3D Shape and Pose Estimation of Deformable Tapes from Multiple Views

Hitoshi Kubota[1], Masakazu Ono[1], Masami Takeshi[1], and Hideo Saito[2]

[1] Yokohama R&D Center, SUZUKI MOTOR Corporation
Yokohama-City 224-0046 Japan
{kubotah,ono,takem}@hhq.suzuki.co.jp

[2] Department of Information and Computer Science, Keio University
Yokohama-City 223-8522 Japan
saito@ozawa.ics.keio.ac.jp

Abstract

In this paper, we propose a method to estimate 3D shape of deformable plastic tapes from multiple camera images. In this method, the tape is modeled as serial connection of multiple rectangular plates, where the size of each plate is previously known and node angles of between plates represent the shape of the object. The node angles of the object are estimated by 2D silhouette shapes taken in the multiple images. The estimation is performed by minimizing the difference of the silhouette shapes between the input images and synthesized images of the model shape. For demonstrating the proposed method, 3D shape of a tape is estimated with two camera images. The accuracy of the estimation is sufficient for making the assembling robot in our plant to handle the tape. Computation time is also sufficiently short for applying the proposed algorithm in the assembling plant.

1 Introduction

In this paper, we propose a method to estimate 3D shape of deformable tapes from multiple view images. Vision systems are recently used for assembling products, inspection of products, etc. in automated factories. In our assembling plant, some parts are covered with plastic tape for surface protection. The tape must be taken off before the parts are assembled. We aim to build a vision system that can detect the position and shape of the tape for making a robot take off the tape from a assembled parts. Since the shape and pose of the tape is flexible, it is difficult to apply the methods for pose and position detection of rigid object, which are extensively studied for bin picking with robots [5] and so on. If the object is rigid, estimation of rotation and translation (6DOFs) of the object is sufficient to make the robot handle the object as long as the robot previously knows the object shape. However, if the object is non-rigid, not only the object rotation and translation, but also the shape of the object must be estimated in 3D, because the shape changes by the object motion. For reducing the DOFs in recovery of the object shape, 3D shape model

BMVC 2003 doi:10.5244/C.17.14

with some DOFs is generally introduced for non-rigid object shape recovery [1, 4, 6]. In this paper, our objects are deformable plastic tapes. For modeling of such tape shape, we introduce skeleton-based model, in which multiple rectangular planar patch with known width and length are connected at nodes with 3DOFs rotation.

For estimating the rotation angles at the nodes in the model shape, we use multiple cameras so that 3D information of the scene can be captured. In general, point-correspondence between multiple cameras (at least two cameras) is required for obtaining 3D information of the scene. Determination of such pixel-correspondence between different camera images is extensively studied in computer vision research field [3, 9, 10]. However, it is not easy problem especially if there are few features in the images, or a part of the object is occluded. The tape of our object has flat texture surface, so it is very difficult to determine corresponding points between different views. Only silhouette image can be obtained from input images in our case. Therefore, instead of determining corresponding feature pointes between views, we take analysis-by-synthesis based approach, in which model parameters that give closest silhouette model images to the input images are searched.

There are some related researches to the presented paper. Torresani et.al. [11] proposed a method for tracking non-rigid object from image sequence by using rank constraint. This method requires some number of the feature points to be detected and tracked through input image sequence. Smith [7] proposed a method for manipulation planning for non-rigid object based on psychological studies. Although the focus of Smith's research is different from shape and pose estimation of non-rigid object, he suggests one alternative approach for making robot to handle non-rigid object. Sudo et.al. [8] proposed a method for modeling shape and elastic property of a string from images of different pose and shape string. They use a skeleton-based model in which a number of short linear strings are connected at some nodes. Their future application is representation of non-rigid object in computer graphics, rather than handling non-rigid object.

2 Estimation Method

Our aim is to build an intelligent robot system for automatic handling of thin deformable objects such as films and tapes in assembling plants. In this paper, we deal with a robot system, which grasps a tape with bending and twisting deformation, and without elastic deformation. The one side of the tape is fixed and the other freely moves. In order to control a robot to handle the tape, position, pose, and surface normal of the tip of the tape object must be measured. Figure 1 shows the scheme of the system.

