VISION-BASED ROBUST CALIBRATION FOR OPTICAL SEE-THROUGH HEAD-MOUNTED DISPLAYS

Naoya Makibuchi, Haruhisa Kato and Akio Yoneyama

KDDI R&D Laboratories Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, Japan
Email: {na-makibuchi, hkato, yoneyama}@kddilabs.jp

ABSTRACT

We propose *Vision-based Robust Calibration (ViRC)* method for OSTHMDs equipped with a camera. In the ViRC method, calibration parameters are decomposed into off-line parameters that remain constant relative to the positional relationship between the camera and the virtual screen, and on-line parameters related to the user's eye. Calculating the off-line parameters beforehand reduces the number of unknown parameters in the on-line phase, giving robust protection against the user's misalignments during calibration. In the off-line phase, the approximate position of the user's eye is calculated using the PnP algorithm. In the online phase, the actual position of the user's eye is estimated from the approximate one by non-linear minimization. In our experiments, we show that the ViRC method can decrease reprojection error by as much as 83% compared with the conventional method based on the DLT algorithm.

Index Terms— Optical See-Through Head-Mounted Display (OSTHMD), HMD Calibration, Augmented Reality

1. INTRODUCTION

Head-mounted displays fall into two categories: video see-through head-mounted displays (VSTHMDs) and optical see-through head-mounted displays (OSTHMDs). Whereas VSTHMDs provide users with the images of the real world captured by the camera, OSTHMDs allow users to observe the real world through the semi-transparent display. The major difference between them is whether scenes input to the user's eye are available as image data. The augmented reality (AR) applications using VSTHMDs are capable of identifying the positions of real objects on the image screen, but have problems with its size and weight, and range of view. On the other hand, those using OSTHMDs have a small and lightweight body, and wide range of view, but difficulty in identifying the positions of real objects on the virtual screen because the image data of user's scenes are unavailable.

When using an OSTHMD, some tracking device must be attached to it for calculating the pose of the user's eye and then for calibration. The goal of OSTHMD calibration is to calculate the positional relationship between the tracking device fixed on the OSTHMD and the user's eye. The relationship is called the calibration parameter and remains constant as long as the user does not move his OSTHMD glasses.

The OSTHMD's calibration is generally performed as follows: users move a cross-hair cursor and then click a mouse button at the positions of real objects on the virtual screen, which is called alignment. The accuracy of calibration depends considerably on the accuracy of the user's alignment. The purpose of this study is to propose a novel vision-based calibration method that is highly robust against user alignment errors.

2. RELATED WORK

Studies of OSTHMD calibration have been conducted with various types of tracking systems: magnetic-based systems [1, 2], ultrasonic-based systems [3], infra-red based systems [4, 5] and vision-based systems [6]. Among these, all except the vision-based systems generally provide high-accuracy tracking and high-speed processing. However, those systems require special, dedicated calibration aids (sensor, transmitter or spherical retro-reflective marker), which lead to high initial cost and limited workspace.

In contrast, the vision-based system is a simple configuration requiring only a fiducial marker and not requiring any additional equipment for calibration. However, there have been few studies of vision-based calibration because of its difficulty. Kato *et al.* have proposed an approach to OSTHMD calibration using the traditional fiducial marker used in the ARToolKit [6]. The approach calculates the transformation matrix from the 3-dimensional (3D) camera coordinate system to the 2-dimensional (2D) virtual screen coordinate system, using the Direct Linear Transformation (DLT) algorithm with the user's alignments. The DLT algorithm requires a minimum of six 2D-3D correspondences, but is very sensitive to mis-correspondences. In practice, it is difficult for users to click a mouse button at the exact positions of multiple targets on the virtual screen during calibration.

