A Stereo Vision-based Mixed Reality System with Natural Feature Point Tracking

Masayuki Kanbara †, Hirofumi Fujii ‡, Haruo Takemura † and Naokazu Yokoya †

†Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara 630-0101, JAPAN
‡Matsushita Communication Industrial Co., Ltd.
600 Saedo-cho, Tsuzuki-ku, Yokohama, Kanagawa 224-8539, JAPAN
E-mail: {masay-ka, takemura, yokoya}@is.aist-nara.ac.jp

Abstract

This paper proposes a method to extend a registration range of a vision-based mixed reality system. We propose to use natural feature points contained in images captured by a pair of stereo cameras in conjunction with pre-defined fixed fiducial markers. The system also incorporates an inertial sensor to achieve a robust registration method which can handle user's fast head rotation and movement. The system first uses pre-defined fiducial markers to estimate a projection matrix between real and virtual coordinate systems. At the same time, the system picks up and tracks a set of natural feature points from the initial image. As a user moves around in MR environment, the initial markers fall out from the camera frame and natural features are then used to recover a projection matrix. Experiments evaluating the feasibility of the method are carried out and show the potential benefits of the method.

Keywords: Augmented Reality, Vision-based Registration, Inertial Sensor, Natural Feature Tracking, Stereo Vision

1 Introduction

Augmented reality produces an environment in which virtual objects are superimposed on user's view of the real environment. Augmented reality has received a great deal of attention as a new method for displaying information or increasing the reality of virtual environments. A number of applications have already been proposed and demonstrated [1, 2, 5, 9]. To implement an augmented reality system, we must solve some problems. Geometric registration is especially the most important problem because virtual objects should be superimposed on the right place as if they really exist in the real world.

One of the major approaches to the registration between the real and virtual worlds is vision-based method [3, 6, 8, 10, 11]. The methods, which are sometimes referred to as vision-based tracking or registration, estimate a position and an orientation of user's viewpoint from images captured by a camera attached at the user's viewpoint. Because the method usually uses fiducial markers placed in the environment, the measurement range is actually limited.

To overcome this limitation, Park et al. proposed a registration method that tracked natural features in addition to markers in the real environment [7]. The method realizes a wide range registration between the real and virtual worlds by tracking markers and natural features. However, tracking the natural features in the captured images is usually difficult because of the following two reasons. One is that a template matching for tracking is not robust or stable enough, especially when a perspective of the scene changes as a user moves in a 3-D environment, although the template matching-based tracking method has been extensively studied in the field of computer vision. The other is that tracking itself is time consuming and makes it difficult for the system to run in real time.

One of solutions for the problem above is to predict positions of features in the next frame using Kalman filter [7]. The method avoids miss tracking of markers and reduces calculation cost by limiting a search area in the image. On the other hand, You et al. proposed a tracking method using an inertial sensor attached to a camera for predicting markers' positions [12]. The method is able to track the features even when the camera moves fast. However, the method cannot accurately predict positions of markers when the camera is translated since an inertial sensor can measure only rotation of the camera. Figure 1 illustrates the problem. When the camera is translated and rotated at the same time, the predicted position of marker is $\mathbf{P}'$ at time t+Δt without considering the camera translation. However, the predicted position $\mathbf{P}'$ differs from the correct position $\mathbf{P}_{t+\Delta t}$ of marker by translation $\mathbf{T}$.

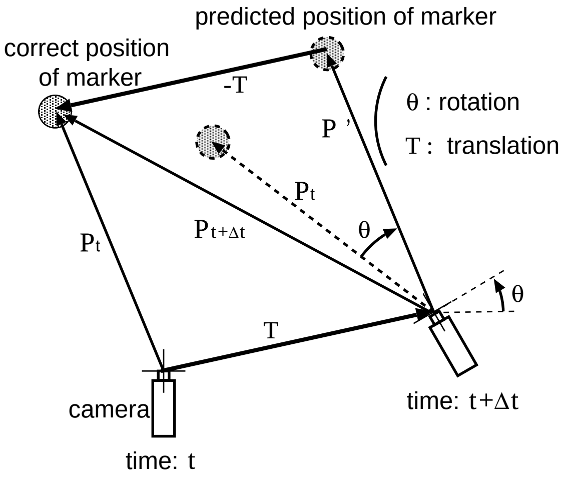

Figure 1 Relationship between a marker and a camera in motion.

