New Generation Computing, 18(2000)127-135
Ohmsha, Ltd. and Springer-Verlag

NEW GENERATION COMPUTING

View-based Virtual Learning and Recognition of 3D Object Using View Model Obtained by Motion-Stereo

Caihua WANG† and Katsuhiko SAKAUE
Real World Intelligence Center, Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba-shi, Ibaraki 305-8568 JAPAN
† *JST domestic research fellow*
c-wang,sakaue@etl.go.jp

Received 31 August 1999
Revised manuscript received 12 October 1999

Abstract View-based approach for learning and recognition of 3D object and its pose detection was proved to be affective and efficient, except its high learning cost. In this research, we propose a virtual learning approach which generates learning samples of views of an object from its 3D view model obtained by motion-stereo method. From the generated learning sample views, features of high-order autocorrelation are extracted, and discriminant feature spaces for object recognition and pose detection are built. Recognition experiments on real objects are carried out to show the effectiveness of the proposed method.

Keywords: 3D View Model, View-based Virtual Learning, High-order Autocorrelation, Object Recognition, Pose Detection.

§1 Introduction

Object recognition is a major task in computer vision. In the passed decades, enormous research efforts were made in this field and there were various methods[1~6] proposed. Generally, most of those methods can be classified into two classes of model-based approach[1~3] and view-based approach.[4~6] In the model-based approach, object recognition is carried out by matching the geometric features extracted from the image of object with its geometric model. This approach requests the geometric model of the object to be available in advance, but in general, building a geometric model of an object is not an easy task, especially when the object has complex shape. Moreover, fault in extraction of geometric features from the images may cause failure in recognition.

On the other hand, the view-based methods take a visual learning approach to learn 3D object from its two-dimensional views. Murase et al.[4] proposed a method called parametric eigenspace for object learning. In this method, a sequence of views of the object, which are taken when the 3D object is turned around on a turn table, is compressed to its eigenspace by K-L expansion. The trajectory of the view sequence in the eigenspace is treated as a appearance model of the object. Object recognition is carried out by projecting the image to be recognized to the eigenspace and finding the closest point in the trajectory. Hasegawa et al.[5] also adopt a similar learning procedure, but they represented the images with high-order autocorrelation features[6] and constructed the feature space for recognition by discriminant analysis.

On contrast to model-based approach, view-based methods do not need geometric models, so they can be applied effectively in spite of complexity of object shapes. However, in order to deal with 3D object appears at arbitrary pose under arbitrary illumination, numerous learning samples are required, so the learning cost will be very high. With regard to this problem, Amano et al.[7] proposed a method which generates the learning samples of range images of an object from its 3D model, where the range data of the object is utilized for object recognition.

In this research, we propose a view-based virtual learning approach which generates learning samples of views of object from its 3D view model obtained by motion-stereo method. In motion-stereo, some views of the object are shown to a static stereo camera by moving the object slightly. The shapes of the object at each view are reconstructed by motion-stereo which utilizes both stereo and motion information. The obtained shapes are integrated to form a 3D view model of the object.

From the 3D view model, we can generate views of the object observed from any view point. In our research, we select 110 view points which uniformly distribute on the spherical surface as the representative poses of the object. For each pose of the object, learning samples of view are generated and the features of high-order autocorrelation are extracted to build discriminant feature spaces for object recognition and pose detection. In order to suppress the influence of illumination, we use directional high-order autocorrelations defined on Gaussian directional differential. With these features, the rotation of images can be represented as a linear transformation in the feature space, so we can deduce one degree of freedom of rotation angle. Finally, recognition experiments on real objects are carried out to show the effectiveness of the proposed method.

§2 3D View Model from Motion-Stereo

Motion stereo[8] is a method to obtain the 3D view model of the object in general environment. Fig. 1 briefly illustrates the procedure.