In this paper, we propose a vision-based calibration

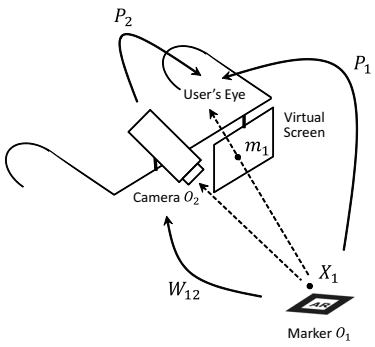


Fig. 1. Configuration of the ViRC method

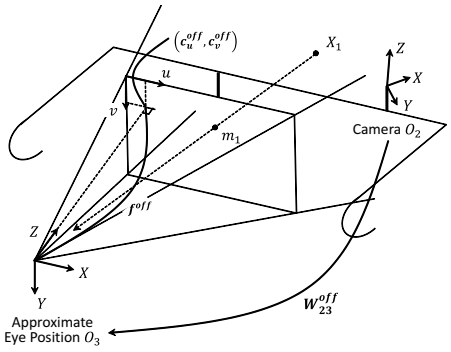


Fig. 2. Off-line calibration of the ViRC method

method that is robust against the user misalignments. The novelty of the proposed method is to apply the PnP algorithm [7] to OSTHMD calibration. In general, the PnP algorithm provides higher-accuracy camera pose than the DLT algorithm. However, the PnP algorithm cannot be applied directly to the OSTHMD calibration because the exact intrinsic parameters are not available in advance. In the proposed method, introducing a novel mathematical model that corrects the approximate intrinsic parameters enables the proposed method to employ the PnP algorithm. Consequently, it is capable of inheriting the high performance of the PnP algorithm.

3. PROPOSED METHOD: VIRC

We develop *Vision-based Robust Calibration (ViRC)* method for the AR system using monocular OSTHMDs equipped with a camera. The configuration of the ViRC method is shown in Fig. 1. Note that the virtual screen is depicted under the glasses' frame for convenience, but in practice, it is positioned the focal length away. The projection matrix P_1 of the 3D point X_1 on the marker coordinate system onto the 2D point m_1 on the virtual screen is decomposed as follows:

$$sm_1 = P_1 X_1 \quad (1)$$
$$= P_2 W_{12} X_1 \quad (2)$$

where s is a scale factor, and W_{12} is a transformation matrix from the marker coordinate system O_1 to the camera coordinate system O_2, and P_2 is a projection matrix from O_2 to the virtual screen. W_{12} is calculated by the camera pose estimation algorithm of the ARToolKit. P_2 has 11 DoF parameters that remain constant under the position of the user's eye.

While P_2 is calculated at a time in the conventional DLT-based method, it is identified step-by-step in the ViRC method in two phases: the off-line phase and the on-line phase. Decreasing the number of unknown parameters in the on-line phase permits a more robust calibration compared with the conventional method. The coordinate system of the user's eye cannot be transformed directly from O_2, so it is calculated via some intermediate coordinate systems from O_2. P_2 is decomposed into four matrices according to Eq. 3. The details of each matrix are explained in the following sections.

$$P_2 = A^{on} W_{45}^{on} W_{34}^{on} W_{23}^{off} \quad (3)$$

3.1. Off-line phase

In the off-line phase, the approximate position of the user's eye is calculated by the Perspective-n-Point (PnP) algorithm [7], which is used to calculate camera pose from the camera intrinsic parameters and a minimum of four 2D-3D correspondences. The model of this phase is shown in Fig. 2.

Assuming that the virtual screen is the image plane, the normal from the user's eye to the virtual screen is the optical axis, and the user's eye is the optical center in the perspective projection model, let A^{off} be the intrinsic parameter in this phase, and be denoted as follows:

$$A^{off} = \begin{pmatrix} f^{off} & 0 & c_u^{off} \\ 0 & f^{off} & c_v^{off} \\ 0 & 0 & 1 \end{pmatrix} \quad (4)$$

where f^{off} is the distance between the user's eye and the virtual screen, and (c_u^{off}, c_v^{off}) is the point on the virtual screen coordinate (u, v) where the normal and the virtual screen intersect. Three parameters are set as the approximate values computed according to the specification of OSTHMDs. For example, they are computed in the experiments as follows: when using OSTHMDs that place f millimeters ahead the virtual screen whose resolution is $w \times h$ pixels and the length of whose diagonal is d millimeters, f^{off}, c_u^{off} and c_v^{off} are respectively set as fk^{off}, $\frac{w}{2}$ and $\frac{h}{2}$ where $k^{off} = \frac{\sqrt{w^2+h^2}}{d}$.