In this paper, we propose a stereo vision-based augmented reality with a wide range of registration by using pre-defined fiducial markers and natural features. In addition, we discuss a robust tracking method using an inertial sensor, which predicts positions of markers and natural features for tracking.

The following part of the paper is structured as follows. Section 2 describes the stereo vision-based registration method with an inertial sensor using natural feature points. In Section 3, experimental results with the proposed method and discussion about a prototype system are described. Finally, Section 4 summarizes the present work and describes future work.

2 Geometric registration using markers and natural features

We assume in this study that a pair of stereo cameras are virtually located at viewer's two eyes in an augmented reality system. Figure 2 shows the flowchart of the proposed method. First, to memorize positions of markers and natural features, the markers and natural features, both of which may be simply called features hereafter, are detected from a pair of stereo images (A in Fig. 2). In order to track the features, the positions are predicted using a pair of stereo images and an inertial sensor (B and C in Fig. 2). At the same time, the predictions are evaluated because tracking of natural features contains errors (D in Fig 2). Finally, the results above are used for estimating a model-view matrix which represents a relationship between the real and virtual coordinate systems (E in Fig. 2).

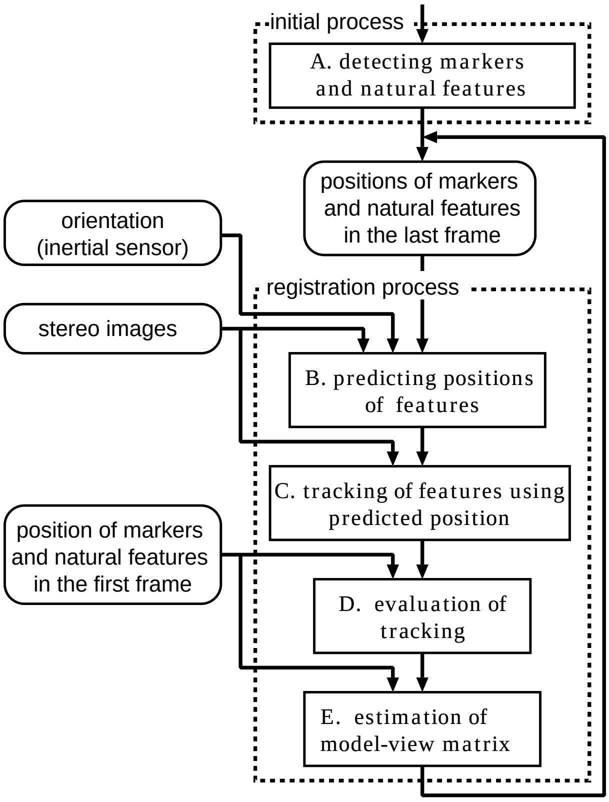
initial process
A. detecting markers and natural features
positions of markers and natural features in the last frame
orientation (inertial sensor)
registration process
stereo images
B. predicting positions of features
C. tracking of features using predicted position
position of markers and natural features in the first frame
D. evaluation of tracking
E. estimation of model-view matrix

Figure 2 Flow diagram of the registration method.

2.1 Detecting positions of markers and natural features

In the first frame, markers and natural features are detected from a pair of stereo images. The markers are detected by color matching [3]. On the other hand, the natural features are detected by using Moravec's interest operator [4]. The interest operator can detect characteristic points as natural features that are easily matched between the two consecutive images. In the proposed method, the natural features are detected from every evenly spaced region by the interest operator to distribute detected features evenly in the images. Figure 3 shows a N × M search window layout for the interest operator which is applied to the left image of stereo pair.