We represent the structure of a thin deformable object by skeleton-based model. The model parameters are optimized by an evaluation function calculated as silhouette matching between a 3D model image and an object image.

2.1 3D Modeling

Figure 2 shows the 3D model structure of a thin deformable object. In order to reduce the number of 3D model parameters, we adopt skeleton-based model represented by nodes, trunks and stems. Each trunk is connected to each node on the center of an object. A pair of stems is connected with each node and rotates around trunk-axis. In fact, a pair of stems is represented with curve segments and the curvature changes according to twist

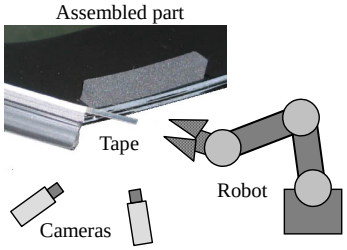


Figure 1: System arrangement.

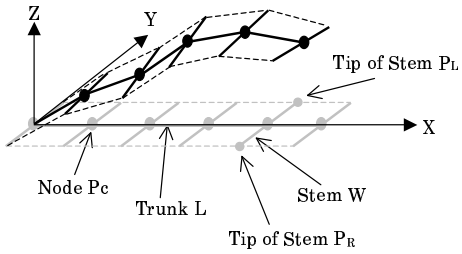


Figure 2: 3D model structure of a thin deformable ojbect.

deformation. In this paper, we assume that twist deformation is small and a pair of stems is linear.

P_{Ci}, P_{Li} and P_{Ri} denote initial coordinates of each node and tips of stems, which are initialized by coordinates value of fixed node P_{C0} and length of each trunk $L_{1\sim n}$ and width of each stem $W_{1\sim n}$. Values of $L_{1\sim n}$ and $W_{1\sim n}$ are initialized manually with object size measured in advance. Coordinates values of each node and tip of stems are transformed by the following translation matrixes:

$$P'(x'_i, y'_i, z'_i)^T = R\theta_{zi} R\theta_{yi} R\theta_{xi} P(x_i, y_i, z_i)^T \tag{1}$$

where θ_{xi}, θ_{yi}, θ_{zi} denote rotation angles about the X, Y, Z-axis. Bending is represented by the rotation matrix $R\theta_{yi}$ and twisting is represented by the rotation matrices $R\theta_{xi}$ and $R\theta_{zi}$. Each rotation matrix is given by:

$$R\theta_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos\theta_x & -sin\theta_x \\ 0 & sin\theta_x & cos\theta_x \end{pmatrix} \tag{2}$$

$$R\theta_y = \begin{pmatrix} cos\theta_y & 0 & sin\theta_y \\ 0 & 1 & 0 \\ -sin\theta_y & 0 & cos\theta_y \end{pmatrix} \tag{3}$$

$$R\theta_z = \begin{pmatrix} cos\theta_z & -sin\theta_z & 0 \\ sin\theta_z & cos\theta_z & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{4}$$

2.2 3D Model Projection and Camera Calibration

We generate a 3D shape model image to make a comparison with an object image. The projection image is represented by linkage of tips of stems whose 2D coordinates are calculated from 3D coordinates using camera parameters. The outline is described on an image plane filled with black, and then pixels inside of the outline are filled with white. It is processed by graphic library functions.

For accurate 3D model projection, it is important to obtain extrinsic and intrinsic parameters of cameras. We adopt Zhang's calibration method [12] to get these camera

parameters. This method calculates all camera parameters from already known 3D points in world coordinates and 2D points coordinates in image coordinates.

When we apply a machine vision system to a robot control system, we need to correspond between world coordinates and robot coordinates exactly. In this paper, we give 3D points using for camera calibration as robot coordinates, and these points are indicated by a cross pattern plate attached to a robot arm.

2.3 Model Optimization

For deformable tapes, it is difficult to extract local features such as edge, vertex and surface, used for reconstruction 3D features, which are occluded by bending and twisting deformation.

The proposed estimation method is based on silhouette image matching between a 3D model image and an object image that does not need to extract of local features. The 3D model parameters are estimated by a multidimensional minimization method with an evaluation function defined by silhouette matching.