Let W_{23}^{off} be a transformation matrix from the camera coordinate system O_2 to the approximate coordinate system O_3 corresponding to the user's eye. It is calculated by the PnP algorithm according to Eq. 5 with a minimum of four 2D-3D correspondences obtained through the user's alignments.

$$sm_1 = A^{off} W_{23}^{off} W_{12} X_1 \quad (5)$$

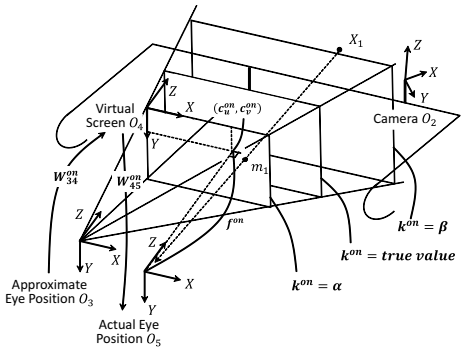
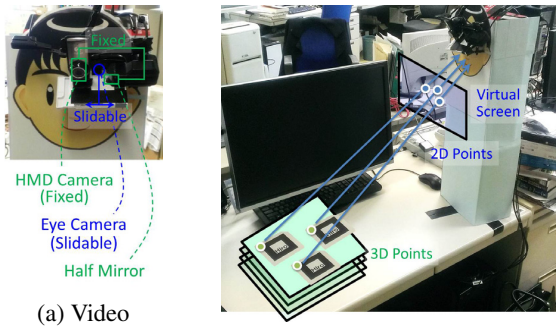


Fig. 3. On-line calibration of the ViRC method



(a) Video see-through system

(b) 2D-3D correspondences

Fig. 4. Experiment environment

Then, the off-line parameters in the ViRC method are defined as f^{off}, c_u^{off}, c_v^{off} and W_{23}^{off}. Note that the position of the user's eye is not always estimated accurately because of the approximate intrinsic parameter.

3.2. On-line phase

In the on-line phase, the actual position of the user's eye is estimated from the approximate one by non-linear minimization under the given off-line parameters. Let W_{34}^{on} be the transformation matrix from O_3 to the virtual screen coordinate system O_4, and W_{45}^{on} be the transformation matrix from O_4 to the actual position of the user's eye coordinate system O_5. W_{34}^{on} and W_{45}^{on} are introduced as correction terms converting the approximate position of the user's eye into the accurate one. The model of this phase is shown in Fig. 3.

W_{34}^{on} is denoted as follows:

$$W_{34}^{on} = \begin{pmatrix} 1 & 0 & 0 & \frac{c_u^{off}}{k^{on}} \\ 0 & 1 & 0 & \frac{c_v^{off}}{k^{on}} \\ 0 & 0 & 1 & -\frac{f^{off}}{k^{on}} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6}$$

where k^{on} is the pixel density that is introduced as one of the on-line parameters in the ViRC method. Estimating k^{on} means identifying the position of the virtual screen in the view frustum calculated in the off-line phase. As examples, the positions of the virtual screen with $k^{on} = \alpha$ and $k^{on} = \beta$ are shown in Fig. 3. Ideally, k^{on} is set as *true value* of Fig. 3.

