Recovering 3-D gaze scan path and scene structure from inside-out camera

Yuto Goto[1] , Hironobu Fujiyoshi[2]
Dept. of Computer Science, Chubu University
1200 Matsumoto, Kasugai, Aichi 487-8501
Japan
supica@vision.cs.chubu.ac.jp[1],
hf@cs.chubu.ac.jp[2]

ABSTRACT

First-Person Vision (FPV) is a wearable sensor that takes images from a user's visual field and interprets them, with available information about the user's head motion and gaze, through eye tracking [1]. Measuring the 3-D gaze trajectory of a user moving dynamically in 3-D space is interesting for understanding a user's intention and behavior. In this paper, we present a system for recovering 3-D scan path and scene structure in 3-D space on the basis of ego-motion computed from an inside-out camera. Experimental results show that the 3-D scan paths of a user moving in complex dynamic environments were recovered.

Keywords

3-D gaze scan path, inside-out camera, ego-motion

1. INTRODUCTION

The human eye enables a person to instantly absorb information and act accordingly. As a consequence, information on a person's gaze, which reveals what objects in the outside world a person is looking at, can be valuable in determining that person's behavioral intentions. Noton et al. discovered that similar scan paths are used when a person is shown the same object at different times [2]. Obtaining information on eye movement in this way can therefore be expected to clarify higher cognitive processes in humans [3].

First-Person Vision (FPV) is a wearable sensor that takes images from a user's visual field and interprets them, with available information about the user's head motion and gaze, through eye tracking [1]. Eye tracking systems have been commercially available for some time [4, 5, 6, 7]. When these head-mounted eye-tracking systems are used, point-of-regard (POR) is generally measured as a point on the image plane. To acquire a 3-D POR, an inside-out camera system was proposed by Shimizu et al. [8]. The inside-out camera system can a recover 3-D POR only under static state. We are interested in the gaze measurement of a user moving dynamically in 3-D space. It is therefore difficult to analyze 3-D gaze trajectories quantitatively due to user's head movements.

AH '13, March 07 - 08 2013, Stuttgart, Germany.

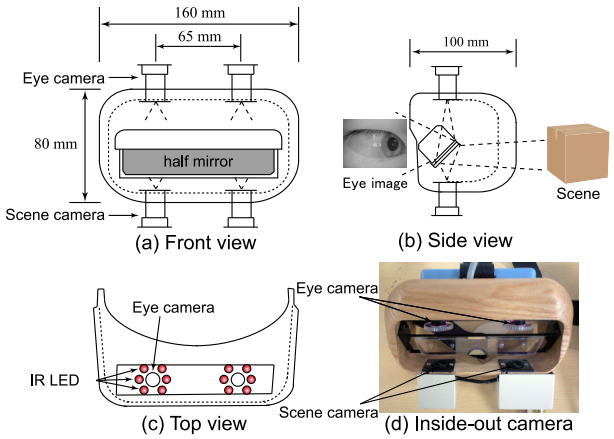


Figure 1: Inside-out camera system

To recover the 3-D trajectory of a POR (here after, we called as 3-D scan path), we need to estimate a user's head motion ("ego-motion"). In this paper, we present a system for recovering a 3-D scan path and scene structure in 3-D space on the basis of ego-motion obtained from an inside-out camera system.

2. INSIDE-OUT CAMERA SYSTEM

We here describe the configuration of the inside-out camera system and our goal with using the system.

2.1 Inside-out camera system

Our prototype inside-out camera system has the shape of goggles, as shown in Figure 1. It consists of two eye cameras installed at the top of the unit for capturing images of the user's eyeballs and two scene cameras installed at the bottom of the unit for capturing the user's visual field. The equipment measures 160(W) $\times$ 80(H) $\times$ 100(D) mm and weighs about 200 g. As shown in Figure 1, the inside-out camera achieves an optical configuration in which transparent cameras seem to exist. The following describes the eye and scene cameras in more detail.

Eye camera The eye camera system consists of an infrared mirror, two infrared cameras for capturing the left and right eyeballs, and six infrared LEDs arranged around each camera. Each infrared camera captures a near-infrared image of the user's eyeball from in front of that eye via an infrared mirror at a resolution of 640 $\times$ 480 pixels. The LEDs arranged around each camera emit near-infrared light in a wavelength range of 750 - 900 nm. Since infrared light is invisible, it provides no visual stimuli, enabling images of the eyeballs to be captured unhindered.