The evaluation function for silhouette matching s which denote the similarity between the object image $f(u,v)$ and the model image $g(u,v)$ is defined as:

$$s = \sum_u \sum_v f(u,v) \oplus g(u,v) \tag{5}$$

$$f(u,v) \oplus g(u,v) = \begin{cases} 0, & f(u,v) = g(u,v) \\ const, & f(u,v) \neq g(u,v) \end{cases} \tag{6}$$

where, $\oplus$ denotes exclusive-OR between left and right term. s denotes the sum of these calculation values at each coordinates. The smaller s is, the more similar the two images are.

The evaluation function s is influenced by percentage of the object's area in the image. Then s is normalized by the object's area r_d in the image as:

$$s' = s/r_d \tag{7}$$

In this case, the relation between model parameters θ_x, θ_y,θ_z and the evaluation function s' is represented as a nonlinear programming problem. However, the description of the relation is not available. We use the Downhill Simplex method [2] originated by Nelder and Mead. This method requires only an evaluation function and not of its derivatives. It is simple algorithm and very robust to noise.

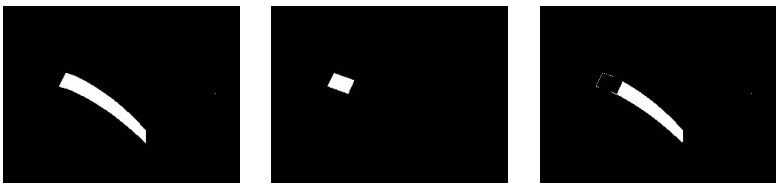
3 Strategy of Robust and High Performance Estimation

3.1 Sequential Optimization

The set of model parameters composed of n nodes is represented by $N(=3n)$ dimensions vector $\vec{V}$ as:

$$\vec{V} = \{\theta_{x1}, \theta_{y1}, \theta_{z1}, \theta_{x2}, \theta_{y2}, \theta_{z2}, \cdots\cdots, \theta_{xn}, \theta_{yn}, \theta_{zn}\} \tag{8}$$

The Downhill Simplex method optimizes the value of parameters (N dimensions) by using the simplex which is a convex with $N+1$ vertices in N dimensional search space.



(a) (b) (c)

Figure 3: Images for Sequential Optimizatin. (a) An object image for node P_1. (b) A result model image of node P_1. (c) An object image for node P_2. (the result model image is deleted from the object image.)

The evaluation values at the vertices of the simplex determine the next searching point of the optimum. This computation is iterated until the evaluation value converges. In each step of the iteration, the positions of the vertices of the simplex are updated. At the first step of the iteration, one of the vertices of the initial simplex is given by the initial guess model parameters $\vec{V}_0$, and the positions for the other vertices are

$$\vec{V}_i = \vec{V}_0 + \lambda_i \vec{e}_i, \tag{9}$$

where the $\vec{e}_i$'s are N unit vectors, and where λ_i's are constant coefficient values for updating.

In order to improve system performance, we take a sequential optimization strategy that optimizes in order of the node number from P_1 to P_n. Only three parameters at one node are optimized in one time. This strategy can reduce the number of updating variables. Then the vector to be optimized can be reduced as:

$$\vec{V}_i' = \{\theta_{xi}, \theta_{yi}, \theta_{zi}\} \tag{10}$$

In the sequential method, each 3D model image doesn't correspond to the object image. Each 3D model image only occupies a part of the object image. Therefore the minimum value of the evaluation function doesn't become zero. Although, there is a tendency that the correspondence of each 3D model to object image is so high that the value of evaluation function becomes smaller. We use the expression 5 - 7 as it is. Before estimation of node P_i, the region of result model image of node P_{i-1} is deleted from the object image. (Figure 3)

3.2 Correspondence of Multiple Images

This method is able to estimate 3D shape of a thin deformable object by a monocular image. Actually, this method can be well adapted to deformation about U-V-axis directions in object image coordinates (X-Z-axis directions in 3D model coordinates). However, it is difficult to estimate accurately in depth direction of a camera.

In order to improve estimation accuracy, we use multiple camera images. In case of using n cameras, the evaluation function is described by the following expression:

$$s' = \sqrt{\left(\frac{s_1}{r_{d1}}\right)^2 + \left(\frac{s_2}{r_{d2}}\right)^2 + \cdots + \left(\frac{s_n}{r_{dn}}\right)^2} \tag{11}$$

where s' denotes the sum of the area of exclusive-OR calculated from multiple images.

