IMAGE–BASED AUGMENTATION OF VIRTUAL OBJECT FOR HANDY CAMERA VIDEO SEQUENCE USING ARBITRARY MULTIPLE PLANES

Yuko Uematsu Hideo Saito

Keio University
Dept. of Information and Computer Science
Yokohama, Japan
{yu-ko, saito}@ozawa.ics.keio.ac.jp

ABSTRACT

We propose a novel vision-based registration method for Augmented Reality with merging of arbitrary multiple planes. In our approach, we require neither artificial markers nor sensors, and estimate the camera rotation and translation by an uncalibrated image sequence in which arbitrary multiple planes in the real world exist. Since the geometrical relationship of those planes is unknown, for merging of them, we assign a 3D coordinate system for each plane independently and construct projective 3D space defined by projective geometry of two reference images. By merging with the projective space, we can use arbitrary multi-planes, and achieve high-accurate registration for every position in the input images.

1. INTRODUCTION

Augmented Reality (AR) / Mixed Reality (MR) allow users to see the real world with virtual objects superimposed onto the real world. Thus AR can provide the users with more effective view [1, 2].

One of the most important issues for AR is geometrical registration between the real and the virtual world. In order to achieve correct registration, accurate measurements of the camera rotations and translations (corresponding to the user's view) are required. For the measurements, some kind of sensors such as magnetic or gyro sensors may be used. The registration by such sensors is stable against a change in light conditions and is especially effective when a camera moves rapidly. However, the rotations and translations obtained from sensors are not accurate enough to achieve perfect geometrical registration. Furthermore, the use of sensors has some limitations in practice: user's movable area, perturbation caused by the environment, and so on.

On the other hand, vision-based registration does not require any special devices except cameras. Therefore an AR system can be constructed easily. This kind of registration relies on the identification of features in the input images. Kutulakos et al. has proposed one of the earliest works of vision-based registration for real-time AR system without camera calibration [3] in which they used artificial markers and an affine camera model for simplification. Recent works extend this method to a perspective camera model and also use known-3D model and / or natural features for registration. Related works based on natural features have used various features: feature points [4, 5], edges [6] or curves [7]. However, it is also true that few features are available for registration in the real world. As a result, the augmentation becomes unstable and generates tracking jitters. For reducing such instability more effective and stronger constraints should be employed.

Registration using planes [8, 9, 10, 11] has attracted attention recently. Using planar structures of a scene gives effectively restricted conditions, because a lot of planes exist in doors or urban environment. Simon et al. have proposed related AR systems using multiple planes such as a room's floor and walls or the wall surfaces of buildings [9, 10]. In [9], they estimated the projection matrix by multiple planes which are perpendicular to the reference plane, using an uncalibrated camera. In [10], they estimated the projection matrix using a calibrated camera from multiple planes of arbitrary position and pose. In their method, the geometrical relationship between these planes and motion of the camera are calculated by bundle adjustment which is carried out over all frames.

In this paper, we propose a registration method with arbitrary multiple planes, which does not require any information on the physical relationship of the planes and can estimate the camera motion frame by frame. Fig.1 describes an overview of the proposed method. The input image sequence is taken with an uncalibrated handy video camera. The main contribution of our method is "constructing projective space" with two reference images for estimation of geometrical relationship among the planes and a camera. The constructed projective space provides the geometrical relationship of the planes even if they are not perpendicular

0-7803-9134-9/05/$20.00 ©2005 IEEE

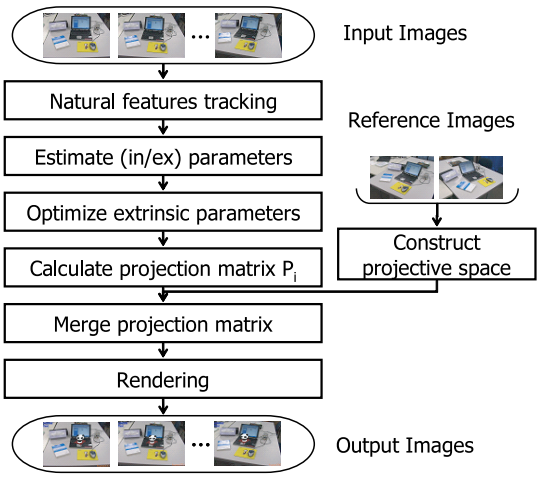

Fig. 1. Overview of the proposed method.

to each other and the intrinsic parameters of a camera are unknown.