In motion stereo, the object is moved to show some views of it to a static stereo camera. At each view, two frames of stereo images which are temporally continued are use to recovery the shape of the object. Using such two temporally continued stereo frames, we can utilize both stereo information and motion

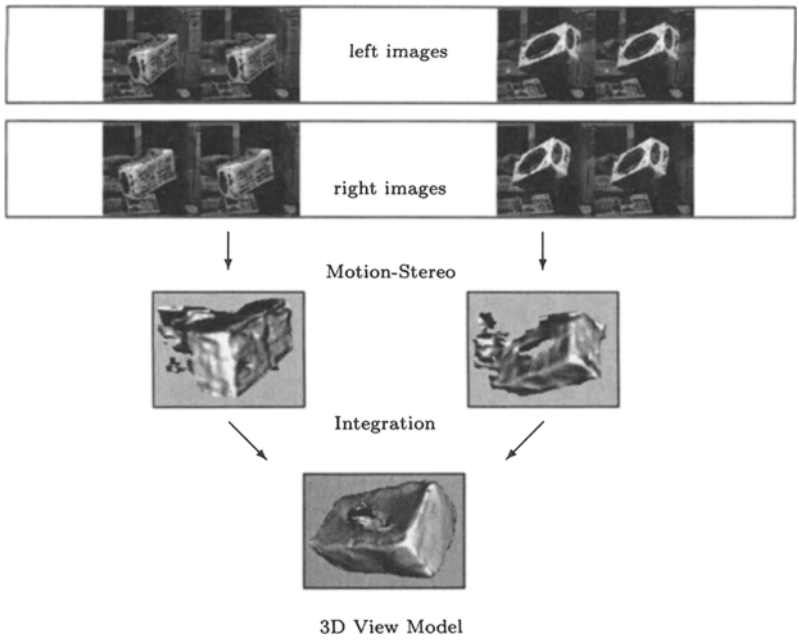

Fig. 1 Obtain 3D View Model from Motion-Stereo

information to find more reliable correspondences in the images than just using only one. On the assumption that the background is static and the object is rigid, we can extract the object region easy by checking the condition of rigidity of motion.

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 partial shapes obtained at the selected views. First, we estimate the relative pose parameters of the partial shapes using a few initial matching points and refine them using texture matching. Then, the partial shapes are merged by transforming them to a common coordinate system. Thirdly, we eliminate some irrelevant parts in the partial 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 unique and thin surface and the 3D view model of the object is obtained.

§3 Generation of Learning Samples

The 3D view model obtained by motion stereo method contains both shape and texture of the object, so the view of any view-point can be generated using this model. In this research, in order to deal with any views of the object, we generate view points which uniformly distribute on the spherical surface with enough density, and regard the views from those view points as representative

poses of the object.

Fig. 2 briefly illustrates how to generate such a view point distribution. First, we generate an initial view point distribution using 8 vertex of the inscribed cube of a sphere and 6 view points which pass through the centers of six faces of the cube. The initial spherical surface is stretched with triangle patches using over the 14 view points, as shown in Fig. 2 (a). View point distribution with any density can be generated by iteratively adding new view points using the centers of triangle patches and stretching the sphere surface again, as shown in Fig. 2 (b) and 2 (c). In this research, we select 110 view points shown in Fig. 2 (c) as the representative poses of the object. Fig. 2 (e) shows the views generated for some poses from the 3D view models of two objects shown in Fig. 2 (d).

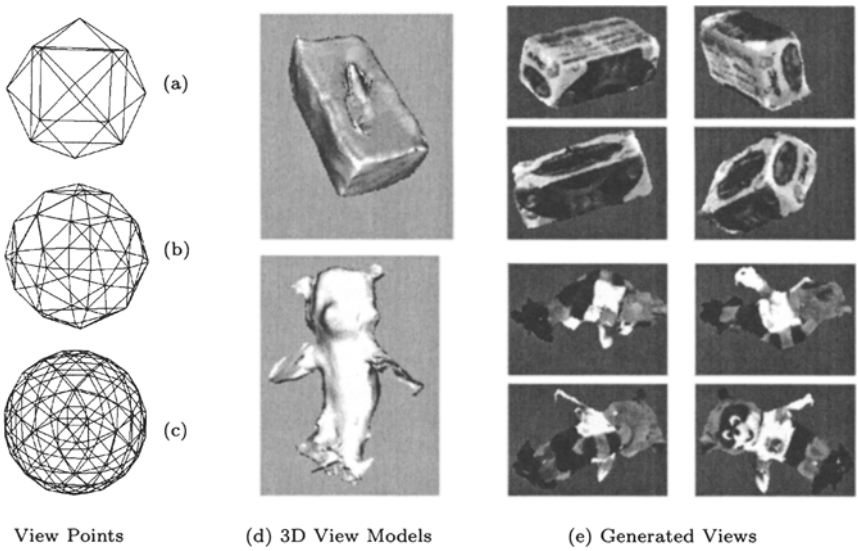

Fig. 2 Some Sample Views Generated from 3D View Models

For each pose, we select some view points close to it and generate views as its learning samples. In experiment, we select 24 viewpoints uniformly distributed on a circle around view point of the pose on the spherical surfaces.

