A quick 3D-2D registration method for a wide-range of applications

Yasuyo Kita* Dale L. Wilson† J. Alison Noble‡ Nobuyuki Kita*
* Intelligent Systems Division, Electrotechnical Laboratory, Tsukuba, Japan 305-8568, {ykita, nkita}@etl.go.jp
†Mathematical & Information Sciences, CSIRO, Sydney, Australia, Dale.Wilson@cmis.CSIRO.AU
‡Dept. of Engineering Science, University of Oxford, Oxford,UK OX1 3PJ, noble@robots.ox.ac.uk

Abstract

A method for quick determination of the position and pose of a 3D free-form object with respect to its 2D projective image(s) is proposed. It is a precondition of the method that a 3D model of the object and an initial estimation of the state are given. In [1], we have proposed a 3D-2D registration method for the real-time registration of a 3D model of a cerebral vessel tree to a X-ray image of the vessel taken during an operation. The method is robust and fully automated from raw input image to the final result. In this paper, we extend this method to meet a more general purpose. First, the formulae for obtaining the 3D model transformation from 3D-2D point pairs are generalized by describing camera coordinates independently from world coordinates. This enables easy integration of multiple images and simple treatment of moving camera coordinates. The second generalization is to use the occluding contour instead of the skeleton of the blood vessel(tube-like shape) as the feature for matching. To quickly obtain the 3D model points corresponding to the occluding contours in an observed image, we take a 3D graphics system like OpenGL into our 3D-2D registration method. The applicability improvements are shown using two applications: 1) position and pose estimation of a 3D vessel model using multiple views, 2) visual feedback on the position and pose of an active camera head.

1. Introduction

The determination of the position and pose of a 3D object from its 2D view is a fundamental and important problem in computer vision research. One typical approach for this purpose is feature-based one: it first extracts features(eg. edges, corners) and matches them between the 3D model and its 2D view. Once the 3D-2D point correspondences are obtained, the position and pose of the model can be quickly calculated [2]. Usually, however, neither robust feature extraction nor robust feature matching are easy. Especially, for a free-form object, finding the correspondences on an observed image is extremely difficult. In the case that the position and pose of the object can be roughly estimated in advance, it is an effective strategy to iteratively transform the 3D model towards the correct position and pose using the corresponding pairs between the observed and the model points, which are matched on the basis of the closeness at each state[3]. Some methods have been developed for 3D-2D registration according to this strategy using the closeness between the observed 2D points and the projections of the 3D model points at the state[4] or the closeness between the projection rays from the observed 2D points and the 3D model points[5][6]. Before the matching, however, features corresponding to 3D model points should be well extracted from observed images. This is still difficult and is a bottleneck. Actually, the methods treating free-from objects[5] [4] needed manual helps for feature extraction processes in their experiments using actual scenes. It is not clear how long the total processes through from raw input image take.

To achieve more robust feature extraction and matching for free-form objects, Kita et al. [1] adopt a model-based approach for the whole processes and realized a totally quick method. The method gave good results for registration of a 3D model of a cerebral vessel tree to its X-ray image. There are, however, the following two conditions are obstacles to apply this method for more general situations.

1) The relationship between camera and world coordinates were not considered in a general manner.

2) The object should be a tube-like since its skeleton used as a corresponding feature.

In this paper, we generalize this method concerning the above two points. First, camera coordinates are clearly described with respect to world coordinates. According to this, the formulae for obtaining the 3D model transformation from 3D-2D point pairs are neatly modified. This generalization enables easy integration of multiple images and simple treatment of moving camera coordinates. Second,

Proceedings of the International Conference on Pattern Recognition (ICPR'00)
1051-4651/00 $10.00 @ 2000 IEEE

