pairs of images before and after motion, respectively.

2. Rigidity of motion and stationariness of background: all the 3D points on the object surface should obey the same motion parameters and the 3D points in the background should be stationary. that is,

$$P'_{obj} = RP_{obj} + T \quad \text{and} \quad P'_{bck} = P_{bck}. \qquad (2)$$

where P'_{obj}, P_{obj}, P'_{bck} and P_{bck} can be computed easily from two sets of stereo matchings. R and T are rotation matrix and translation vector of the object motion.

Under these conditions, we can integrate the stereo matchings and the optical flow, which are obtained from $I_k, k = 1, 2, 3, 4$ individually and may satisfy neither of the above conditions, to generate reliable correspondences among the four images (here I_1 is the base image). This is carried out by the following procedure:

1) Select the correspondences which have obvious motion, nearly satisfy condition 1 and have relative motion rigidity. That is, the correspondences should satisfy: a) $\sqrt{u_l^2 + v_l^2} \geq 0.5$, b) $\sqrt{(v_l - v_r)^2 + (u_r - u_l - d_2 + d_1)^2} \leq 0.1$ and c) $\sum_{k=1}^{n} |d(P_m, P_k) - d(P'_m, P'_k) \leq T_g$. Where P_k and P'_k are the 3D points before and after motion determined by the correspondence, and T_g is a threshold which can be computed automatically with thresholding method such as that proposed by Kittler[9].

2) Estimate the parameters of object motion using the correspondences selected above. This can be done in the 3D space[10] because the 3D coordinates of the correspondences before and after motion can be computed easily from stereo matchings.

3) For each pixel $\mathbf{r}_1 = (x_1, y_1)$, the following four correspondences in the four images with motion and stereo consistency are derived using the stereo matings and the optical flow.

C_1: $\mathbf{r}_1 \rightarrow \mathbf{r}_2 = d(\mathbf{r}_1) \rightarrow \mathbf{r}_4 = f(\mathbf{r}_2) \rightarrow \mathbf{r}_3 = d(\mathbf{r}_4)$

C_2: $\mathbf{r}_1 \rightarrow \mathbf{r}_3 = f(\mathbf{r}_1) \rightarrow \mathbf{r}_4 = d(\mathbf{r}_3) \rightarrow \mathbf{r}_2 = f(\mathbf{r}_4)$

C_3: $\mathbf{r}_3 = f(\mathbf{r}_1) \leftarrow \mathbf{r}_1 \rightarrow \mathbf{r}_2 = d(\mathbf{r}_1) \rightarrow \mathbf{r}_4 = f(\mathbf{r}_2)$

C_4: $\mathbf{r}_2 = d(\mathbf{r}_1) \leftarrow \mathbf{r}_1 \rightarrow \mathbf{r}_3 = f(\mathbf{r}_1) \rightarrow \mathbf{r}_4 = d(\mathbf{r}_3)$

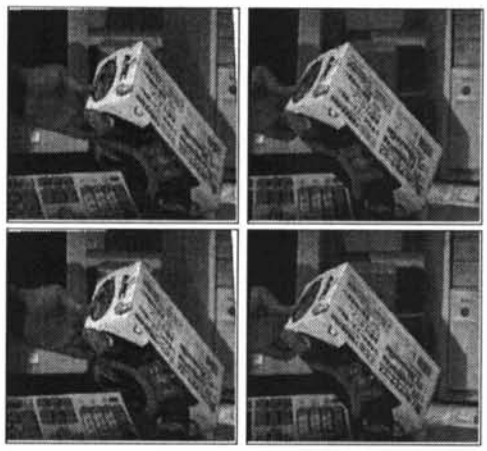
Where $d()$ and $f()$ are pixel mappings by stereo matching and optical flow, and $\mathbf{r}_k = (x_k, y_k), k = 1, 2, 3, 4$ are the coordinate in image I_k.

From above four correspondences with motion stereo consistency, we select the one which has the minimal motion rigidity error as the correspondence at pixel $\mathbf{r}_1$, supposed that $\mathbf{r}_1$ is on the object. As the motion rigidity error at object pixels will be much lower that that at other places, the reliable correspondence of the pixel on the object can be extracted by thresholding the motion rigidity error. Similarly, the the reliable correspondences of background pixels can also be obtained. For the pixels whose correspondence can not be determined reliably, we use the average depth of the reliable pixels in their neighborhood to determine their correspondences.

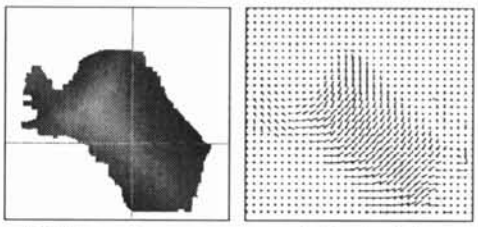
4) The obtained correspondences in four images are represented by Active Net[11], and are refined

to the optimal matchings by minimizing the energy function defined on Active Net.