2.2 Predicting positions of features

2.2.1. Analysis of camera motion. Figure 4 illustrates the relationship between the positions of stereo cameras at time t and t+Δt. In general, the motions of cameras contain rotation and translation. Since the center of rotation differs from the position of stereo cameras, the translation $\mathbf{T}$ of camera can be decomposed into two translations $\mathbf{T_r}$

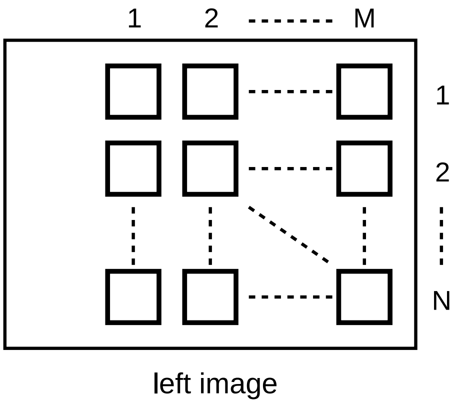

Figure 3 Search windows of natural features in the first frame.

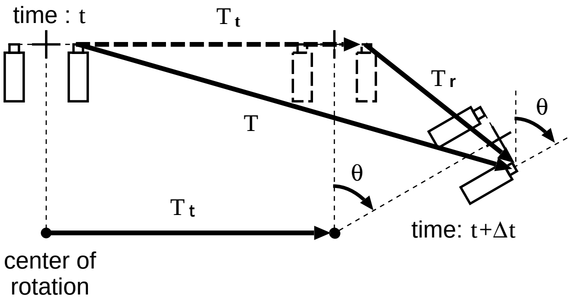

Figure 4 Decomposition of camera motion.

and $\mathbf{T_t}$. Between $\mathbf{T_t}$ and whole translation $\mathbf{T}$, the following equation stands:

$$\mathbf{T} = \mathbf{T_t} + \mathbf{T_r}, \tag{1}$$

where $\mathbf{T_r}$ is the translation caused by the rotation of the camera which is obtained by the inertial sensor, and $\mathbf{T_t}$ represents the user's viewpoint translation, that is, the translation of rotation center. In this paper, the positions of features in the next frame are predicted by estimating both $\mathbf{T_t}$ and $\mathbf{T_r}$.

2.2.2. Estimating camera motion and predicting positions of features. When θ , $\mathbf{T_t}$ and $\mathbf{T_r}$ are estimated, the positions of features in the current frame can be predicted as follows.

$$\mathbf{P}_{\mathrm{t}+\Delta\mathrm{t}} = \mathbf{RP}_{\mathrm{t}} - \mathbf{T_r} - \mathbf{T_t}, \tag{2}$$

where $\mathbf{R}$ is a transformation matrix of the rotation θ which is obtained by the inertial sensor.

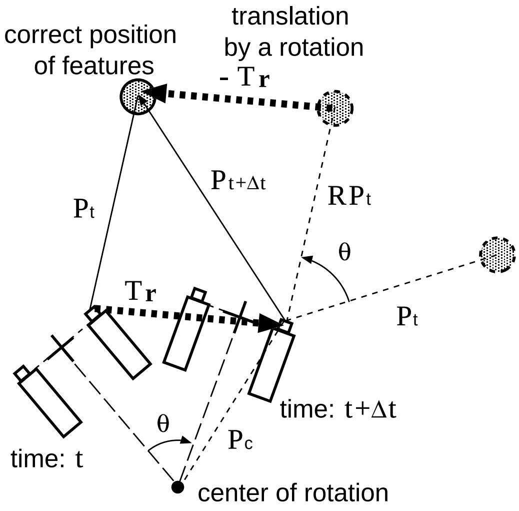

Figure 5 Translation of camera in rotation by a displacement.

.

The stereo cameras are actually translated even when the camera is rotated because of the displacement between the center of rotation and the position of stereo cameras, as shown in Figure 5. Because the displacement is constant, the translation $\mathbf{T_r}$ can be calculated with the following equation.

$$\mathbf{T_r} = \mathbf{RP}_{\mathrm{c}} - \mathbf{P}_{\mathrm{c}}, \tag{3}$$

where $\mathbf{P}_{\mathrm{c}}$ represents the relationship between the center of rotation and the position of stereo cameras. Then, the transformation matrix $\mathbf{R}$ is represented by the forementioned matrix since the rotation of camera is the same as the rotation obtained by the inertial sensor. The calculation is applied to both cameras.