This method acquires a result equivalent to stereo matching without extraction and corresponding of local features from object images.

4 Experiments and Discussion

4.1 Experimental results

Table 1 shows experimental specifications. In order to perform the presented estimation method, we used a rectangle tape that flexibly deformed.

Figure 4 shows some of the input images from camera–1 and camera–2 that these poses are set at random. We applied our method to these images in above initial condition. The 3D wire frame model images are overlaid by red line. 2D image points for overlaid are calculated from 3D points in the world coordinates of nodes and tip of stems with estimated parameters θ_x, θ_y, θ_z.

It can be seen that result shapes of the 3D model almost correspond with contour of the sample tape on all images. The pose and shape of the object is successfully estimated even if the feature correspondence between two images cannot easily be taken because of self-occlusion, especially in the images (e) where the left image looks almost a line, the images (g) where the object bends greatly, and the images (h) where the back side appears by twisting.

Figure 5 shows results of 3D reconstruction for the case of the images (e) and (h) using the usual stereo-matching method which search of corresponding features between two images. By bending and twisting deformation, it cannot reconstruct 3D shape and pose of the object. In the image (e), it cannot correspond vertexes or shapes in the tip of the object. In the image (h), the 3D reconstruction result is incorrect that describes crossed shape. Alternatively, the presented method can reconstructs 3D shape and pose of these images. This result shows that the presented method is applicable for occluded conditions that cannot obtain feature correspondence.

To investigate the result further, we show error between true coordinates data and estimated coordinates data of node P_5 located in the tip of the object by following expression:

Camera	Number	2
	Resolution	W640 $\times$ H480pixel
	View	W280 $\times$ H210mm
Distance	cameras — tape	600mm
	camera–1 — camera–2	400mm
Tape	L150 $\times$ W30mm	
Model	Nodes	6（P_0 is fixed）
	Trunks	L=30mm
	Stems	L=15mm
Computer	Pentium4 3.0GHz	

Table 1: Experimental specifications.

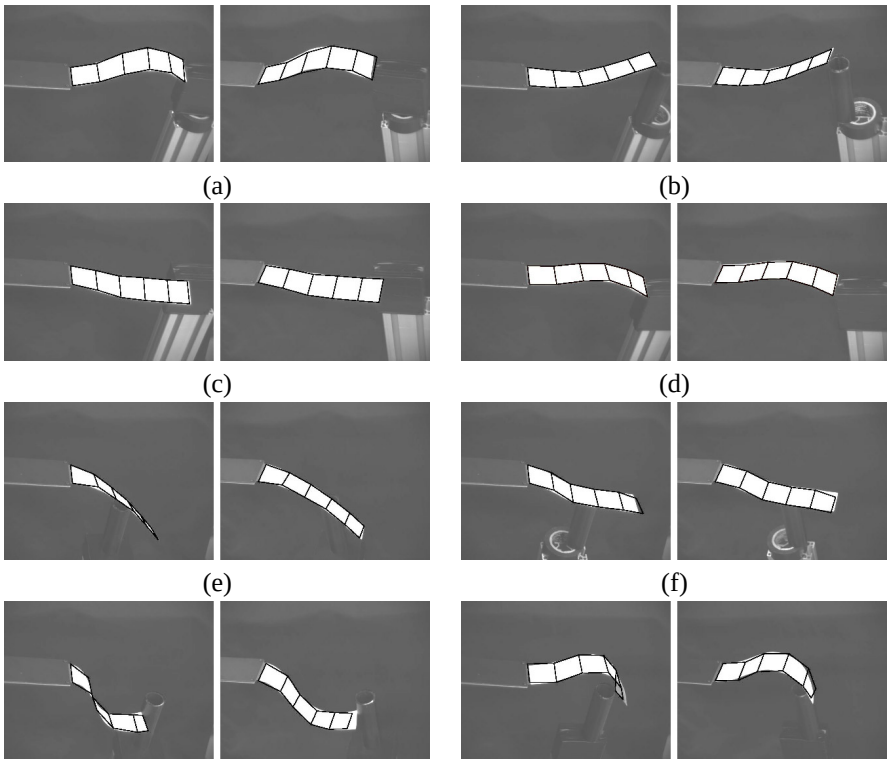


Figure 4: The input and result images. Estimation result is drawn on the input images from camera–1(left) and camera–2(right).