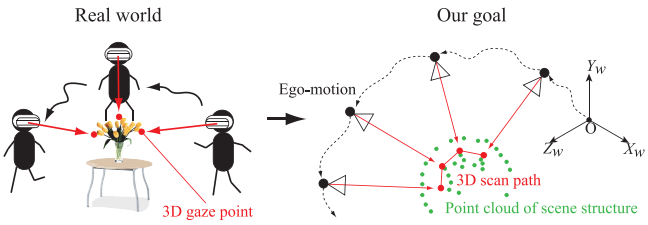


Figure 2: Our goal

Scene camera The scene camera system consists of a half mirror and two compact CCD cameras for capturing the left and right visual fields. The viewing angle of each CCD camera is about 80 degrees, and the focal length is about 4 mm. The half mirror reflects 50% of incident light and allows the rest to pass. Such use of a half mirror makes it possible to capture images with a transparent camera from a position that is optically nearly the same as the user's view point. Furthermore, as this is a stereo camera system, it is relatively easy to calibrate it by using the Tsai model [9] or Zhang model [10] and to estimate the 3-D position of the gaze point in the visual field.

Relationship between eye camera and scene camera The eye cameras and scene cameras are placed opposite each other with half mirrors in between. The image planes configured by each type of camera are therefore parallel to each other. Now, for an object observed by an eye camera that moves in a similar manner to an object observed by the scene camera, it is clear that a correlation exists between the distance moved by the object observed in the eye-camera video and the distance moved by the object observed in the scene-camera image. The relationship between these two types of cameras is therefore easy to work with.

2.2 Our goal

The goal of FPV is to make the system environmentally-aware in order to recognize the key elements of a scene by using the camera, including static objects and the 3-D structure of a scene [1]. Combining these key elements (scene structure) and the gaze point of a user makes it easier to understand a user's intention and behavior. Our goal is to recover the 3-D scan path of a user moving dynamically, and to recover a scene structure in 3-D space by using the inside-out camera system, as shown in Figure 2.

Figure 3 shows the process for recovering a 3-D scan path and scene structure. First, we recover the ego-motion of a user moving dynamically in 3-D space. Second, the 3-D gaze point is estimated from the eye cameras and scene cameras. We also recover the scene structure by identifying image features in common between the two scene cameras and by matching features across images in a video sequence. Finally, by converting to the global coordinate system from the camera coordinate system by using the ego-motion, the system can output a 3-D gaze trajectory ("3-D scan path") with a 3-D scene structure as a point cloud.

3. EGO-MOTION ESTIMATION

In this section, we describe a method for recovering the ego-motion proposed by Bandino et al. [11]. Kalman filters estimate the position and velocity of the world points in 3-D space. The six degrees of freedom of the ego-motion are obtained by minimizing the projection error of the current and previous clouds of static points.

3.1 System model and measurement model

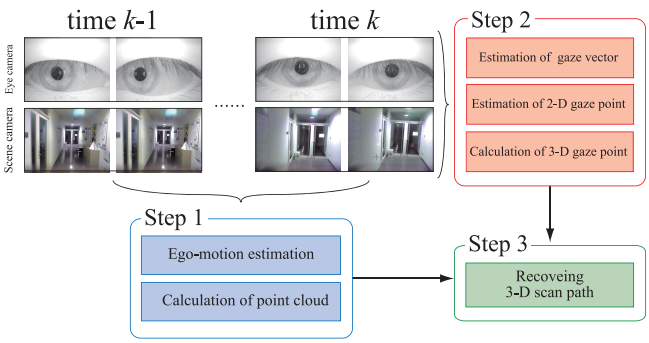


Figure 3: Process for recovering 3-D scan path and scene structure

System Model Let $\boldsymbol{p}_{k-1} = (X, Y, Z)^{\mathrm{T}}$ represent the coordinate vector of a world point observed by the system at time $k-1$ and $\boldsymbol{v}_{k-1} = (\dot{X}, \dot{Y}, \dot{Z})^{\mathrm{T}}$ represent the velocity vector. The camera moves with a given translational and angular velocity. After a time, $\triangle t_k$, the new position of the point from the camera point of view is given by