§4 High-order Directional Autocorrelation

In this research, we adopt the high-order autocorrelation features proposed by Otsu et al.[6] for image representation, because this representation has the advantages of shift invariance and simplicity in computation. The definition of high-order autocorrelation is modified by using Steerable Gaussian Differential Filter, that is, the high-order autocorrelation is defined on Gaussian directional differentials of the image rather than on its grey value. With the modified definition, the influence of illuminations can be suppressed, and because the steerability of Gaussian differential filter, the rotation of images can be represented as a linear transformation in the feature space, so we can deduce

one degree of freedom of rotation angle.

4.1 Steerable Gaussian Differential Filter

Gaussian differential filter of arbitrary orientation can be represented by a linear combination of a set of basis filters in particular orientations. The filter with such property is called steerable filter. Steerable Gaussian differential filter is described briefly as following.

Given a Gaussian function $G(x,y) = \frac{1}{2\pi\sigma^2} exp(-\frac{x^2+y^2}{2\sigma^2})$, the 1-order differential filter in x direction (0°) and y direction (90°) are as following.

$$G_1^{0^\circ} = \frac{\partial}{\partial x}G(x,y) = -\frac{1}{2\pi\sigma^4} x \exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{1}$$

$$G_1^{90^\circ} = \frac{\partial}{\partial y}G(x,y) = -\frac{1}{2\pi\sigma^4} y \exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{2}$$

The first differential filter of Gaussian $G_1^{\theta^\circ}$ in direction θ can be represented linearly by $G_1^{0^\circ}$ and $G_1^{90^\circ}$ as following.

$$G_1^{\theta} = \cos(\theta) G_1^{0^\circ} + \sin(\theta) G_1^{90^\circ} \tag{3}$$

Similarly, the second differential filter of Gaussian G_2^{θ} in direction θ can be represented by a linear combination of that of $\theta_1 = 0°$, $\theta_2 = 60°$ and $\theta_3 = 120°$. That is,

$$G_2^{\theta} = k_1(\theta) G_2^{0^\circ} + k_2(\theta) G_2^{60^\circ} + k_3(\theta) G_2^{120^\circ} \tag{4}$$

where

$$k_j(\theta) = \frac{1}{3}(1 + 2\cos(2(\theta - \theta_j))), \qquad j = 1,2,3 \tag{5}$$

In general, nth differential Gaussian filter of arbitrary direction can represented as a linear combination of a set of basis filters. In this research, we only use first and second differential filters of Gaussian.

4.2 High-order Directional Autocorrelation Features

For image $f(x,y)$, its nth differential filter of Gaussian F_n^{θ} in direction θ can be represented as

$$F_n^{\theta}(x,y) = I(x,y) * G_n^{\theta}(x,y) \tag{6}$$

where $*$ stands for convolution operator.

The high-order directional autocorrelation features are defined on filtered images $F_n^{\theta}(x,y)$ as following.

$$R(F_{n_1}^{\theta_1}, F_{n_2}^{\theta_2}, ..., F_{n_m}^{\theta_m}) = \iint F_{n_1}^{\theta_1}(x,y) F_{n_2}^{\theta_2}(x,y) ... F_{n_m}^{\theta_m}(x,y) dxdy \tag{7}$$

Using the distributivity of the convolution operator and the steerability of $G_n^{\theta}(x,y)$, we can show that any given $R(F_1^{\theta_1}, F_1^{\theta_2})$ can be represented as a linear combination of $K(R(R(F_1^{\theta_1}, F_1^{\theta_2})) = \{R(F_1^{0^\circ}, F_1^{0^\circ}), R(F_1^{0^\circ}, F_1^{90^\circ}), R(F_1^{90^\circ}, F_1^{90^\circ})\}$, which is called the basis features of $R(F_1^{\theta_1}, F_1^{\theta_2})$. The prove is omitted here.