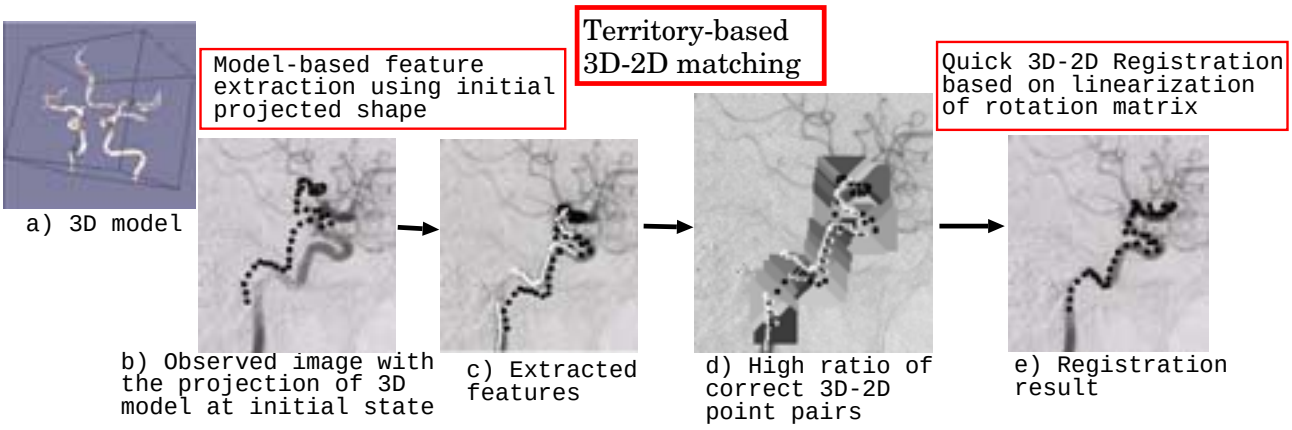


Figure 1. Basic scheme of our 3D-2D registration method

to deal with more general shapes, it is taken into consideration to use the occluding contour as the feature for 3D-2D matching. To quickly calculate the contour generator that is the 3D line on the object's surface corresponding to the occluding contour in the observed image, we take a 3D graphics system like OpenGL into our 3D-2D registration method and effectively use the depth image supplied by it. In the following sections, first we briefly explain our basic strategy, and then describe the details of these improvements with lots of experiments showing their effects.

2. Basic strategy

The basic strategy is shown in Fig. 1. The details can be found in [1]. Fig. 1b is a X-ray image of the right internal carotid circulation (the right side of the cerebral vessels) in Fig. 1a taken after injecting contrast medium into only the right part. The acquisition angle is given from the graduations of the X-ray system but includes some errors. As a result, the projection of the 3D model points sampled from the skeleton of the 3D model vessel at the state is deviated from the observed vessel as shown by the black points in Fig. 1b. First, the projected shape is two-dimensionally translated on the image to the position which gives optimal overlap on the dark regions (possible vessel regions) as shown by the black points in Fig. 1c. The corresponding features on the image, that is the skeletons of 2D vessels in this case, are extracted in the neighborhood of the projected shape in a model-based way(white lines). At this stage, perfect feature extraction is not required. Lack of corresponding features can be allowed, due to the characteristics of the following territory-based 3D-2D matching. The territory-based 3D-2D matching uses anisotropic search regions determined from the projected shape of the model when making 3D-2D corresponding pairs based on the closeness in the image. As shown in Fig. 2, this territory-based search restriction well removes out the model point whose corresponding 2D feature has