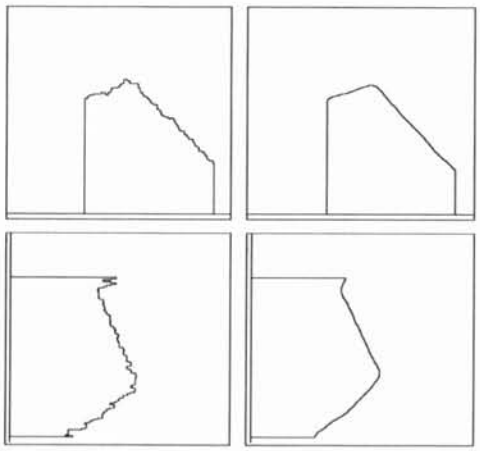
Figure 2 shows the experimental results obtained by motion stereo from a motion stereo frame of a box.



a) Input images



b) Disparity c) Optical flow



d) Shape on the sections in b). left: by stereo, right: by motion stereo.

Figure 2: Experimental results of motion stereo

3 Optimal Motion-Stereo Frames

The above method can be applied to the motion stereo frames at the time points initially selected for the typical views of the object to obtain the 3D shapes. However, due to the difference of object motion in each motion stereo frames, influence of illumination and the mistakes in stereo matching and optical flow, the initial motion stereo frame of a typical view may be not the optimal one from which the 3D shapes can be obtained most reliable and the object can be extracted most easily by motion stereo. Instead, some motion stereo frames around the initial one also contains the same view of the object, and may be more suitable to be used by motion stereo. In this section, we consider how to select the optimal motion stereo frame around the initial view.

On the assumption that the object is rigid and the background is stationary, if the correspondences of the pixels in four images are correct, they must satisfy either rigidity condition of object motion or stationariness condition of background. Therefore, the proportion of the correspondences satisfying either of these two conditions, which can be extracted by the method described in the previous section, can be used to evaluate the reliability of correspondences obtained by motion stereo. The proportion $R_c(F)$ can be computed as:

$$R_c(F) = \frac{P_b(F) + P_r(F) - P_o(F)}{P(F)} \tag{3}$$

Here $P_r(F)$, $P_b(F)$, $P_o(F)$ and $P(F)$ are the numbers of correspondences with motion rigidity, the number of correspondences with stationariness, the number of overlapping correspondences and total number of correspondences at motion stereo frame F, respectively. $R_c(F)$ should be maximized for the optimal motion stereo frame.

On the other hand, in order to extract the object, we expect that all the region of the object has a distinct motion from stationary background. In other words, we expect that correspondences satisfying motion rigidity condition and that with stationaries are exclusive. Therefore, the overlapping proportion $R_o(F)$ of them should be minimized.

$$R_o(F) = \frac{P_o(F)}{P_b(F) + P_r(F) - P_o(F)} \tag{4}$$

Because $R_c(F)$ and $R_o(F)$ may be conflicting objective functions, in experiment we select the optimal frame by

$$\hat{F} = arg \min_{k=1}^{n} \{R_c(F_k) - R_o(F_k)\} \tag{5}$$

Figure 3 shows $R_c(F) - R_o(F)$ computed for 14 motion stereo frames around an indicated time points. We can see that the 10th frame will be the optimal frame.

4 Integrating partial 3D shapes

When the views of the object are selected so that the full view of the object can be composed from them, the 3D model of the object can be obtained by integrating the 3D shapes obtained at each selected view. First, we estimate the relative pose parameters of the 3D shapes using a few initial matching points and refine them using texture matching.

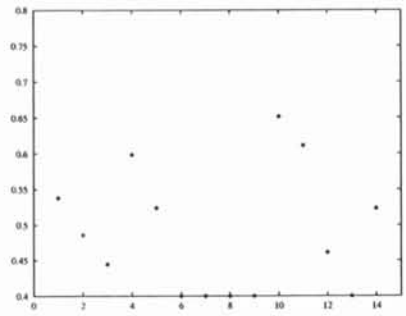


Figure 3: $R_c - R_o$ of continuous motion-stereo frames

Then, the 3D shapes are merged by transforming them to a common coordinate system. Thirdly, we eliminate some irrelevant parts in the 3D shapes such as hands which touch to but not belong to the object, by extracting the overlapped surfaces, supposed that the views are selected such that each surface of the object appears in more than one views. Finally, the overlapped surfaces are filtered by a median filter to generate an unique and thin surface.

The details are described in the following.