Next, the translation $\mathbf{T_t}$ of the center of rotation can be determined as follows.

$$\begin{aligned} \mathbf{T_t} &= \mathbf{RP}_{\mathrm{t}} - \mathbf{T_r} - \mathbf{P}_{\mathrm{t}+\Delta\mathrm{t}} \\ &= \mathbf{P}' - \mathbf{P}_{\mathrm{t}+\Delta\mathrm{t}} \end{aligned} \tag{4}$$

The equation means that the translation of cameras $\mathbf{T_t}$ can be calculated by the relationship $\mathbf{P}_{\mathrm{t}+\Delta\mathrm{t}}$ between positions of a feature and camera in the current frame and the predicted position $\mathbf{P}'$.

The proposed method assumes that the farthest marker from cameras has small movement in adjacent images. Hence, the translation $\mathbf{T_t}$ is predicted by estimating $\mathbf{P}_{\mathrm{t}+\Delta\mathrm{t}}$ of the farthest marker from the camera.

2.3 Tracking of features

In this section, the tracking of markers and natural features with the prediction above is described.

2.3.1. Tracking of markers. This section describes two cases of features: inside or outside of the current frame. In the case where the predicted position is inside of the image, a search window is determined by the predicted position. Then, the markers are tracked by detecting marker's region in the search window based on color information. Next, the 3D positions of the markers are calculated with a stereo matching algorithm. Note that the farthest marker is tracked without the predictions.

In the case where the predicted position is outside of the image, tracking is realized by assuming that the prediction position is correct in the current frame. Therefore, the markers can be tracked continuously even when markers that have once gone outside come back into sight again.

2.3.2. Tracking of natural features. The natural features are tracked by using a standard template matching technique applied to two consecutive images. Note that a template is made from a neighboring region of the natural features in the previous frame and a similarity measure is a normalized cross correlation. When the cameras rotate on roll direction, the rotation of the template is considered in the matching process.

2.3.3. Evaluation of tracking results. The tracking results are evaluated by the following constraints because the natural feature positions drift by updating of the template. When one of the following constrains is not satisfied, the tracking is discontinued.

- Correlation between two consecutive images.

  The normalized cross correlation between two consecutive frames is used to evaluate the tracking error caused by mismatchings or occlusions by camera motion. If the correlation is under a given threshold, the tracking of natural features is discontinued.

- Epipolar constraint.

  The second constraint is the epipolar constraint, which means that corresponding points in stereo pair should exist on the epipolar line which is intersection of two image planes and a plane determined by the feature point in 3D and the two centers of lenses. If the corresponding points are not exist on the epipolar line, the tracking of natural features is discontinued.

- Displacement of 3D positions.

  The third constraint is that the 3D positions of the features in the current frame are compared to their positions recorded in the first frame. If the distance between the positions of natural features in the current frame and their positions in the first frame is larger than a threshold, the tracking of natural features is discontinued.

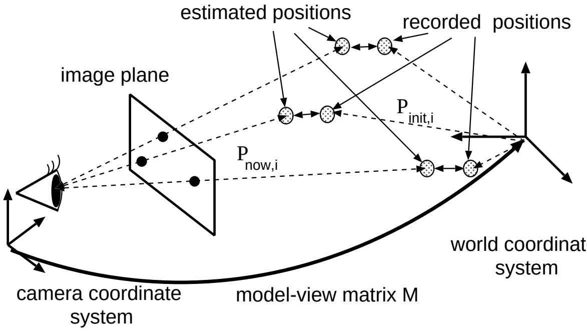

Figure 6 Registration using features.

2.4 Estimating model-view matrix using features

A model-view matrix that represents the relationship between the world and camera coordinate systems is estimated using the features. Figure 6 illustrates a relationship between the positions of features in the first and current frames. The model-view matrix is calculated by matching their 3D positions. Provided that the position of the *i*-th feature in the world coordinate system recorded in the first frame is $\mathbf{P}_{init,i}$ and the position of *i*-th feature in the camera coordinate system in the current frame is $\mathbf{P}_{now,i}$, the model-view matrix **M** can be estimated by minimizing a sum of square differences (SSD) as follows:

$$SSD = \sum_i w_i(\mathbf{P}_{init,i} - \mathbf{M}\mathbf{P}_{now,i})^2$$

where w_i is a parameter that represents the credibility of each feature. The credibility is defined by the constrains described in Section 2.3.3.