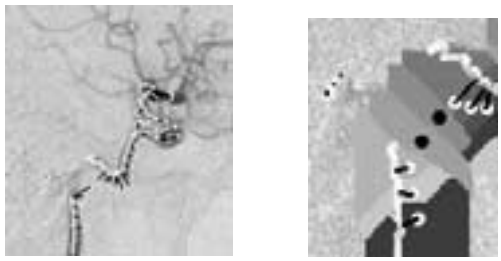
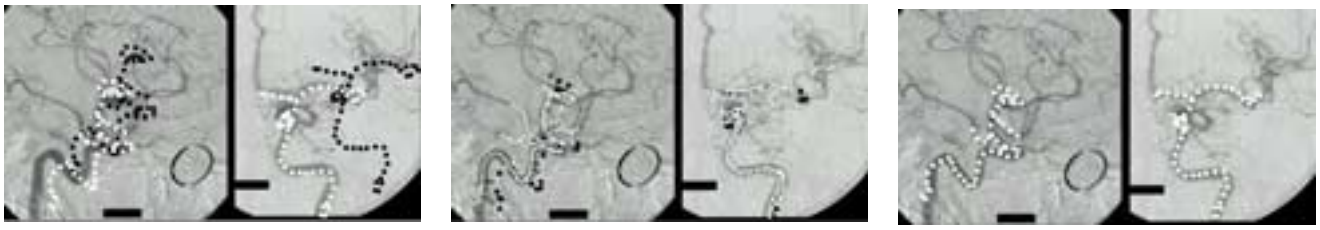


Figure 2. Territory-based search region removing out wrong pairs caused by lack of observed features

not been extracted in the observed image. Finally, taking advantage of the high-ratio of viable 3D-2D point pairs obtained by the previous processes, the 3D transformation of the model is quickly calculated by separating the translation effect out and linearizing the rotation matrix[2]. Although the correct position and pose of the model is not obtained at once, because of inaccurate matching pairs and linearization errors, the 3D model quickly converges to the correct state by iterating the point matching and model transformation processes. In this example, the processes were iterated 30 times for convergence and the total processing time was 2.7 sec on a Pentium II(333MHz) machine.

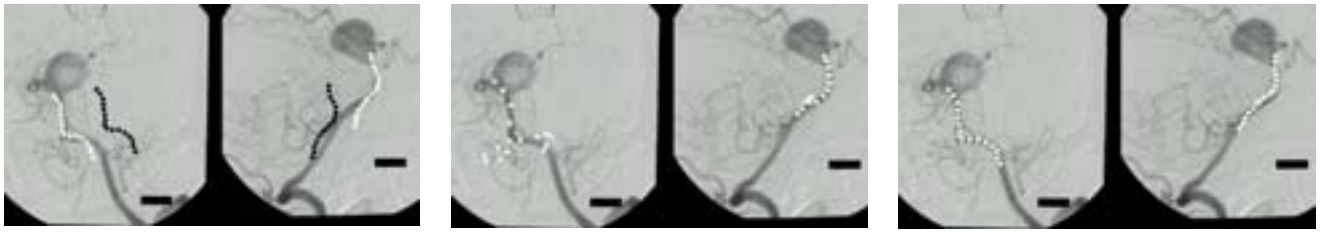
3. Generalization of the camera coordinates

To calculate the 3D transformation of the model from 3D-2D corresponding point pairs, we take the strategy presented in [2]. Supposed that we have m corresponding pairs of the observed point $\mathbf{p}_i, (x_i, y_i, f)^\top$ and a 3D model point $\mathbf{P}_i, (X_i, Y_i, Z_i)^\top$. First, the following minimization criteria based on only rotation is built by separating out the translation effect and by linearizing the rotation matrix represented with quaternions $\mathbf{q} = (q_0, q_1, q_2, q_3)^\top$.



(a) (b) (c)

Figure 3. Experiment using two views of the left internal carotid circulation in Fig. 1a; (a) projections of the 3D model at the initial state; (b) correspondences using territory-based search regions; (c) projections of the 3D model at the resultant position and pose.



(a) (b) (c)

Figure 4. Experiment using two views of a basilar artery: (a) projections of the 3D model at the initial state; (b) results using one image respectively; (c) result using the two images simultaneously .

$$\min_{\mathbf{q}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} ((\mathbf{p}_i \times \mathbf{p}_j) \cdot (q_0(\mathbf{P}_i - \mathbf{P}_j) + 2\mathbf{q}' \times (\mathbf{P}_i - \mathbf{P}_j)))^2 \tag{1}$$

where $\mathbf{q}' = (q_1, q_2, q_3)^\top$.