Similarly as above, we can derive the basis features for other form of high-order directional autocorrelation features like $R(F_{n_1}^{\theta_1}, F_{n_2}^{\theta_2}, ..., F_{n_m}^{\theta_m})$. Here we give the basis features of second-order and third-order directional autocorrelation features with first and second differential filters of Gaussian.

$$
\begin{aligned}
K(R(F_1^{\theta_1}, F_2^{\theta_2})) &= \{R(F_1^{b_1}, F_2^{b_2}),\ b_1 = 0°, 90°;\ b_2 = 0°, 60°, 120°\} \\
K(R(F_2^{\theta_1}, F_2^{\theta_2})) &= \{R(F_2^{b_1}, F_2^{b_2}), b_1, b_2 = 0°, 60°, 120°; b_1 \le b_2\} \\
K(R(F_1^{\theta_1}, F_1^{\theta_2}, F_1^{\theta_3})) &= \{R(F_1^{b_1}, F_1^{b_2}, F_1^{b_3}), b_1, b_2, b_3 = 0°, 90°; \\
&\qquad b_1 \le b_2 \le b_3\} \\
K(R(F_1^{\theta_1}, F_1^{\theta_2}, F_2^{\theta_3})) &= \{R(F_1^{b_1}, F_1^{b_2}, F_2^{b_3}), b_1, b_2 = 0°, 90°; \\
&\qquad b_3 = 0°, 60°, 120°;\ b_1 \le b_2\} \\
K(R(F_1^{\theta_1}, F_2^{\theta_2}, F_2^{\theta_3})) &= \{R(F_1^{b_1}, F_2^{b_2}, F_2^{b_3}), b_1 = 0°, 90°; \\
&\qquad b_2, b_3 = 0°, 60°, 120°;\ b_2 \le b_3\} \\
K(R(F_2^{\theta_1}, F_2^{\theta_2}, F_2^{\theta_3})) &= \{R(F_2^{b_1}, F_2^{b_2}, F_2^{b_3}), b_1, b_2, b_3 = 0°, 60°, 120°; \\
&\qquad b1 \le b_2 \le b_3\}
\end{aligned}
$$

In the above, we get 50 directional autocorrelation features which can be arranged to a vector of 50 dimensios to represent a gray image. For color images, a vector of 150 dimensions is used.

4.3 Linear Transform for Image Rotation

Let $h(x', y')$ be the image rotated by γ degree from image $f(x, y)$. $F_n^{\theta}(x, y)$ and $H_n^{\theta}(x', y')$ be nth-order Gaussian differential in θ direction computed for $f(x, y)$ and $h(x', y')$ respectively. Then $F_n^{\theta}(x, y)$ and $H_n^{\theta}(x', y')$ have following relation.

$$H_n^{\theta}(x', y') = F_n^{\theta+\gamma}(x, y) \tag{8}$$

Let $\{R(F_{n_1}^{\theta_1}, F_{n_2}^{\theta_2}, ..., F_{n_m}^{\theta_m})\}$ be the high-order Gaussian directional differential on the original image$f(x, y)$ and let $\{R'(H_{n_1}^{\theta_1}, H_{n_2}^{\theta_2}, ..., H_{n_m}^{\theta_m})\}$ be that of the rotated image $h(x', y')$. We can show that

$$
\begin{aligned}
&R'(H_{n_1}^{\theta_1}, H_{n_2}^{\theta_2}, ..., H_{n_m}^{\theta_m}) \\
&= \int_{E_{x'}} \int_{E_{y'}} H_{n_1}^{\theta_1}(x', y') H_{n_2}^{\theta_2}(x', y') ... H_{n_m}^{\theta_m}(x', y') dx' dy' \\
&= \int_{E_x} \int_{E_y} F_{n_1}^{\theta_1+\gamma}(x, y) F_{n_2}^{\theta_2+\gamma}(x, y) ... F_{n_m}^{\theta_m+\gamma}(x, y) dx dy \\
&= R(F_{n_1}^{\theta_1+\gamma}, F_{n_2}^{\theta_2+\gamma}, ..., F_{n_m}^{\theta_m+\gamma})
\end{aligned} \tag{9}
$$