3 Implementation and Experiments

3.1 Prototype system

We have constructed a prototype of video see-through augmented reality system using two small CCD cameras (Toshiba IK-UM42) and an inertial sensor (InterSense IS-300) mounted on a HMD (Olympus Media Mask), as shown in Figure 7, for demonstrating the proposed geometrical registration algorithm. The baseline length between two cameras is set to 6.5 cm. The optical axes of the cameras are set to be parallel to the viewer's gaze direction (actually the head direction). The images captured by the cameras are fed into a graphic workstation (SGI Onyx2 IR: 16CPU MIPS R10000 195MHz) through the digital video interface (DIVO). The orientation of the camera (head) obtained by the inertial sensor is also fed into the workstation with serial interface. The incoming real world images are merged with virtual objects and output from the DIVO interface to the HMD. The hardware

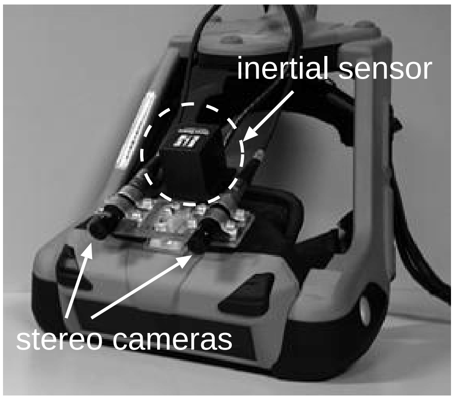

Figure 7 Appearance of stereo video see-through HMD with inertial sensor.

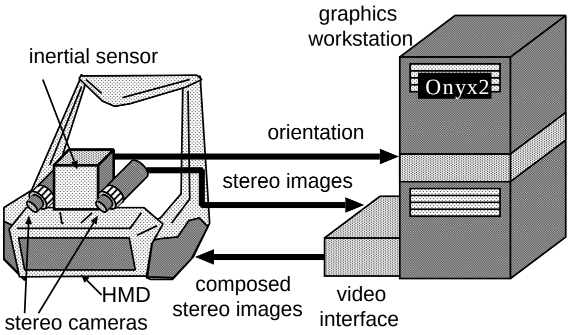

Figure 8 Configuration of prototype system.

configuration of the whole system is illustrated in Figure 8. Figure 9 shows an appearance of experiment using the prototype system.

3.2 Evaluation of effectiveness of translation prediction in marker tracking

The experiment uses four blue markers placed on a desktop as fiducial markers. Figure 10 shows the results of prediction of markers' position with translation prediction of user's head in comparison with the prediction without translation prediction. The squares of solid represent search windows with translation prediction. The dotted squares represent those without translation prediction. Note that the user's head translation is predicted using only the marker farthest from cameras. As shown in Figure 10, the case which does not consider user's head translation cannot accurately track the markers. On the other hand, it is confirmed that the proposed method can accurately predict the markers' positions by only using one tracked marker even when user's head is translated or a marker is occluded by another object.

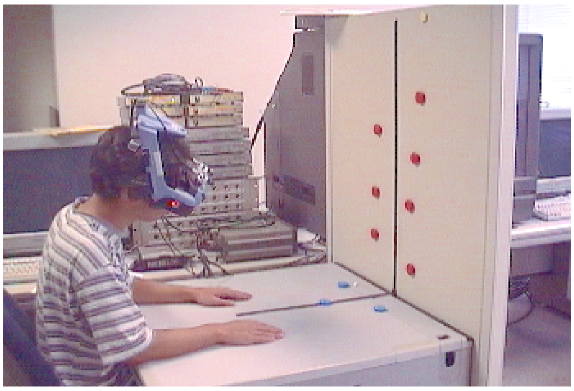

Figure 9 Appearance of the experiment.

Figure 11 shows the results of tracking markers in rapid camera motion. The squares of solid line and dotted line represent search windows with and without translation prediction, respectively. The markers in each image are tracked successfully with translation prediction, while the tracking without translation prediction is not complete because a marker in the current frame does not exist in the search window as is clearly observed in Figure 11.