The position of the user's eye relative to the virtual screen in the on-line phase is, in the same way as in Eq. 4, denoted by A_{on} as follows:

$$A^{on} = \begin{pmatrix} f^{on} & 0 & c_u^{on} \\ 0 & f^{on} & c_v^{on} \\ 0 & 0 & 1 \end{pmatrix} \tag{7}$$

where f^{on} and (c_u^{on}, c_v^{on}) are the remaining on-line parameters in the ViRC method. Then, W_{45}^{on} is denoted as follows:

$$W_{45}^{on} = \begin{pmatrix} 1 & 0 & 0 & -\frac{c_u^{on}}{k^{on}} \\ 0 & 1 & 0 & -\frac{c_v^{on}}{k^{on}} \\ 0 & 0 & 1 & \frac{f^{on}}{k^{on}} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{8}$$

As above, the projection formula of X_1 onto m_1 in the ViRC method is given by substituting Eqs. 3, 6, 7, 8 and W_{23}^{off} in Eq. 5 into Eq. 2. The f^{on}, c_u^{on}, c_v^{on} and k^{on} of the unknown on-line parameters are estimated by the Levenberg-Marquardt algorithm to minimize reprojection error through the projection formula, when the initial values are set to be the values of f^{off}, c_u^{off}, c_v^{off} and k^{off} used in the off-line phase. This is calculated with a minimum of four 2D-3D correspondences because of the number of unknown parameters. Thus, the ViRC method estimates the positions of both the user's eye and the virtual screen simultaneously under a small number of unknown parameters.

4. EXPERIMENTS

Evaluating an OSTHMD's calibration has also been a troublesome problem because only the user wearing it can observe the superimposed scenes from the calibration result. In this paper, we develop a video see-through system instead, as shown in Fig. 4, which permits the recording of superimposed scenes, and in addition, the quantitative evaluation of OSTHMDs independently of the user's skill at alignment.

As shown in Fig. 4 (a), we used AiRScouter made by Brother as a monocular OSTHMD and implemented QuickCam Pro for Notebooks made by Logicool. Additionally, we placed another camera in the position of the user's eye (hereinafter called "eye camera"), a Logicool HD Webcam C525, mounted so as to be slidable independently of the OSTHMD to simulate use by another user.

The video see-through system is fixed on a desk as shown in Fig. 4 (b). In practice, the alignments are performed

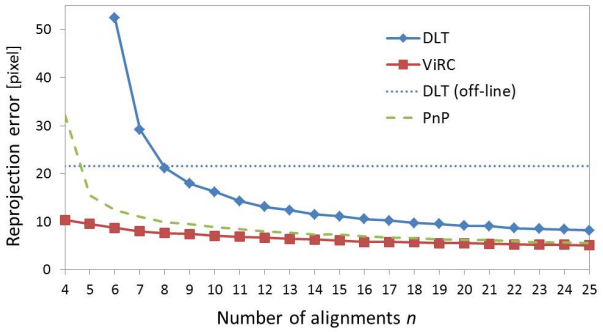
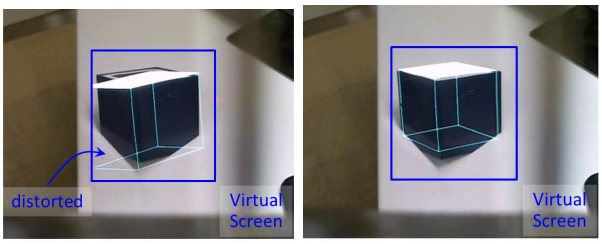


Fig. 5. Reprojection error



(a) DLT (error = 52.4 pixels) (b) ViRC (error = 8.8 pixels)

Fig. 6. Superimposed cube

through the user's mouse operation, but in the experiments, conducted automatically for markers at various distances as follows: the marker (3D point) is placed at some elevation above the desk, then the position of its corresponding projected point (2D point) on the virtual screen is identified by applying the marker detection algorithm used in the ARToolKit onto images captured by the eye camera (a procedure henceforth called "automatic alignment"). In this way, accurate 2D-3D correspondences are available that are independent of the user's ability to perform alignment and aim at markers at different distances.