$$\boldsymbol{p}_k = \boldsymbol{R}_k \boldsymbol{p}_{k-1} + \boldsymbol{t}_k + \triangle t_k \boldsymbol{R}_k \boldsymbol{v}_{k-1},$$

where $\boldsymbol{R}_k$ and $\boldsymbol{t}_k$ are the rotation matrix and translation vector of the static scene with respect to the camera. Combining position and velocity in the state vector $\boldsymbol{x}_k = (X, Y, Z, \dot{X}, \dot{Y}, \dot{Z})^{\mathrm{T}}$ leads to the discrete linear system model equation $\boldsymbol{x}_k$ with the state transition matrix $\boldsymbol{A}_k$ and input vector $\boldsymbol{B}_k$ as :

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{A}_k \boldsymbol{x}_{k-1} + \boldsymbol{B}_k + \boldsymbol{\rho}_k, \\ \boldsymbol{A}_k &= \begin{bmatrix} \boldsymbol{R}_k & \triangle t_k \boldsymbol{R}_k \\ \boldsymbol{0_{3\times 3}} & \boldsymbol{R}_k \end{bmatrix}, \qquad (1) \\ \boldsymbol{B}_k &= (\boldsymbol{t}_k^{\mathrm{T}}, 0, 0, 0)^{\mathrm{T}}, \qquad (2) \end{aligned}$$

where $\boldsymbol{\rho}_k$ is assumed to be Gaussian white noise.

Measurement Model A measurement is defined by the vector $\boldsymbol{m} = (u, v, d)^{\mathrm{T}}$, where (u, v) corresponds to the image position of the feature point and d is its disparity. The (u, v) components are obtained from the KLT tracking algorithm [12], while the disparity d is obtained from the stereo algorithm. The non-linear measurement equation h for the state vector is

$$\boldsymbol{h}(\boldsymbol{x}_k) = \begin{bmatrix} u \\ v \\ d \end{bmatrix} = \frac{f}{Z} \begin{bmatrix} X \\ Y \\ B \end{bmatrix} + \boldsymbol{\nu}, \qquad (3)$$

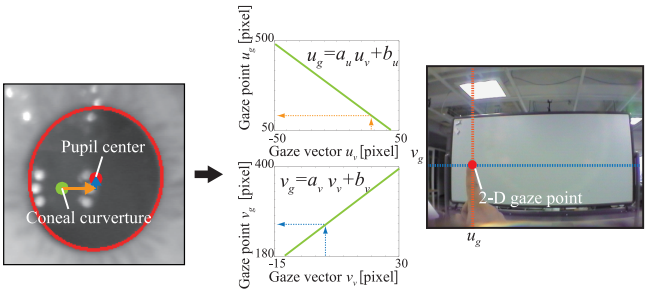
where f is the focal length of the camera and B is the baseline of the stereo system. The term $\boldsymbol{\nu}$ is assumed to be Gaussian white noise. Since the measurement equation is non-linear, the extended Kalman filter[13] is used.

3.2 Ego-motion estimation

Given a set of tracked feature points, $\boldsymbol{m}_i = (u_i, v_i, d_i)^{\mathrm{T}}$, for $i = 1, 2, ..., n$, in the current frame, and the set of corresponding feature points $\boldsymbol{m}'_i = (u'_i, v'_i, d'_i)^{\mathrm{T}}$ in the previous frame, we seek to estimate the rotation matrix $\boldsymbol{R}$ and translation vector $\boldsymbol{t}$. We minimize them in the image space, where the noise level is similar for all components of the measurement vector:

$$E = \arg\min_{\{\boldsymbol{R},\boldsymbol{t}\}} \frac{\sum_{i=1}^{n} \omega_i^2 (\boldsymbol{m}'_i - \boldsymbol{h}(\boldsymbol{R}\boldsymbol{g}(\boldsymbol{m}_i) + \boldsymbol{t}))^2}{\sum_{i=1}^{n} \omega_i^2}, \qquad (4)$$

where ω_i is a weighting factor that determines the contribution of the measurement to the least square solution. To minimize Equa-