The rate of miss tracking using only stereo cameras is 7.47%. The rates of miss tracking using both stereo cameras and an inertial sensor with and without considering the translation of the user's head are 1.26% and 0.34%, respectively. Note that the rate of miss tracking is defined as follows:

$$rate\ of\ miss\ tracking = \frac{number\ of\ miss\text{-}tracked\ markers}{number\ of\ frames \times number\ of\ observed\ markers}.$$

Note that the total number of frames is 150 and the number of observed markers is 4 in the experiment shown in Figure 11.

3.3 Evaluation of effectiveness of using natural features.

Figure 12 illustrates the results of natural feature tracking using the proposed method, where the mark '+' represents positions of the natural features. The marks in the top of the image sequence represents the position of natural features detected by Moravec's interest operator. The only reliabe natural features are tracked by considering the constrains mentioned in the Section 2.3.3.

Figure 13 shows results of geometric registration using both markers and natural features. The solid and dotted lines are drawn connecting the 3D positions of markers by using the estimated model-view matrix. The solid and dotted lines represent results of registration using markers and natural features and using only markers, respectively.

In the experiment, the markers and natural features captured in the right half of image are used for registration to verify the robustness of the proposal method when the markers go outside of images. As shown in Figure 13, the solid line is more accurate than the dotted line.

4 Conclusion

This paper has proposed a stereo vision-based augmented reality system with a wide range of registration. We have used natural feature points contained in images captured by a pair of stereo camera in conjunction with pre-defined fiducial markers. In addition, the method realizes a robust feature tracking by using an inertial sensor which predicts positions of features. The feasibility of the prototype system has been successfully demonstrated through experiments. Our future work will include automatic detection of new natural features that come into sight for a wider range of registartion.

Acknowledgments
This work was supported in part by Core Research for Evolutional Science and Technology (CREST) of Japan Science and Technology Corporation (JST).

References

[1] R. Azuma. A Survey of Augmented Reality. In *Presence*, volume 6(4), pages 355–385, 1997.

[2] S. Feiner, B. Maclntyre, and D. Seligmann. Knowledge-based Augmented Reality. In *Commun. of the ACM*, volume 36(7), pages 52–62, 1993.

[3] M. Kanbara, T. Okuma, H. Takemura, and N. Yokoya. A Stereoscopic Video See-through Augmented Reality System Based on Real-time Vision-based Registration. In *Proc. IEEE Virtual Reality 2000*, pages 255–262, 2000.

[4] H. Moravec. Visual Mapping by a Robot Rover,. In *Proc. 6th IJCAI*, pages 598–600, 1979.

[5] T. Ohshima, K. Satoh, H. Yamamoto, and H. Tamura. AR^2 Hockey: A Case Study of Collaborative Augmented Reality. In *Proc. VRAIS'98*, pages 14–18, 1998.

[6] T. Okuma, K. Kiyokawa, H. Takemura, and N. Yokoya. An Augmented Reality System Using a Real-time Vision Based Registration. In *Proc. ICPR'98*, volume 2, pages 1226–1229, 1998.

[7] J. Park, S. You, and U. Neumann. Natural Feature Tracking for Extendible Robust Augmented Realities. In *Proc. of Int. Workshop on Augmented Reality*, 1998.

[8] J. Rekimoto. Matrix: A Realitime Object Identification and Registration Method for Augmented Reality. In *Proc. APCHI*, pages 63–68, 1998.

[9] A. State, A. Livingston, F. Garrett, G. Hirota, and H. Fuchs. Technologies for Augmented Reality Systems: Realizing Ultrasound-Guided Needle Biopsies. In *Proc. SIGGRAPH'96*, pages 439–446, 1996.

[10] M. Uenohara and T. Kanade. Vision-Based Object Registration for Real-time Image Overlay. In *Proc. CVRMed'95*, pages 13–22, 1995.

[11] Y. Yokokohji, Y. Sugawara, and T. Yoshikawa. Accurate Image Overlay on See-through Head-mounted Displays Using Vision and Accelerometers. In *Proc. IEEE Virtual Reality 2000*, pages 247–254, 2000.