Once the rotation is known, the translation can be calculated by solving $2m$ simultaneous equations on (t_x, t_y, t_z):

$$\frac{f(X_i' + t_x)}{Z_i' + t_z} = x_i \text{ and } \quad \frac{f(Y_i' + t_y)}{Z_i' + t_z} = y_i \tag{2}$$

where $\mathbf{X}'$=$\mathbf{RX}$ using the calculated $\mathbf{R}$.

We extend these formulae to treat arbitrary numbers of arbitrary camera coordinates in world coordinates. From now on, we use the "c" superscript to represent the values in the camera coordinates, whose origin is (X_0, Y_0, Z_0) and whose $x^c(y^c, z^c)$ axis has direction cosine $(\lambda_{x(y,z)}, \mu_{x(y,z)}, \nu_{x(y,z)})$ in world coordinates. In Equation (1), $\mathbf{R}$ in the world coordinates is calculated in the same way by using the world coordinates for all variables in the equation. In Equation (2), (t_x^c, t_y^c, t_z^c) can be transformed into one involving (t_x, t_y, t_z) using the relation of $t^c_{x(y,z)} = F(\lambda_{x(y,z)}, \mu_{x(y,z)}, \nu_{x(y,z)}, t_x, t_y, t_z)$. Once we calculate $\mathbf{R}$ and $\mathbf{T}$ in the world coordinates, it is easy to integrate the information from multiple images to determine the position and pose of a 3D model. As a result, general formulae to determine $\mathbf{R}$ and $\mathbf{T}$ using n cameras become as follows.

On rotation:

$$\min_{\mathbf{R}} \sum_{k=1}^{n} \sum_{i=1}^{m(c_k)-1} \sum_{j=i+1}^{m(c_k)} (F_{c_k}(\mathbf{p}_i^{c_k} \times \mathbf{p}_j^{c_k}) \cdot \mathbf{R}(\mathbf{P}_i - \mathbf{P}_j))^2 \tag{3}$$

Where $F_{c_k}(\mathbf{p})$ is the function to transform the kth camera coordinates to the world coordinates, that is $\mathbf{P} = F_{c_k}(\mathbf{P}^c)$; n is the number of cameras.

On translation:

$$\begin{bmatrix} f^{c_1}\lambda_x^{c_1} - x_1^{c_1}\lambda_z^{c_1} & f^{c_1}\mu_x^{c_1} - x_1^{c_1}\mu_z^{c_1} & f^{c_1}\nu_x^{c_1} - x_1^{c_1}\nu_z^{c_1} \\ f^{c_1}\lambda_y^{c_1} - y_1^{c_1}\lambda_z^{c_1} & f^{c_1}\mu_y^{c_1} - y_1^{c_1}\mu_z^{c_1} & f^{c_1}\nu_y^{c_1} - y_1^{c_1}\nu_z^{c_1} \\ \vdots & \vdots & \vdots \\ f^{c_1}\lambda_x^{c_1} - x_{m(c_1)}^{c_1}\lambda_z^{c_1} & f^{c_1}\mu_x^{c_1} - x_{m(c_1)}^{c_1}\mu_z^{c_1} & f^{c_1}\nu_x^{c_1} - x_{m(c_1)}^{c_1}\nu_z^{c_1} \\ f^{c_1}\lambda_y^{c_1} - y_{m(c_1)}^{c_1}\lambda_z^{c_1} & f^{c_1}\mu_y^{c_1} - y_{m(c_1)}^{c_1}\mu_z^{c_1} & f^{c_1}\nu_y^{c_1} - y_{m(c_1)}^{c_1}\nu_z^{c_1} \\ f^{c_2}\lambda_x^{c_2} - x_1^{c_2}\lambda_z^{c_2} & f^{c_2}\mu_x^{c_2} - x_1^{c_2}\mu_z^{c_2} & f^{c_2}\nu_x^{c_2} - x_1^{c_2}\nu_z^{c_2} \\ f^{c_2}\lambda_y^{c_2} - y_1^{c_2}\lambda_z^{c_2} & f^{c_2}\mu_y^{c_2} - y_1^{c_2}\mu_z^{c_2} & f^{c_2}\nu_y^{c_2} - y_1^{c_2}\nu_z^{c_2} \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}$$