2. REGISTRATION METHOD

2.1. Assign 3D coordinate systems

We first assign a 3D coordinate system for each plane in the 3D real world independently because the geometrical relationship among the planes is unknown. These coordinate systems will be merged by the projective space which is constructed with two reference images. As shown in fig.2, each coordinate system is defined by setting each plane to $Z = 0$. This is for computing a homography and a projection matrix from each plane. The detail of the computation is described in the next section.

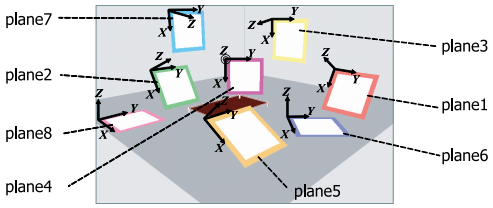

Fig. 2. Example of assigning 3D coordinate systems.

2.2. Calculate projection matrix (estimate intrinsic / extrinsic parameters)

Natural feature points are tracked using the KLT-feature-tracker [12] for the input image sequence in which n planes exist. We assume that the extracted feature points are segmented into areas of planes. (In this paper, we do not focus on the method for the segmentation, so we segment them manually.) Using the features on each real world plane, n homographies among the real planes and the input image plane are computed independently. Next, n projection matrices that relate the 3D coordinate systems on the real planes to the image are computed by extending the homographies from 2 dimensions to 3 dimensions via the following method.

When each real plane's Z coordinate is set to 0, the relationship between the real plane and the image plane is

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \simeq \boldsymbol{P} \begin{pmatrix} X \\ Y \\ 0 \\ 1 \end{pmatrix} \simeq \begin{bmatrix} p_{11} & p_{12} & p_{14} \\ p_{21} & p_{22} & p_{24} \\ p_{31} & p_{32} & p_{34} \end{bmatrix} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \quad (1)$$

The 3×3 matrix (called $\hat{\boldsymbol{P}}$), which is the deleted third column vector of $\boldsymbol{P}$, is equivalent to a planar homography $\boldsymbol{H}$. Thus, when the homography is calculated, we can compute the projection matrix from it.

$\boldsymbol{P}$ is also represented by the intrinsic and extrinsic parameters

$$\boldsymbol{P} = \boldsymbol{A}\,[\boldsymbol{R} \mid \boldsymbol{t}] = \boldsymbol{A}\,[\boldsymbol{r_1}\ \boldsymbol{r_2}\ \boldsymbol{r_3}\ \boldsymbol{t}], \quad \hat{\boldsymbol{P}} = \boldsymbol{H} = \boldsymbol{A}\,[\boldsymbol{r_1}\ \boldsymbol{r_2}\ \boldsymbol{t}]$$

$$\therefore \boldsymbol{A}^{-1}\boldsymbol{H} = [\boldsymbol{r_1}\ \boldsymbol{r_2}\ \boldsymbol{t}] \quad (2)$$

By fixing the skew to 0, the aspect ratio to 1 and the principal point to the center of the image, the intrinsic parameters can be defined as in eq.(3). According to property of rotation matrix $\boldsymbol{R}$, the inner product of $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ is equal to 0. We can calculate the focal length f as eq.(4).

$$\boldsymbol{A} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \qquad \begin{matrix} (c_x, c_y) & : & \text{principal point} \\ f & : & \text{focal length} \end{matrix} \quad (3)$$

$$f^2 = \frac{(h_{11} - c_x h_{31})(h_{12} - c_x h_{32}) + (h_{21} - c_y h_{31})(h_{22} - c_x h_{32})}{-h_{31} h_{32}} \quad (4)$$

The extrinsic parameters of a camera consist of a rotation matrix $\boldsymbol{R}$ and a translation vector $\boldsymbol{t}$. Since $\boldsymbol{r_1}$, $\boldsymbol{r_2}$ (the first and second column vectors of $\boldsymbol{R}$) and $\boldsymbol{t}$ are already known, we only need to estimate $\boldsymbol{r_3}$. According to the property of $\boldsymbol{R}$, we can compute $\boldsymbol{r_3}$ from the cross product of $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$. Therefore, $\boldsymbol{R}$ is

$$\boldsymbol{R} = [\boldsymbol{r_1}\ \boldsymbol{r_2}\ (\boldsymbol{r_1} \times \boldsymbol{r_2})] \quad (5)$$

The extrinsic parameters are optimized by the steepest descent method.