[12] S. You, U. Neumann, and R. Azuma. Hybrid Inertial and Vision Tracking for Augmented Reality Registration. In *Proc. IEEE Virtual Reality '99*, pages 260–267, 1999.

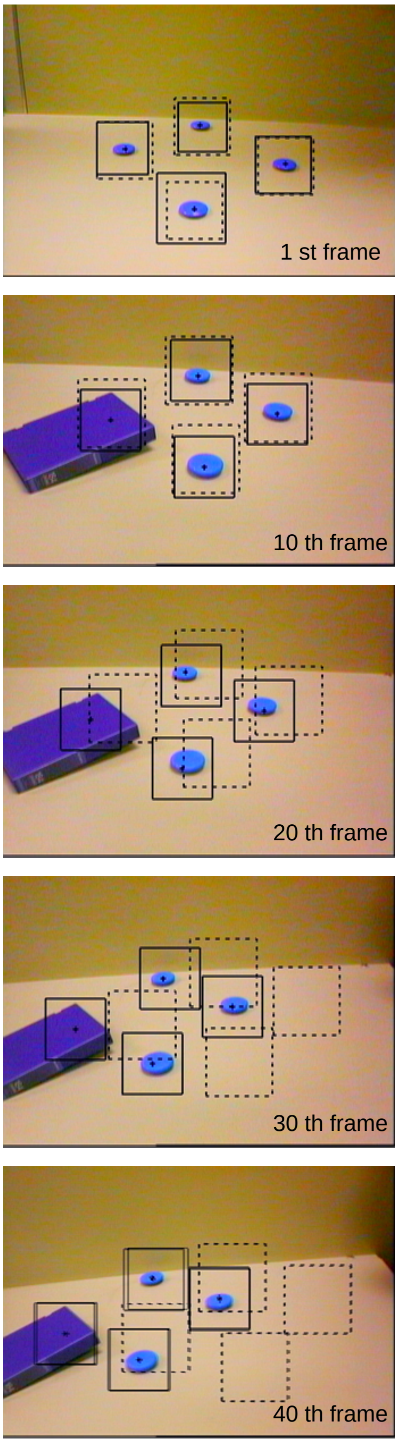

Figure 10 Position prediction of markers with and without considering a translation (solid square: search window with translation prediction, dotted square: search window without translation prediction).

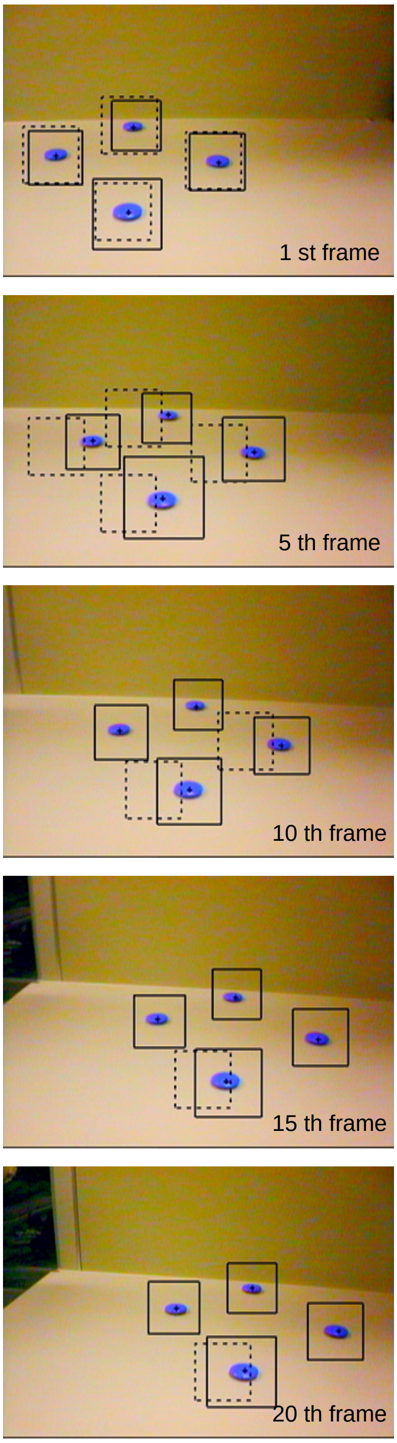

Figure 11 Marker tracking with and without position prediction in rapid motion (solid square: search window of tracked marker with translation prediction, dotted square: search window of tracked marker without translation prediction).