(a) Estimate gaze vector

(b) Calculate gaze point coordinates from components

Figure 4: Relation between 2-D gaze point and gaze vector

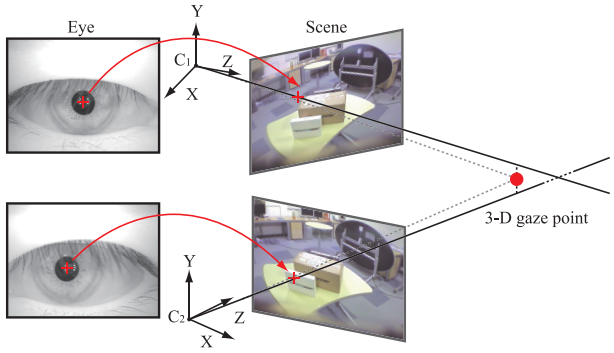


Figure 5: Calculation of 3-D gaze point

tion 4, the rotation matrix $\boldsymbol{R}$ is parameterized by the pseudo-vector $\boldsymbol{r} = (\omega_x, \omega_y, \omega_z)^{\mathrm{T}}$. The matrix $\boldsymbol{R}$ is obtained by rotating the identity matrix $|\boldsymbol{r}|$ radians around the axis $\boldsymbol{r}/|\boldsymbol{r}|$. Assuming $\boldsymbol{t} = (t_x, t_y, t_z)^{\mathrm{T}}$, the parameter for minimization is then the six-dimensional vector $\mathbf{x} = (\omega_x, \omega_y, \omega_z, t_x, t_y, t_z)^{\mathrm{T}}$. This calculation of the rotation matrix and the translation vector is done for each time k.

4. 3-D GAZE SCAN PATH ESTIMATION

In this section, we describe a method for recovering a 3-D scan path on the basis of the ego-motion computed in previous section.

4.1 3-D gaze point estimation

First, we need to estimate the 3-D gaze point by using the inside-out camera system proposed in [8]. The process of estimating the 3-D gaze point in the camera coordinate system has the following flow.

Step 1. Estimation of gaze vector

Our proposed technique estimates the gaze vector from the corneal curvature and pupil center. The inside-out camera that we use features six light sources on the periphery of the camera, and we can assume that the center of these light sources corresponds to the optical axis of the camera. The center of the Purkinje-image group can therefore be taken to be the cornea curvature center. Purkinje images can be extracted from the techniques proposed in [14, 15], and the cornea curvature center $\boldsymbol{c}_k = (u_c, v_c)^{\mathrm{T}}$ can be estimated from the group of Purkinje images determined in this way.

Next, the pupil center is estimated. We use the technique proposed by Sakashita et al. to calculate the pupil center $\boldsymbol{p}_k = (u_p, v_p)^{\mathrm{T}}$ [16]. The gaze vector $\boldsymbol{v}_k = (u_v, v_v)^{\mathrm{T}}$ at time k, which base is taken to be the cornea curvature center $\boldsymbol{c}_k$, can be calculated by the equation $\boldsymbol{v}_k = \boldsymbol{p}_k - \boldsymbol{c}_k$.

Step 2. Calculation of 2-D gaze point

The 2-D gaze point $\boldsymbol{g}_k = (u_g, v_g)^{\mathrm{T}}$ on the image plane of the scene camera can be calculated by using the estimated gaze vector and a conversion equation $\boldsymbol{g}_k = \boldsymbol{a}\boldsymbol{v}_k + \boldsymbol{b}$, where $\boldsymbol{a} = (a_u, a_v)^{\mathrm{T}}$ is the slope and $\boldsymbol{b} = (b_u, b_v)^{\mathrm{T}}$ is the intercept of the linear equation, as shown in the Figure 4(b). Note that the parameters of linear conversion $\boldsymbol{a}$ and $\boldsymbol{b}$ are obtained in advance by using the calibration process.

Step 3. Calculation of 3-D gaze point

Since these 2-D gaze points are points on two scene cameras, the 3-D gaze point can be calculated as a problem in stereo matching, as shown in the Figure 5. We can solve the problem of the 2-D gaze vectors not intersecting in the 3-D space by treating it as a problem in stereo matching. We therefore correct the positions of the two gaze points on the scene cameras so that they coincide by using the optimal correction technique proposed in [17]. Finally, the 3-D gaze point $\boldsymbol{G}_k^c = (X^c, Y^c, Z^c)^{\mathrm{T}}$ in the camera coordinate system is calculated by using stereo matching. The accuracy of its 3-D gaze point is equal to that of stereo matching.