As described in the previous section, $R(F_{n_1}^{\theta_1+\gamma}, F_{n_2}^{\theta_2+\gamma}, ..., F_{n_m}^{\theta_m+\gamma})$ can be represented as a linear combination of $K(R(F_{n_1}^{\theta_1}, F_{n_2}^{\theta_2}, ..., F_{n_m}^{\theta_m}))$. That is,

$$R'(H_{n_1}^{\theta_1}, H_{n_2}^{\theta_2}, ..., H_{n_m}^{\theta_m})^T = A(\gamma) \vec{K}(R(F_{n_1}^{\theta_1}, F_{n_2}^{\theta_2}, ..., F_{n_m}^{\theta_m}))^T \tag{10}$$

where $A(\gamma)$ is a vector which contains a parameter γ. $\vec{K}(R(F_{n_1}^{\theta_1}, F_{n_2}^{\theta_2}, ..., F_{n_m}^{\theta_m}))$ is a vector by lining up the basis high-order Gaussian directional differential features in $K(R(F_{n_1}^{\theta_1}, F_{n_2}^{\theta_2}, ..., F_{n_m}^{\theta_m}))$.

To recognize the rotated image with the learning images without rotation, we just need to solve the following problem.

$$arg \min_i\{\min_\gamma \|\vec{R}'(H_{n_1}^{\theta_1}, H_{n_2}^{\theta_2}, ..., H_{n_m}^{\theta_m})A^T(\gamma)E - M_i\|^2\} \quad (11)$$

where E is the projecting matrix obtained by discriminant analysis which project the the high-order Gaussian directional differential features to the discriminant space, in which the classes can be separated most easy. M_i stands for the centers of class i in the discriminant space.

As $\|\vec{R}'(H_{n_1}^{\theta_1}, H_{n_2}^{\theta_2}, ..., H_{n_m}^{\theta_m})A^T(\gamma)E - M_i\|^2$ is polynominal of $\sin(\gamma)$ and $\cos(\gamma)$, it is usually difficult to find an analytic solution for the minimalizing problem. In our research, we seek for a approximate minimal solution by following method.

$$\min\{\|\vec{R}'(H_{n_1}^{\theta_1}, H_{n_2}^{\theta_2}, ..., H_{n_m}^{\theta_m})A^T(\gamma_i)E - M_i\|^2, \\ \gamma_i = -180, -180 + s, ..., 180\} \quad (12)$$

where s is the step of rotation degrees in seeking the minimal solution. In experiment s takes a value of 5.

§5 Object Recognition and Pose Detection

In order that both object recognition and pose detection can be carried out efficiently, we construct the discriminant space hierarchically by bottom up approach.

First, we generate 110 poses of the object from uniformly contributed view points as shown in Fig. 2 (c). For each pose P_k and its neighboring poses $N_k = \{P_{k_j}, j = 1, ..., n_k\}$, we find the ellipse with largest density of neighboring poses by random sampling method. Then we iteratively extract the the ellipse with largest density of neighboring poses obtained at each pose. The poses contained in the extracted ellipse are treated as a cluster of poses.

For a given object, discriminant spaces of two hierarchies are built, where the top one is built for the clusters obtained above. For each cluster which contains more than one poses, a discriminant space is built where a pose is regarded as a class. In pose detection, we first project the view of the object to the top discriminant spaces and find the cluster to which the view of the object belongs. If the obtained cluster contains only one pose, we get the pose of the view at the same time. Otherwise, we project the view to the discriminant space of the found cluster, and detecting the pose of the view.

In the case of more than one objects, the discriminant spaces for object recognition must also be built. One simple way is to regard an object as a class. However, if the view of an object changes much when the view point changes, a class may contain samples without similarity in feature. In this research, we represent each object with 20 classes. These classes are obtained by iteratively merging two closet clusters of poses. For n objects, $20n$ classes are used to build the discriminant space for object recognition. In this space, a view is recognized object O_n if it is classified to any class of object O_n.

§6 Experimental Results

We applied our method to three real objects, one doll of bear and two boxes. The results of object recognition and pose detection are summarized in Table 1. In the experiment, the background is not taken into account.

Table 1 Recognition Results

Item	box 1	doll	box 2
Number of test images	13	14	12
Correct object recognition	13	14	12
Correct pose detection	12	12	10
Incorrect pose detection	1	2	2

The detected pose is judged to be correct or not by human based on difference between input view and the view of detected pose. If they are not different from each other obviously, the pose detection result is regarded to be correct. Fig. 3 some shows examples of correct results of pose detection.