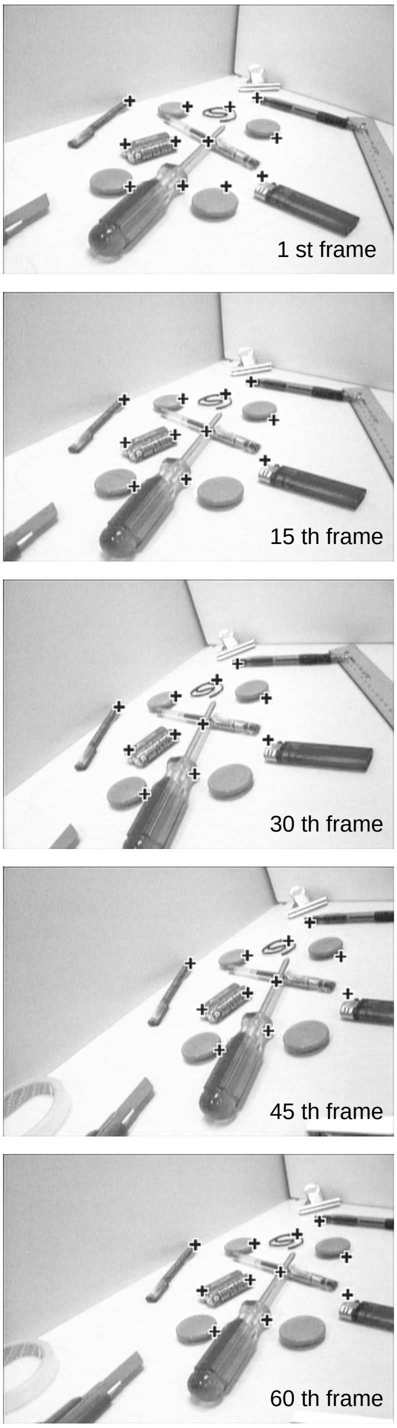

Figure 12 Result of natural feature tracking.

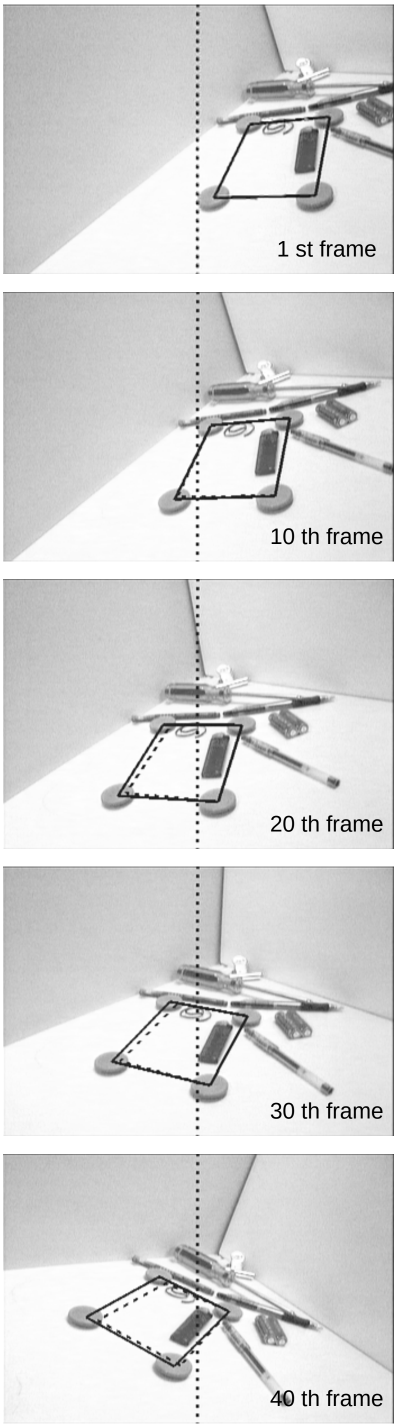

Figure 13 Result of registration using both markers and natural features (solid line) compared with that using only markers (dotted line).