$$= \begin{bmatrix} x_1^{c_1} Z_1^{\prime c_1} - f^{c_1} X_1^{\prime c_1} \\ y_1^{c_1} Z_1^{\prime c_1} - f^{c_1} Y_1^{\prime c_1} \\ \vdots \\ x_{m(c_1)}^{c_1} Z_{m(c_1)}^{\prime c_1} - f^{c_1} X_{m(c_1)}^{\prime c_1} \\ y_{m(c_1)}^{c_1} Z_{m(c_1)}^{\prime c_1} - f^{c_1} Y_{m(c_1)}^{\prime c_1} \\ x_1^{c_2} Z_1^{\prime c_2} - f^{c_2} X_1^{\prime c_2} \\ y_1^{c_2} Z_1^{\prime c_2} - f^{c_2} Y_1^{\prime c_2} \\ \vdots \end{bmatrix} \tag{4}$$

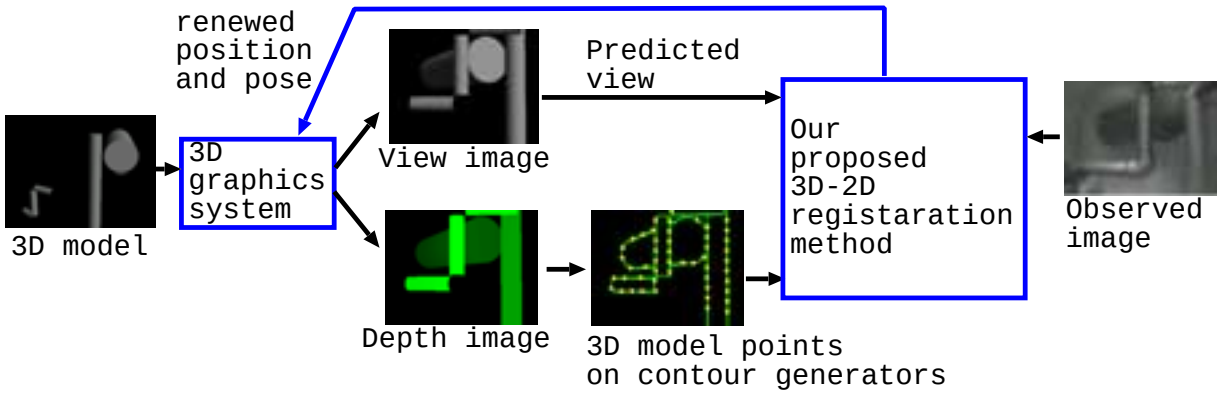


Figure 5. Use of 3D graphics system

4. Position and pose estimation of 3D vessel model using multiple X-ray images

To examine the effect of the new formulae, calibration of a 3D vessel model with a X-ray system using two X-ray images were conducted. The position and pose of the 3D model relative to the system known from its graduations include about $\pm$ 20 degrees error in rotation and about ($\pm$100,$\pm$100,$\pm$200)(mm) in translation, since the position and pose of the head is not calibrated to the X-ray system. Although the 3D relative geometry between the two X-ray images given by the graduations also includes small errors because of the deflection of the arm of the X-ray system, the errors are enough small to ignore comparing to the errors in the relation between the 3D model and the X-ray system. Fig. 3 shows a result of registration of the left internal carotid circulation in Fig. 1a. In Fig. 3a the black and white points respectively represent the projection of the 3D model skeleton at its initial state and after translating it on the image so as to overlap the dark regions. In Fig. 3b, correspondences using territory-based search restriction at the first iteration are shown. After 27 iterations using Equations (3) and (4), the 3D model is converged to the position and pose so as to produce the projection as shown as white points in Fig. 3c. The transformation of the 3D model is a 13.0 degree rotation around the axis (0.98, -0.06 , 0.19) and a (19.6, 3.3, -10.1)(mm) translation.

