IEEE Instrumentation and Measurement
Technology Conference
Budapest, Hungary, May 21–23, 2001

A new Optical Tracking System for Virtual and Augmented Reality Applications

Miguel Ribo
VRVis Competence Center for Virtual Reality and Visualization
Vienna, Austria
email: ribo@vrvis.at

Axel Pinz
Institute of Electrical Measurement and Measurement Signal Processing
Graz University of Technology, Austria
email: pinz@emt.tu-graz.ac.at

Anton L. Fuhrmann
VRVis Competence Center for Virtual Reality and Visualization
Vienna, Austria
email: fuhrmann@vrvis.at

Abstract – ***A new stereo vision tracker setup for virtual and augmented reality applications is presented in this paper. Performance, robustness and accuracy of the system are achieved under real-time constraints. The method is based on blobs extraction, two-dimensional prediction, the epipolar constraint and three-dimensional reconstruction. Experimental results using a stereo rig setup (equipped with IR capabilities) and retroreflective targets are presented to demonstrate the capabilities of our optical tracking system. The system tracks up to 25 independent targets at 30 Hz.***

Keywords – ***Digital Image Analysis, Real Time Tracking, 6 DoF Tracking, Virtual and Augmented Reality***

I. INTRODUCTION

In Virtual Reality (VR) and Augmented Reality (AR), *tracking* denotes the process of tracing the scene coordinates of moving objects in real-time. Typical items to be tracked are head-mounted displays (HMDs) and hand-held 3D interaction devices. In many cases, *position and orientation* of an object have to be recovered, which can be represented by 6 independent variables (3 translational coordinates and 3 rotational angles). Such systems are called *six degrees of freedom (6 DoF) tracking* systems. Since VR and AR applications demand on-line generation of 3D graphics at frame rate, *real-time tracking* means tracking of all objects in the scene at rates of at least 30 Hz. Further requirements of VR / AR tracking systems are related to the specific environment and to what sometimes is called 'immersive' perception of the human user:

- Accuracy: Especially in AR very high accuracy is required, because virtual content has to be *registered* with a real scene. In scene coordinates, position should be within 1mm, and orientation errors should be less than 0.1 degrees.
- Jitter: When no motion occurs, tracker output should be constant.
- Robustness: Small motions should always lead to small changes in tracker output. Outliers have to be suppressed. Tracking operation should be continuous over time (e.g. no loss of targets).
- Mobility: Human users should be unrestricted in their mobility. It would be desirable to have no cables attached to the VR / AR devices, and all mobile parts of a tracking system should be very lightweight. For full mobility, no stationary parts should be required at all.
- Prediction: Since the rendering of a new view requires some time, prediction of the near future is necessary. Otherwise a 'lag' will be perceived, especially during periods of faster motion, where virtual objects appear to be 'behind'. Furthermore, linear prediction may cause an overshooting behavior (virtual objects starting to move too late, then moving too far) in cases of fast accelerations / decelerations.

A further categorization of VR / AR tracking systems is the distinction between *outside-in* and *inside-out* tracking. Outside-in means, that the object (e.g. HMD) is observed from outside, typically by a stationary tracking device mounted to a fixed location in the scene. Inside-out

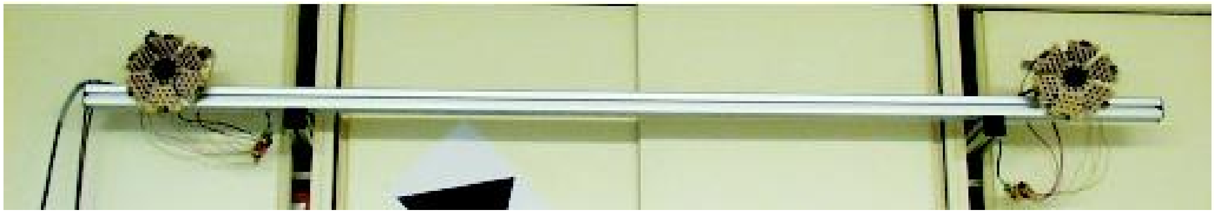


Fig. 1. Stereo rig with adjustable baseline up to 2 m, progressive scan cameras for synchronous stereo grabbing at 30 Hz, IR lighting, and IR filters.

systems take the necessary measurements from a mobile platform, typically from the object itself. In the case of optical tracking, an inside-out tracking system might use a camera which is attached to the HMD and calculate its position and pose with 6 DoF from the tracking of stationary visual landmarks attached in the scene.