4.2 Recovering 3-D scan path over time

Finally, the 3-D gaze point $\boldsymbol{G}_k^c$ at time k in the camera coordinate system is converted to the global coordinate system by using ego-motion which consists of the rotation matrix $\boldsymbol{R}_k$ and the translation matrix $\boldsymbol{t}_k$ computed in section 3.2, by the following equation;

$$\boldsymbol{G}_k^w = \boldsymbol{R}_k \boldsymbol{G}_k^c + \boldsymbol{t}_k + \boldsymbol{p}_{k-1},$$

where $\boldsymbol{p}_{k-1}$ is the world location of the camera (head) at time $k-1$. The point cloud, which is the 3-D structure of the scene is recovered by identifying image features between the two scene cameras. This point cloud recovered at time k, is also converted to the global coordinate system the same as the 3-D gaze point.

5. EXPERIMENTAL RESULTS

In this section, we describe an example of recovering the 3-D scan path by using our system.

“ Corridor ” data set

A video sequence containing 256 images was captured in a corridor. Figure 6 shows the 3-D scan path with the ego-motion of a user walking straight in the corridor and the scene structure obtained by the accumulation of all observed static points.

“ Table ” data set

A video sequence containing 202 images was captured . Figure 7 shows the 3-D scan path of a user walking around a table. It can be seen that these scan paths and the point cloud are precise enough to understand their spatial relations.

6. CONCLUSION

In this research, we described a system for recovering 3-D scan paths and scene structures in 3D space on basis of ego-motion obtained from an inside-out camera. The 3-D scan paths and scene structures obtained from our system can be used to understand a user's intentions, such as the level of interest in an object. Our future work includes finding applications for the 3-D information of gaze and scene structures by using our inside-out camera system.