2.3. Merge projection matrices

Each projection matrix is reliable around its corresponding plane. However, as the position of a virtual object moves

away from each plane, the accuracy becomes lower. Therefore, we merge the projection matrices in order to compute one accurate matrix over the whole image and reduce registration errors.

Two reference images are chosen from the input image sequence (usually the first image and last image) or taken in advance to construct a 3D projective space. The projective space defined by the reference images is a common 3D coordinate system for the whole input image sequence. Eq.6and fig.3 show the relationship of the real world, the projective space and the image coordinate systems.

$$\boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{PI}} = \boldsymbol{P}_{\boldsymbol{k}} \boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{WP}-1} \tag{6}$$

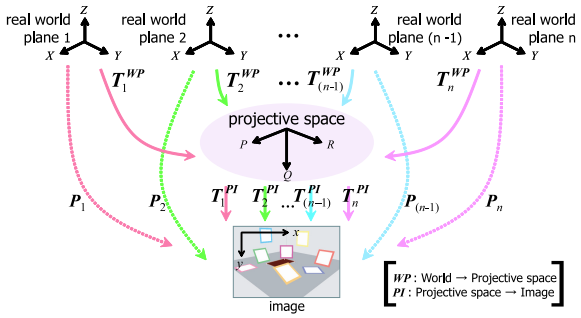

Fig. 3. Relationship among 3 coordinate sysmtes.

2.3.1. *Construct projective space*

We construct projective space by "projective reconstruction" as shown in fig.4. By epipolar geometry between the reference images (cameras), the relationship between the projective space and the reference images is given by

$$\boldsymbol{P}_{\boldsymbol{A}} = [\mathbf{I} \mid \mathbf{0}], \quad \boldsymbol{P}_{\boldsymbol{B}} = [\boldsymbol{M}\boldsymbol{e}_{\boldsymbol{B}}], \quad \boldsymbol{M} = -\frac{[\boldsymbol{e}_{\boldsymbol{B}}]_{\times} \boldsymbol{F}_{\boldsymbol{AB}}}{\|\boldsymbol{e}_{\boldsymbol{B}}\|^2} \tag{7}$$

where $\boldsymbol{F}_{\boldsymbol{AB}}$ is a fundamental matrix of image A to B, and $\boldsymbol{e}_{\boldsymbol{B}}$ is an epipole on the image B. Consider $\boldsymbol{C}_{\boldsymbol{p}}$ as a point in the projective space, $\boldsymbol{C}_{\boldsymbol{A}}(u_A, v_A)$ as on the image A, $\boldsymbol{C}_{\boldsymbol{B}}(u_B, v_B)$ as on the image B, we can write

$$\boldsymbol{K}\boldsymbol{C}_{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{p}_{\boldsymbol{A}}^{1} - u_A \boldsymbol{p}_{\boldsymbol{A}}^{3} \\ \boldsymbol{p}_{\boldsymbol{A}}^{2} - v_A \boldsymbol{p}_{\boldsymbol{A}}^{3} \\ \boldsymbol{p}_{\boldsymbol{B}}^{1} - u_B \boldsymbol{p}_{\boldsymbol{B}}^{3} \\ \boldsymbol{p}_{\boldsymbol{B}}^{2} - v_B \boldsymbol{p}_{\boldsymbol{B}}^{3} \end{bmatrix} = \mathbf{0} \tag{8}$$

$\boldsymbol{p}^{\boldsymbol{i}}$ is the i th column vector of $\boldsymbol{P}$. Then, we obtain $\boldsymbol{C}_{\boldsymbol{p}} \simeq [p, q, r, 1]^{\top}$ by the singular value decomposition of $\boldsymbol{K}$.