In this case, results using one image only are not very different since the vessel has enough complexity in shape to determine its position and pose from one view. However, in the case that an object has a simple shape as shown in Fig. 4, it is fairly effective to use multiple images. The black and white points in Fig. 4a show the projection of the 3D model in the same way as Fig. 3a. The result when we use only one view is shown in Fig. 4b. In the left-hand image, the projection is largely deviated from the observed vessels after a 63.6 degree rotation around the axis (0.63, 0.47, 0.61). Although, in the right-hand image, the projection is on the observed vessels after a 62.5 degree rotation around the axis (0.78, -0.44, -0.43), the rotation angle is clearly far from the actual value, which is known up to 20 degree. On the other hand, in Fig. 4c, the result using two views simultaneously gives reasonable projections on the both images. Although, unfortunately, we do not have the ground truth data for the experiments, a 16.9 degree rotation around the axis (0.81, -0.01, 0.58) seems proper.

5. Visual feedback for active camera head

When a 3D model of the environment surrounding a camera is given, the position and pose of the camera can be determined using registration of the 3D model to the image taken by the camera. Based on this principle, we apply our proposed method to visual feedback for an active camera head. An active camera head can usually know its position and pose from its control driving modules, but typically the values include some errors. Our 3D-2D registration method can be used for estimating the error.

In the case of the registration of the blood vessel, its 3D skeleton can be used as the matching feature independently of the view direction. In general, the occluding contour is typically the observed feature. The contour generator, that is the 3D line on the object's surface corresponding to the occluding contour, is largely changed depending on the view direction. That is, the 3D model points corresponding to the observed features should be calculated at each state of the model. To quickly obtain such 3D model points even for complex 3D scene model, we propose to combine a 3D graphics system like OpenGL with our 3D-2D registration method as shown in Fig. 5, since it can rapidly supply both the view and depth image of a scene.

The basic flow for obtaining the estimation error is follows:

i) The edges in the observed image are extracted in a model-based way using the view image (predicted view) supplied by the 3D graphics system.

ii) 3D model points on the contour generators are derived from the differentiation of the depth image supplied by the

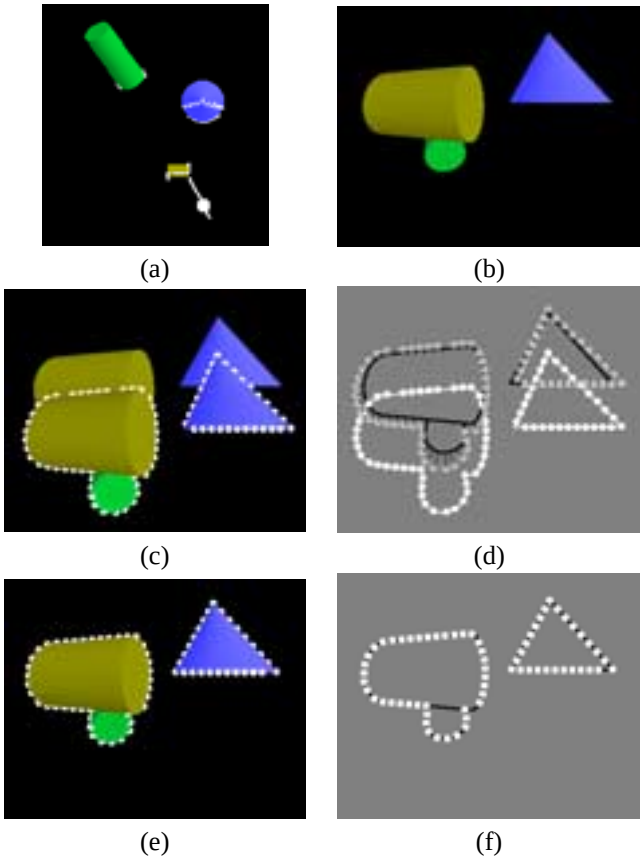


Figure 6. Experiment on visual feedback using synthetic data

3D graphics system.

iii) The 3D-2D point correspondences are found using territory-based search restrictions.

iv) The transformation of the 3D model determined from the 3D-2D correspondences is calculated using Equations (3) and (4) with the condition that $n = 1$.

v) The estimation on the position and pose of the camera head is renewed by applying the transformation inverse to the one obtained in iv).