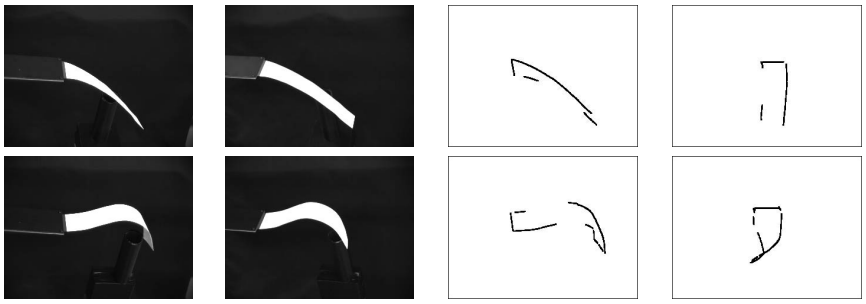


Figure 5: Resutls of 3D reconstruction by the usual method. (corresponding features with stereo camera)

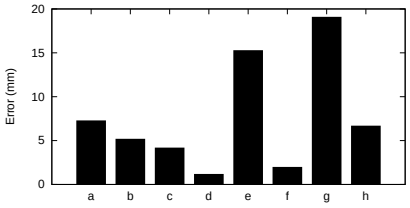


Figure 6: Error between true data and estimated data.

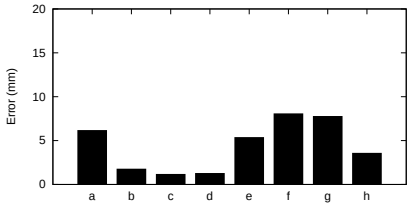


Figure 7: Error between true data and estimated data. (Re-calculation)

$$e = \sqrt{(x_e - x_t)^2 + (y_e - y_t)^2 + (z_e - z_t)^2} \tag{12}$$

where (x_e, y_e, z_e) denotes an estimated coordinates, and (x_t, y_t, z_t) denotes a true coordinates of node P_5 (Figure 6). We acquired a true coordinates from the robot controller when directing with the thin needle attached at the tip of the robot arm. The error is within the limited of 20mm in all images.

On a multidimensional minimization problem, a solution may stagnate in a local minimum. In order to escape from a local minimum, it is shown validity of re-initialization by a claimed minimum and re-calculation. In the Downhill Simplex method, we reinitialize N of the $N+1$ vertices of the simplex again by the expression 9, with P_0 being one of the vertices of the claimed minimum. Figure 7 shows the error by re-calculation. The error is within the limited of 10mm in all images. This result shows that it is efficient to calculate twice for high accuracy. However, it takes twice calculation time. We need to decide estimation times or model parameters according to a specification of a target system.

In this experiment, it took approximately one second for estimation on a Pentium4–3.0GHz machine. This is enough to use this method for a robot control system in assembling plants.

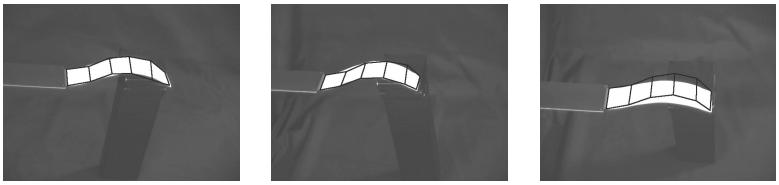
4.2 Estimation with Multiple Views

We tried estimation by using three cameras. One more camera (camera–3) is set about 400mm above camera–1. Figure 8 shows result images by using two cameras (camera–1,2) and three cameras (camera–1,2,3). In the result by using two cameras, the obtained 3D shapes are almost same with the tape images on camera–1 and camera–2. The projected shape on the camera–3 image doesn't fit to the tape image. However, in the result by using three cameras, the obtained 3D shapes fit to the tape image on all camera images.

Figure 9 shows error between true coordinates data and estimated coordinates data in each node. The error decreases in the node P_2 to P_4 by increasing the number of cameras. This result shoes that this method can estimate over the whole tape by increasing the number of cameras.