7. REFERENCES

[1] T. Kanade and M. Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

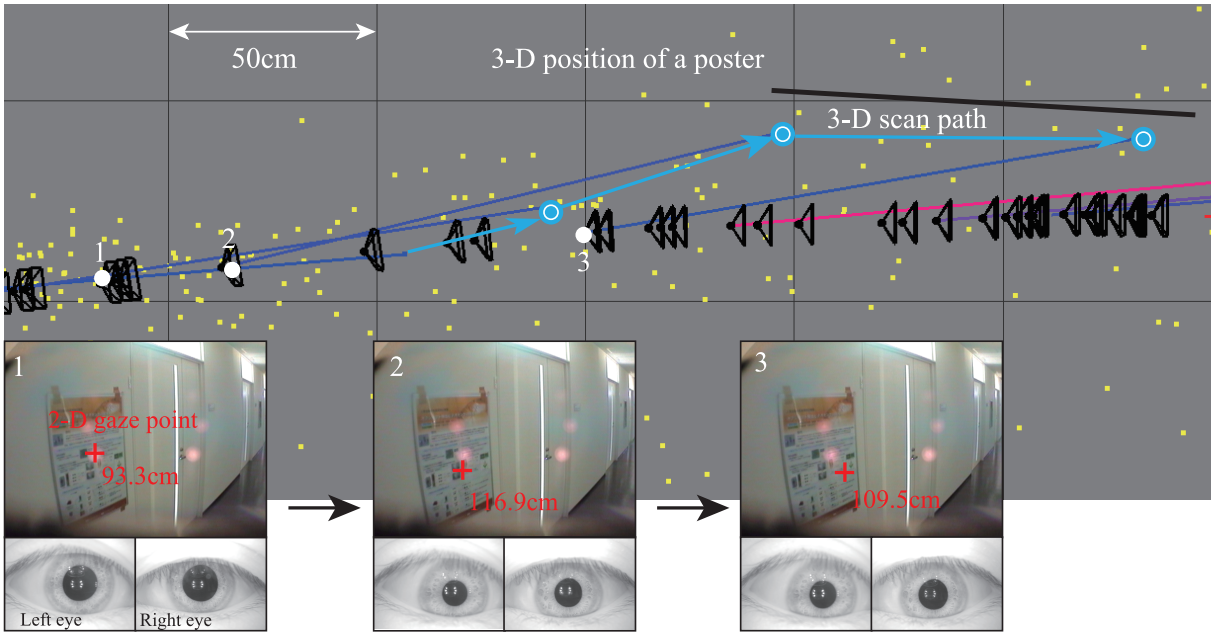


Figure 6: Experimental result 1 ("Corridor" dataset) The 3D points of the scene structure are shown with yellow color. The blue circles show the estimated 3-D gaze points, and the blue arrows show the 3-D scan path obtained by connecting the 3-D gaze points at each frame. We see that the 3-D gaze point is not located on the poster in 3D space, although 2-D gaze point is on the poster.

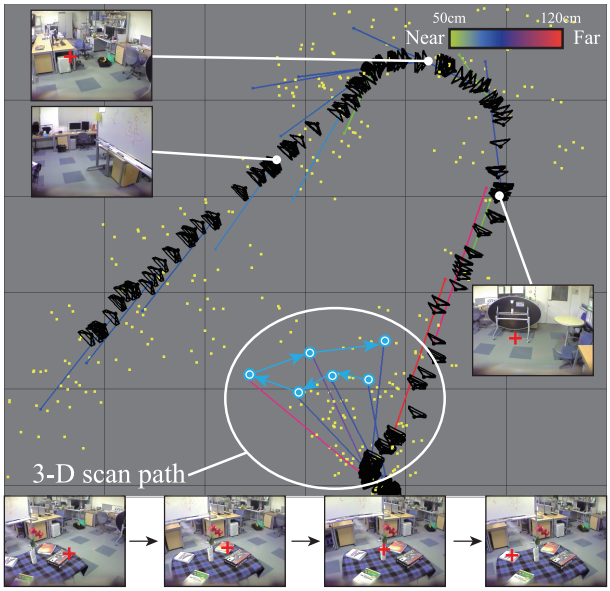


Figure 7: Experimental result 2("Table" dataset)

[2] D. Noton. Eye movements and visual perception. *Scientific American*, 6:34–43, 1971.

[3] T. Ono. What can be learned from eye movement?: Understanding higher cognitive processes from eye movement analysis. *Cognitive Studies*, 9(4):565–579, 2002.

[4] Arrington Research, inc. EyetrackerViewPoint.

[5] ISCAN. AA-ETL-500B.

[6] NAC Image Technology Inc. eyemark recorderEMR-9.

[7] Tobii. Tobii XL.

[8] S. Shimizu and H. Fujiyoshi. Acquisition of 3d gaze information from eyeball movements using inside-out camera. In *Proceedings of the 2nd Augmented Human International Conference*, page 6, 2011.

[9] R.Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, 1922.

[10] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11):1330–1334, 2000.

[11] H. Badino and T. Kanade. A head-wearable short-baseline stereo system for the simultaneous estimation of structure and motion. In *IAPR Conference on Machine Vision Applications (MVA)*, 2011.

[12] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[13] Z. Zhang and O. Faugeras. 3d dynamic scene analysis: a stereo based approach. 1992.

[14] T. Ohno, T. Mukawa, and A. Yoshikawa. Freegaze: a gaze tracking system for everyday gaze interaction. In *the symposium on ETRA 2002: eye tracking research & applications symposium*, pages 125–132, 2002.

[15] S. Tanaka, H. Hikita, T. Kasai, and T. Takeda. Eye-gaze detection based on the position of the cornea curvature center using two near-infrared light sources. *IEICE technical report. ME and bio cybernetics*, 108(479):177–180, 2009.

[16] Y. Sakashita, H. Fujiyoshi, Y. Hirata, and N. Fukaya. Real-time measurement system of cyclodction movement based on fast ellipse detection. *IEEJ Transactions on Electrical and Electronic Engineering*, 127-C:591–598, 2007.

[17] Y. Kanatani, K. Sugaya and H. Niitsuma. Triangulation from two views revisited: Hartley-sturm vs. optimal correction. *Proceedings of the 19th British Machine Vision Conference (BMVC'08)*, 8:173–182, 2008.