First, the off-line calibration for the ViRC method is performed when the eye camera is in a certain position. There is no off-line phase in the conventional DLT-based method. The 30 correspondences are obtained through automatic alignments, then Gaussian noise with zero mean and standard deviation $\sigma = 10$ pixels is added to the coordinates of the 2D points on the assumption that there are user misalignments. Next, after the eye camera is slid about 5 millimeters, on-line calibrations using both the ViRC method and the conventional method are performed in the same way with a maximum of 25 correspondences with noise. This completes both calibrations. Last, an additional 30 correspondences are collected for use in testing both calibrations.

Fig. 5 shows the median of reprojection errors during 500 trials using each calibration method under n alignments. The processing times of the ViRC method and the conventional DLT-based method were 1.93 ms and 0.05 ms, respectively. Both methods were executed on Intel Core i7-2600K 3.4 GHz. The ViRC method requires more computational cost than the DLT method because of iterative computation, but the processing time is sufficiently practical because calibration is executed only once after the user wears the HMD glasses.

"DLT (off-line)" is the calibration result of the DLT method with 30 alignments obtained in the off-line phase, which simulates the case in which one user takes the calibration result performed by another user. "PnP" is the calibration result of the PnP method with on-line alignments using the same approximate intrinsic parameter used in the ViRC method. The ViRC method achieved more positive results overall against both the DLT method and the PnP method. Specifically, compared with the DLT method, it could decrease reprojection error by 83% when $n = 6$, from 52.4 pixels to 8.8 pixels. The reason for comparing with $n = 6$ is that the minimum value of n required for the ViRC method is 4, while that required for the DLT method is 6. Fig. 6 shows a superimposition of the computer graphic of a cube onto the virtual screen using each of its each calibration results when $n = 6$. The shape of the superimposed cube seems to have no error in the ViRC method while it seems to be distorted in the DLT method. The demonstration movie is available here[1].

Both the ViRC method and the DLT method are defined as estimating both the position and posture of the user's eye. However, the ViRC method estimates only the position of the user's eye in the on-line phase on the assumption that the posture of the user's eye against the virtual screen is invariant. Estimating the accurate posture by the PnP algorithm in the off-line phase produces a robust calibration in the on-line phase. The better result from the PnP method than from the DLT method means that misalignment (simulated by noise) has a more significant impact on the reprojection error than does the approximation error of the intrinsic parameter. The reason the ViRC method gets a better result than the PnP method is that the position of the user's eye is modified from the initial approximate intrinsic parameter to an accurate value.

5. CONCLUSION

In this paper, we proposed the ViRC method for OSTHMDs equipped with a camera. The ViRC method is a simple configuration requiring only a traditional fiducial marker and is quite robust against user misalignments during calibration. In the experiments, we showed that it could decrease reprojection error by up to 83% compared with the conventional method based on the DLT algorithm.

[1] http://www.youtube.com/watch?v=44eWZwhEn68

6. REFERENCES

[1] M. Tuceryan, Y. Genc, and N. Navab, "Single-point active alignment method (spaam) for optical see-through hmd calibration for augmented reality," *Presence: Teleoperators and Virtual Environments*, vol. 11, no. 3, pp. 259–276, 2002.

[2] Y. Genc, M. Tuceryan, and N. Navab, "Practical solutions for calibration of optical see-through devices," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR '02)*, 2002, pp. 169–175.

[3] C.B. Owen, J. Zhou, A. Tang, and F. Xiao, "Display-relative calibration for optical see-through head-mounted displays," in *Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR '04)*, 2004, pp. 70–78.

[4] S.J. Gilson, A.W. Fitzgibbon, and A. Glennerster, "Spatial calibration of an optical see-through head-mounted display," *Journal of Neuroscience Methods*, vol. 173, no. 1, pp. 140–146, 2008.

[5] F. Kellner, B. Bolte, G. Bruder, U. Rautenberg, F. Steinicke, M. Lappe, and R. Koch, "Geometric calibration of head-mounted displays and its effects on distance estimation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 4, pp. 589–596, 2012.

[6] H. Kato and M. Billinghurst, "Marker tracking and hmd calibration for a video-based augmented reality conferencing system," in *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR '99)*, 1999, pp. 85–94.

[7] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.