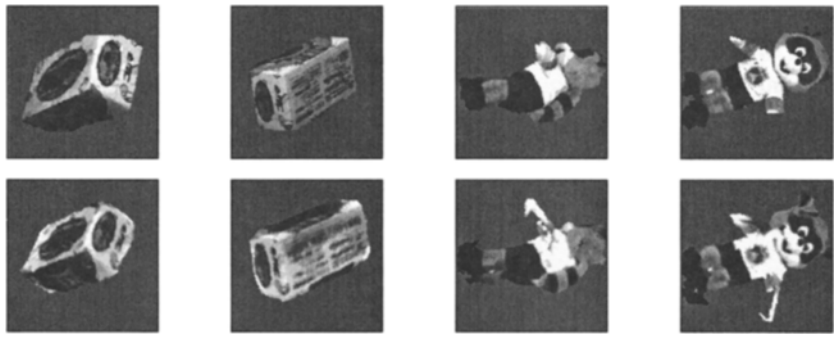

Fig. 3 Examples of Correct Recognition

§7 Conclusion and Future Work

In this research, we proposed a view-based virtual learning approach which uses the 3D view model of object obtained by motion-stereo method. Using the 3D view model, we generated the learning sample views of the object. From the generated learning sample views, features of high-order autocorrelation are extracted, and hierarchical discriminant spaces for object recognition and pose detection are built. Recognition experiments on real objects have been carried out to show the effectiveness of the proposed method.

In current experiment, only three objects are used. It is necessary to test our method on more objects. Also the background in the scene has not be taken into account. This should be studied in future work.

References

1) Ikeiuchi, K. and Kanade, T., "Towards Automatic Generation of Object Recognition Program," *Proc. IEEE*, *76*, *8*, pp. 1016–1035, 1988.

2) Lamdan, Y. and Wolfson, H. J, "Geometric Hashing: A General and Efficiet Model-Based Recognition Scheme," *Proc. Image Understanding Workshop*, pp. 238–249, 1988.

3) Beveridge, J. R. and Riseman, E. M., "Optimal Geometric Model Matching under Full Perspective," *Computer Vision and Image Understanding*, *61*, *3*, pp. 351–364, 1995.

4) Murase, H. and Nayar, S., "Visual Learning and Recognition of 3-D Object from Appearance," *International Journal of Computer Vision*, *14*, *1*, pp. 5–24, 1995.

5) Hasegawa, O., Kurita, T. and Sakaue, K., "A Proposal of Scene Understanding by Learning and Basic Experiments," *Proc. SII'97*, pp. 129–132, 1997.

6) Otsu, N. and Kurita, T., "A New Scheme for Practical Flexible and Intelligent Vision Systems," *Proc. IAPR Workshop on Computer Vision*, pp. 341–435, 1988.

7) Amano, T., Yamaguchi, A. and Inokuchi, S., "Eigenspace Approach for Object Recognition and its Pose Detection," *IEICE Technical Report on PRMU98-20*, pp. 71–78, 1998.

8) Wang, C. and Sakaue, K., "Acquiring 3D Model of Object by Motion-Stereo," *Proc. IAPR Workshop on Machine Vision Application*, pp. 301–305, 1998.

9) Freeman, W. T. and Adelson, E. H., "The Design and Use of Steerable Filters," *IEEE Trans. on PAMI*, *13*, *9*, pp. 891-906, 1991.

Caihua Wang, Ph.D.: He received his B.S. in mathematics and M.E. in electronic engineering from Renmin University of China, Beijing, China in 1983 and 1986, and his Ph. D. from Shizuoka University, Hamamatsu, Japan in 1996. He is a JST domestic fellow and is doing his post doctoral research at Electrotechnical Laboratory. His research interests are computer vision and image processing. He is a member of IEICE and IPSJ.

Katsuhiko Sakaue, Ph.D.: He received the B.E., M.E., and Ph.D. degrees all in electronic engineering from University of Tokyo, in 1976, 1978 and 1981, respectively. In 1981, he joined the Electrotechnical Laboratory, Ministry of International Trade and Industry, and engaged in researches in image processing and computer vision. He received the Encouragement Prize in 1979 from IEICE, and the Paper Award in 1985 from Information.