By iterating the processes from ii) to v), the correct position and pose of the camera is obtained, as the predicted view converges to match the observed image.

Fig. 6 shows a result of synthetic experiments. Fig. 6a is a top view of the world surrounding a camera head. The white circle in Fig. 6a shows the position of the camera; the two lines represent the view and up directions of the camera. Supposed that Fig. 6b is an observed image. If there is an estimation error on the camera state, the predicted view deviates from the observed image. Fig. 6c shows such an example by superposing the predicted view on the observed image. The white points in Fig. 6a and c represents 3D model points obtained at the stage ii). The white points

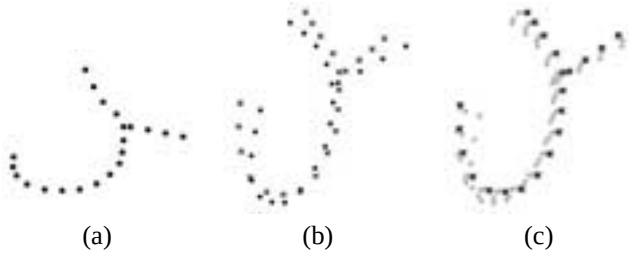


Figure 7. Experiment on convergence: (a) 3D free-form curve; (b) observed 2D points($\times$) and the projection of the model at the initial state($\bullet$); (c) convergence result.

in Fig. 6d shows their projected positions in the observed image. The projected shape is two-dimensionally translated on the image to the position at the grey points which gives optimal overlap on the observed edges(the black lines in Fig. 6d). At the stage iii), the corresponding 2D positions on the observed edges are found for the 3D model points using territory-based search regions as shown with thin lines. Fig. 6e and f show the result after 11 iterations. The processing time is 5.8 sec on a Pentium II(333MHz) machine. A 11.4 degree rotation around the axis (0.89, -0.00 , -0.45) and a (-0.06, -0.13, -0.05) translation were obtained as the estimated error which is actually a 11.5 degree rotation around the axis (0.89, 0.0,-0.45) without translation. For reference to judge the translation, the height of the nearest cylinder is 30.0.

6 Experiments on allowable initial estimate error

When applying the proposed method practically, it is important to know the range of initial estimate error that the method allows. For this purpose, we conducted experiments from several viewpoints.

First, the limitation on the rotation angle caused by the linearization of the rotation matrix is considered. Although the angular error resulting from the linearization of the rotation increases sharply as the rotation angle increases, the error is less than 0.08 degree for the rotation angle under 10 degree[2]. We tested allowable rotation angle for converging to the correct state using a synthetic 3D free-form curve represented by 3D points as shown in Fig. 7a. $\times$ and $\bullet$ in Fig.7b show the observed 2D projection and the projection of the 3D model at the initial state respectively. By changing the initial state, convergence to the goal state, which is the actual position and pose producing the observed 2D projection, is checked. When we gave correct corresponding 2D points for all model points, the model converged to the correct state after a few iterations even from $\pm$ 50 degrees rotated state.

Next, we tested on allowable ratio of correct corresponding point pairs using the same data. This time, correspondences between 3D model points and 2D observed points are determined only based on the closeness in the image so that wrong pairs are easily contained. In this case, we found that the convergence rate depends on the wrong pair ratio ((No. of model points having wrong corresponding pairs)/ (No. of all model points)) rather than the amount of rotation degrees. If the ratio is under around 50 % the model converged to the correct posture. Fig.7c shows a success result when the initial state (Fig.7b) is transformed by a 40 degree rotation around the axis(0.0, 1.0 ,0.0) and a (-10.0, 20.0 -20.0) translation. However, depending on the axis of rotation, even a 10 degree rotation causes a high wrong pair ratio so that the convergence is failed.