2.3.2. *Calculate* $\boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{WP}}$

Consider $\boldsymbol{C}_{\boldsymbol{W}}$ as a point on the kth plane in the real world, and $\boldsymbol{C}_{\boldsymbol{P}}$as a point in the projective space, the relationship of between the two coordinate systems is

$$\boldsymbol{C}_{\boldsymbol{P}} \simeq \boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{WP}} \boldsymbol{C}_{\boldsymbol{W}} \tag{9}$$

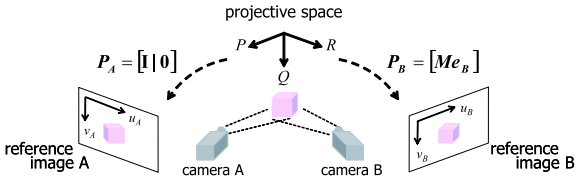

Fig. 4. Projective space by projective reconstruction.

Since $\boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{WP}}$ is 4×4 matrix, we can compute this matrix by 5 (or more) corresponding points, in which any combination of 3 points must not be colinear and 4 points must not also be coplanar.

2.3.3. *Calculate* $\boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{PI}}$

When $\boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{WP}}$ is known, we can compute $\boldsymbol{T}_{\boldsymbol{k}}^{\boldsymbol{PI}}$ by eq.(6), so $\boldsymbol{T}_{\boldsymbol{1}}^{\boldsymbol{PI}} \sim \boldsymbol{T}_{\boldsymbol{n}}^{\boldsymbol{PI}}$ are computed for each plane as fig.3. These matrices should be the same, if every measurement is completely accurate. However, they are slightly different from each other because of inaccuracy errors in previous procedures, such as tracking feature points. Therefore the merging of these matrices will provide a more accurate projection matrix between the projective space and the image plane. We assume that by merging them we can approximately provide the merged matrix, because they are almost the same.

In the merging computation, we take weights $w_1 \sim w_n$, which are determined according to the distance from the position of the virtual object to 3D coordinate origin of each plane.

$$\boldsymbol{T}^{\boldsymbol{PI}} = \frac{1}{n} \left[w_1 \ \cdots \ w_n \right] \left[\boldsymbol{T}_{\boldsymbol{1}}^{\boldsymbol{PI}}, \ \ldots, \ \boldsymbol{T}_{\boldsymbol{n}}^{\boldsymbol{PI}} \right]^{\top} \tag{10}$$

Such merging of the matrices can be regarded as merging of the tracked planes.

3. EXPERIMETAL RESULTS

We implement the Augmented Reality system based on our method using only a PC (OS:Windows 2000, CPU:Intel Pentium Ⅳ 3.20GHz) and a CCD camera (SONY DCR-TRV900). The image's resolution in all the experiments is 720×480 pixels, and graphical views of a virtual object are rendered using OpenGL.

The overlaid result images produced by the augmentation are shown in fig.5. The 3 planes (a mouse pad, a book, and a tissue box) are used and a virtual object is overlaid on the notebook PC. As shown in the figure, our approach can superimpose a virtual object onto the image sequence successfully.

Next, in order to evaluate the registration accuracy in our method, we perform the same implementation for the synthesized image sequence rendered with OpenGL. Since

(a) *Frame0* (b) *Frame33* (c) *Frame74* (d) *Frame99*

Fig. 5. Overlaid image sequence of a virtual object.

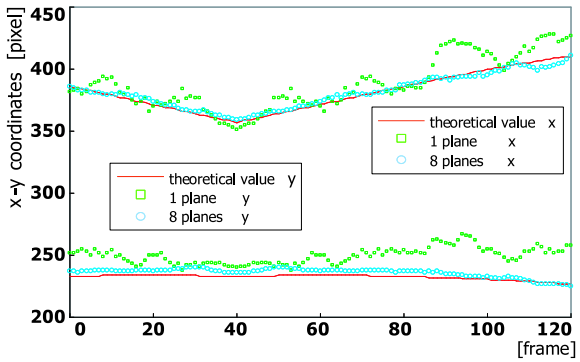

Fig. 6. Comparison of x-y coordinates between 1 plane and 8 planes with theoretical values.