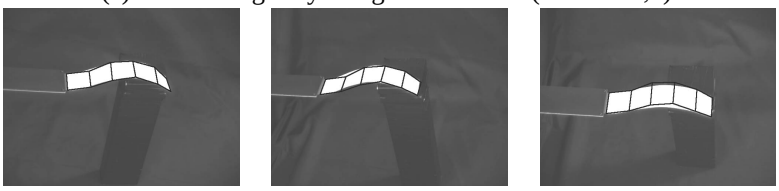
5 Conclusion

We have proposed an original and efficient method for 3D shape and pose estimation of thin deformable objects like a plastic tape. For decreasing the number of model parame-



camera–1 camera–2 camera–3

(a) Result images by using two cameras (camera–1,2).



camera–1 camera–2 camera–3

(b) Result images by using three cameras (camera–1,2,3).

Figure 8: Result images by using two cameras and three cameras.

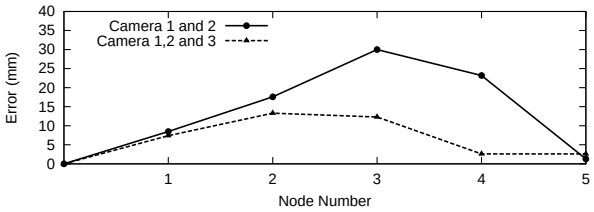


Figure 9: Error between true data and estimated data in each node.

ters, we represented deformation of the object by nodes, trunks and stems rotation. The model parameters are optimized by an evaluation function that is calculated as a silhouette matching area between an input image and a synthesized image of a model shape. Accurate 3D shape estimation is achieved using the multiple camera images and iterative optimization.

The presented experimental results show that our method is able to estimate deformation accurately and robust to occlusions in the situation that feature correspondence cannot be taken between camera images. Further, our method is sufficient for making a robot control system in assembling plants.

References

[1] Douglas DeCarlo and Dimitris Metaxas, "Deformable Model-Based Shape and Motion Analysis from Images using Motion Residual Error," *ICCV'98*.

[2] J.A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, 7, pp.308-313, 1965.

[3] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4), pp.353-63, 1993.

[4] A. Pentland and B. Horowitz, "Recovery of Nonrigid Motion and Structure," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pp.730-742, July, 1991.

[5] K. Rahardja and A. Kosaka, "Vision-Based Bin-Picking: Recognition and Localization of Multiple Complex Objects Using Simple Visual Cues," *Proceedings of IROS'96, Osaka, Japan, IEEE Press*, 1996.

[6] Yoshinobu Sato, Masamitsu Moriyama, Masayuki Hanayama, Hiroaki Naito and Shinich Tamura, "Acquiring 3D Models of Non-Rigid Moving Objects From Time and Viewpoint Varying Image Sequences: A Step Toward Left Ventricle Recovery," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.19, No.3, pp.253-259, March 1997.

[7] P.W. Smith, "Image-Based Manipulation Planning for Non-Rigid Objects," *Proceedings of the IEEE International Conference on Robotics and Automation, Leuven, Belgium*, May 1998.

[8] Katsuhito Sudoh, Koh Kakusho and Michihiko Minoh, "Modeling a String from Observing the Real Object," *Proceedings of 6th International Conference on Virtual Systems and Multimedia*, pp.544-553.

[9] Yasushi Sumi, Yoshihiro Kawai, Takashi Yoshimi and Fumiaki Tomita, "3D Object Recognition in Cluttered Environments by Segment-Based Stereo Vision," *International Journal of Computer Vision*, Vol.46, Issue.1, pp.5-23, January 2002.

[10] P.H.S. Torr and A. Criminisi, "Dense Stereo Using Pivoted Dynamic," *BMVC2002*, pp.414-423, 2002.

[11] Lorenzo Torresani, Danny Yang, Gene Alexander and Christoph Bregler, "Tracking and Modeling Non-Rigid Objects with Rank Constraints," *Proc. IEEE CVPR 2001*.

[12] Zhengyou Zhang, "Flexible Camera Calibration by Viewing a Plane from Unknown Orientations," *ICCV'99*, Vol.1, pp.666-673, 1999.