4.1 Estimate Relative Pose Parameters

The pose of the object in view V_i relative to that in V_k can be estimated initially from a few of matched points in the images of the two views. Because the 3D information of each point in the images is known, the relative pose can be estimated easily in 3D space[10]. The matched points between images of two views are given manually in our experiment, although they can be obtained in real time by feature tracing during the input of the motion stereo images.

To integrate 3D shapes of N views, we only need to estimate relative pose parameters of $N-1$ pairs of views in which any view must appears at least once. Using those parameters, we can derive the relative pose of any view to a base view, under which the partial shapes are integrated.

When the initial pose parameters are approximate to the correct ones, if we transform shapes to a common coordinate system, the points in the views which stand for the same 3D point of the object will be near to each other. Thus the the pose parameters of each view relative to the base view can be refined iteratively using texture matching as following:

1) For each view V_n, transform the shapes of other views $V_m, m \neq n$ to the pose in view V_i. Then project the texture of the surfaces in each view to xy-plane. Note that for views $V_m, m \neq n$, only the surfaces which are overlapped with that in V_n are projected.

2) For each points $p_m(x, y)$ in the projected plane of V_m, find the matching point $p_n(\hat{x}, \hat{y})$ in the projected plan of V_n by

$$C(p_n(\hat{x}, \hat{y}), p_m(x, y)) = \min\{C(p_n(x+i, y+j), p_m(x, y))\} \quad (6)$$

where $C(p_m, p_n)$ stands for the correlation between p_m and p_n, and $i, j = 0, ..., 5$

3) Treat the points in $V_m, m \neq n$ as reference points, and estimate the parameters of perturbation R' and T' of V_n to transformed $V_m, m \neq n$, using the method described in [10].

4) Correct relative pose parameter of V_n.

$$R_n^{k+1} = R_n^k R' \quad \text{and} \quad T_{k+1} = RT' + T_n^k \quad (7)$$

After parameter refination, the 3D shapes at selected views are merged by transforming them to the base view.

4.2 Extract Object Regions

The 3D shapes obtained by motion stereo may contain some parts such as background region adjacent to the object or the hand which hold the object. When the frames are selected such that each surface of the object appears in more than one views, and that the hands holds the object at different place in each view, we can extract the overlapped surfaces to eliminate most of the the hand and the attached background regions which are randomly distributed in each view. This is done as following: for each 3D point P_i on the 3D shape S_i in view V_i, if no point in the other 3D shapes $S_k, k \neq i$ exists in the neighborhood (a cube of 5mm) of P_i, then P_i is deleted.

Suppose enough number of views are selected, the boundary of the object in one view will appear in the interior regions in some other views. Therefore, we can determine the boundary of the object in a view further accurately, using the the interior regions of other views. For each view V_n, we first apply erosion operation to each views $V_m, m \neq n$ so that they contain no irrelevant part. Then transform $V_m, m \neq n$ to the pose of V_n and project them to V_n. Let I_m be the set of the 3D points of the interior region in V_m, and $P_m(I_m)$ be the set of projected points of I_m in V_n, Then the region of the object in V_n can be obtained by

$$R_n = \bigcup_{m \neq n} P_m(I_m) \quad (8)$$

where $\bigcup$ stands for union of sets.

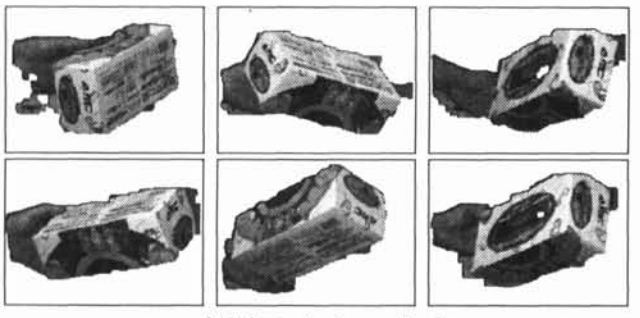
4.3 Thinning the Surfaces

The overlapping surfaces are thinned by applying a median filter in z, x and y direction in order. That is, we first make two depth maps (that is, the z coordinates) on the xy plane, one is for the surfaces which appear in the side of viewer and the other is for the surfaces in the opposite side, by projecting all the 3D points to the xy plane and finding the closest and farthest depths for each 2D point in xy plane. All the 3D points which are not far from closest points are thought as the points in the front surfaces and are filtered by a median filter to generate a thin and unique front surface. The same is done for the surfaces in the opposite side.