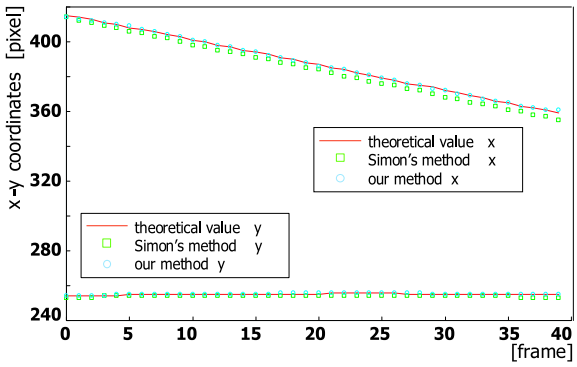

Fig. 7. Comparison of x-y coordinates between related method and our method with theoretical values.

we have to know the exact position and pose of a camera to evaluate accuracy, we use the synthesized images. Fig.6 shows that the result by 8 planes has less registration errors and jitters than using only 1 plane, in spite of no information about the relationship of the planes. This suggests that increasing the number of planar structures in the scene can improve the registration accuracy.

We also evaluate the proposed method by comparing with a related work by Simon [9], in which multiple planes need to be perpendicular to the reference plane. For the comparison, we apply the image sequence, which has 3 orthogonal planes, to Simon's method and our method, and evaluate the registration accuracy. The result of the evaluation is shown in fig.7. Even though our method does not require any geometrical information of the plane, our method achieves almost the same accuracy as Simon's method, in which the planes need to be perpendicular to the reference plane.

4. CONCLUSION

A geometrical registration method for Augmented Reality with uncalibrated camera based on multiple planes has been proposed in this paper. The planes do not need to be perpendicular to each other. This means that any planes at an arbitrary position and pose can be used for registration. Furthermore the registration can be performed frame by frame without using all the frames in input image sequence. Thus we can construct the AR system easily, and overlay a virtual object onto the image sequence correctly.

5. REFERENCES

[1] R. T. Azuma, "A survey of augmented reality," *Presence*, pp. 355–385, 1997.

[2] R. T. Azuma, "Recent advances in augmented reality," *IEEE Computer Graphics and Applications*, vol. 21, no. 6, pp. 34–47, Nov-Dec, 2001.

[3] Kiriakos N. Kutulakos and James R. Vallino, "Calibration-free augmented reality," *IEEE Trans. on Visualization and Computer Graphics.*, vol. 4, no. 1, pp. 1–20, Jan-Mar, 1998.

[4] U. Neumann and S. You, "Natural feature tracking for augmented reality," *IEEE Trans. on Multimadia*, vol. 1, no. 1, pp. 53–64, 1999.

[5] K. W. Chia, A. Cheok, and S. J. D. Prince, "Online 6 dof augmented reality registration from natural features," in *Proc. of the ISMAR*, 2002, pp. 305–313.

[6] D. Gennery, "Visual tracking of known three dimensional objects," *International Journal of Computer Vision*, vol. 7, no. 3, pp. 243–270, 1992.

[7] G. Simon and M. Berger, "A two-stage robust statistical method for temporal registration from features of various type," in *Proc. of 6th ICCV*, 1998, pp. 261–266.

[8] G. Simon, A. Fitzgibbon, and A. Zisserman, "Markerless tracking using planar structures in the scene," in *Proc. of the ISAR*, 2000, pp. 120–128.

[9] G. Simon and M. Berger, "Reconstructing while registering: a novel approach for markerless augmented reality," in *Proc. of the ISMAR*, 2002, pp. 285–294.

[10] G. Simon and M. O Berger, "Real time registration known or recovered multi-planar structures: application to ar," in *Proc. of the BMVC*, 2002, pp. 567–576.

[11] V. Ferrari, T. Tuytelaars, and L. V. Gool, "Makerless augmented reality with a real-time affine region tracker," in *Proc. of the ISAR*, 2001, pp. 87–96.

[12] J. Shi and C. Tomasi, "Good features to track," *IEEE Conf. on CVPR*, pp. 593–600, 1994.