Finally, we examined the allowable initial estimate error in the case of the visual feedback described in Section 5. When the initial estimate includes rotation error only, the 3D model converged to the correct state up to a 20 degree error around arbitrary rotation axis. When the initial estimate includes a translation error only, the 3D model converged to the correct state up to a 20 translation (the height of the nearest cylinder is 30). Although the situations have too much variety to thoroughly examine when the both errors are included, any initial estimate including up to a 15 degree rotation error around any axis with under a 10 translation error gave correct convergences as far as we tested.

7 Conclusion

In this paper, we proposed a quick 3D-2D registration method which is applicable to a wide-range of applications. The attractive general features of our approach are:

1. Few limitations on the shape of an object
   The method can treat both a very complex shape(eg. a vessel tree) and a combinational shape composed of many objects(eg. an environmental model) in a uniform manner. This is largely due to the "territory-based 3D-2D matching(search restriction)", which treats the model points as a pattern without needing to know its structure.
2. Computational quickness
   Existing feature-based 3D-2D registration methods[5][4] have been considered separately from the part of feature extraction processes. The proposed method is quick and totally automatic by consistently connecting the feature extraction and matching processes considering their characteristics: 1) model-based feature extraction produces less outliers though it may include lack of features; 2) territory-based matching is effective in removing the wrong pairs caused by the lack of features.
3. Adaptability for wide-range applications
   By clearly describing the relation of arbitrary camera coordinates to the world coordinates, the method can consistently deal with various situations of the camera, such as simultaneous use of multiple images, moving camera coordinates and so on.

On the subject of the position and pose determination of a 3D model of a vessel tree, the effect of using multiple views especially for objects of a simple shape was shown. Using relative transformation of a 3D environmental model, visual feedback on the position and pose of a camera head can be implemented in the same strategy. Although the experiments presented in this paper are at a fairly preliminary stage, the results show that the approach is promising. The allowable initial estimate errors examined in Section 6 seem to be enough big for ordinary active camera head. We plan to use the method for autonomous robot-based inspection[7] and have already got a good prospect from experiments using actual images of a complex scene composed of pipes[8].

References

[1] Y. Kita, D. L. Wilson and J. A. Noble: "Real-time registration of 3D cerebral vessels to X-ray angiograms", In *Proc. of 1st International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 1125–1133, 1998.

[2] J.J. Heuring and D.W. Murray: "Visual head tracking and slaving for visual telepresence", In *Proc. of IEEE Int Conf. on Robotics and Automation*, pp. 2908–2914, 1996.

[3] P. J. Besl and N. D.Mckay: "A method for registration of 3D shapes", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 14, No.2, pp. 239–256, 1992.

[4] J. Feldmar, G. Malandain, N. Ayache, S. Fernandez-Vidal, E. Maurincomme and Y. Trousset: "Matching 3D MR Angiography Data and 2D X-ray Angiograms", In *Proc. of CVRMed-MRCAS'97*, pp. 129–138, 1997.

[5] S. Lavallée, and R. Szeliski: "Recovering the position and orientaion of free-form objects from image contours using 3D distance maps", *IEEE trans. Pattern Anal. & Mach. Intell.*, **17**, 4, pp. 378–390, 1995.

[6] P. Wunsch and G. Hirzinger: "Registration of CAD-Models to Image by iterative inverse perspective matching", In *Proc. of 13th International Conference on Pattern Recognition*, pp. 78–83, 1996.

[7] N. Kita: "Intelligent Plant Inspection by Using Foveated Active Vision Sensor", *Proceedings of HCI International'99*, 2, 1177–1181, 1999.

[8] Y. Kita and N. Kita: "Pose and position detection using occluding edges in complex scene", In *Meeting Image Recognition and Understanding 2000 (MIRU 2000)*, 2000 [In Japanese](To appear).

1051-4651/00 $10.00 @ 2000 IEEE