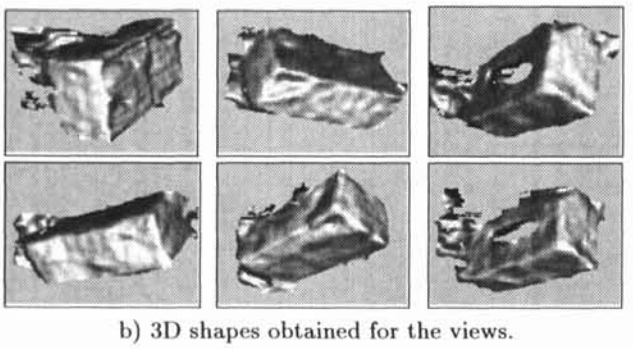
Next, for the part of surfaces which project to the region in the yz plane where no point in the thinned surfaces obtained above is projected on, the above process is also carried out. In other word, the process is only applied on the surfaces which can be see in the x direction but can not be viewed in the z direction. Similar processing is carried out on the xz plane.

5 Experimental Results

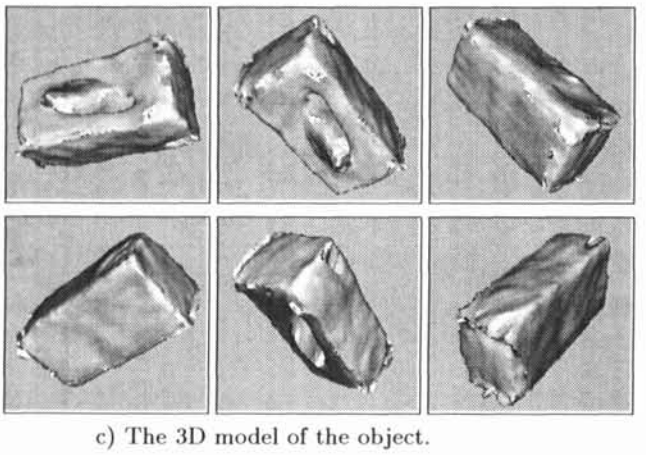
The experimental results of the proposed method are shown in figure 4. Figure 4(a) shows the selected views, and Figure 4(b) shows the 3D shapes obtained by motion stereo method at the optimal frames of each view. The obtain 3D model of the box is shown in Figure 4(c), represented with flat shading.



a) Selected views of a box.



b) 3D shapes obtained for the views.



c) The 3D model of the object.

Figure 4: Object model obtained by motion stereo

6 Conclusion and Future Work

We proposed a passive method that acquires the 3D model of an object from motion-stereo. In the proposed method, we first select a few of views of the object, and recover the 3D shape of each view. Based on the assumption that the background is static and the object is rigid, we can evaluate the reliability of the obtained shapes and thus can select the optimal motion-stereo frame for a view at which the shape of the object can be obtained most reliably. The 3D shape of each view is obtained by motion-stereo method using the optimal motion-stereo frame. Finally, the partial shapes are integrated to build the 3D model of the object.

Future work includes utilizing the obtained 3D model of the object in view-based learning and recognition of the object.

Acknowledgment

This work is carried out under the Real World Computing (RWC) Program.

References

[1] J.Y. Zheng: Acquiring 3-D models from sequences of contours,Trans. on PAMI, Vol 16, pp. 163-178, 1994

[2] W.B. Seales and O.D. Faugeras: Building three-dimensional object models from image sequences, CVIU, Vol. 61, pp. 308–324, 1995

[3] M. Okutomi and T. Kanade: A multiple-baseline stereo, IEEE Trans. on PAMI, Vol. 15, No. 4, pp. 353-363, 1993

[4] Y. Nakamura, T. Matsuura and Y. Ohta: Occlusion detectable stereo—Occlusion patterns in camera matrix, Proc. of CVPR, pp. 371-378, 1996

[5] H. Baker and R. Bolles: Realtime stereo and motion integration for navigation, Proc. of Image Understanding Workshop, pp. 1295-1304, 1994

[6] C. Wang and K. Sakaue: Acquring 3D shape of object by motion stereo, IEICE, Vol.J81-D-II, No.8, pp.1885-1894, 1998(in Japanese)

[7] C. Wang and K. Abe: Stereo matching by integrating piecewise surfaces matched in subranges of depth, Proc. of 13th ICPR A, pp. 451-455, 1996

[8] J. Barron, D. Fleet and S. Beauchemin: Performance of optical flow techniques, IJCV Vol 12, No. 1, pp. 43-77, 1994

[9] J. Kittler and J. Illingworth: Minimum error thresholding, Pattern Recognition, Vol. 19, pp. 41-47, 1986

[10] K. Kanatani: Analysis of 3-D rotation fitting, Trans. PAMI, Vol 16, No. 5, pp. 543-549, 1994

[11] K. Sakaue: "Stereo matching by the combination of genetic algorithm and Active Net", Trans. IEICE, Vol.J77-D-II, No.11, pp.2239-2246, 1994 (in Japanese)