In this paper, we describe a new stereo vision based tracking device, which allows tracking of up to 25 3D target positions at a rate of 30 Hz. In a typical scenario the scene is captured at discrete instances in time, blobs are extracted and a linear predictor is used to speed up the search process, the system computes then 3D positions from 2D image points using epipolar geometry and workspace constraints. A detailed description of the tracking system is given in section II. Afterwards, a concrete application reported in section III demonstrates the capabilities and the potential benefits of our tracking system when dealing with VR / AR environments.

II. TRACKING SYSTEM

None of the tracking systems existing today fulfills all of the above requirements (see e.g. [1] for the description of a commercial system, and [2] for a survey of tracking technologies for virtual environments). Significant improvements have been achieved by using *hybrid tracking systems* which combine the strengths and eliminating the disadvantages of complementary sensing systems (e.g. optical and inertial tracking [3], optical and magnetic tracking [4]).

Tracking constitutes a major current field of research in VR / AR, and many specific approaches to tracking have been reported in recent literature (e.g. [5], [6], [7]). Commercial systems for tracking in VR / AR have been around for several years (e.g. Intersense: www.isense.com, Polhemus: www.polhemus.com) and are constantly improving, but they are very expensive and the mobility of the users is still limited. Optical tracking technologies similar to our approach (but without real-time constraint) are used, for instance, to record 3D motion sequences of actors / persons for animation purposes in film industry or for medical diagnosis in gait analysis (e.g. Vicon: www.vicon.com).

In this paper we propose an optical tracking system for VR / AR applications. The system is based on the *outside-in* approach using a fully calibrated stereo setup with infrared (IR) pass filters. The tracked markers are retroreflective spheres illuminated using IR LED arrays mounted in a radial-symmetric setup. Thus, finding the marker in a 2D search image is accomplished using a simple blob detection approach. Both search-frames (i.e. the left and right images) in our system are quantized into a number of discrete tiles which build the basis for the tracking algorithm. The area being processed in each consecutive frame is predicted using the information extracted from the previous frames. Thus, we reduce unnecessary processing of large areas of the search-frame.

The work most similar to the approach presented here is reported by Dorfmüller ([8] for outside-in and [9] for inside-out tracking). However, major differences can be enumerated as follows.

A. Hardware

The hardware setup consists of a calibrated stereo rig mounted to a fixed frame with adjustable baseline of up to 2 m (see Fig. 1). Its cameras (progressive scan CCD at 30 Hz frame rate) are equipped with optics adjustable to wide viewing angles (up to 50 degrees), infrared pass filters and special infrared lighting devices (see Fig. 2a). Spherical retroreflective targets are attached to the HMD (see Fig. 2b) as well as to the hand-held 3D interaction devices in order to allow for an easier detection of the tracking targets in the pair of the 2D stereo images. The use of versatile 3D interaction devices, in our case the Personal Interaction Panel (PIP) - a tracked and augmented pen and pad combination (see [10], [11] for details) - permits interaction within the Studierstube Augmented Reality environment [12].

B. Methods

Calibration Process. An accurate and fully calibrated stereo rig is mandatory if we want to provide our system with three-dimensional vision, and improve tracking accuracy significantly. Indeed, to perform three-dimensional Euclidean reconstruction from stereo pair images, the *intrinsic parameters* (i.e. internal camera ge-

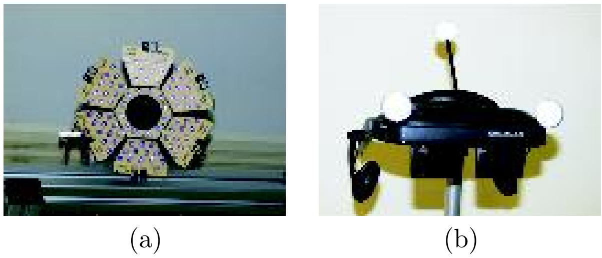


Fig. 2. (a) Closeup view of a camera with mounted IR lighting and IR filter. (b) See through head-mounted display with retroreflective markers.

ometry and optical characteristics) and the *extrinsic parameters* (i.e. three-dimensional position and orientation of the camera frames) need to be known [13]. While Dorfmüller uses a flexible calibration method first introduced by the MIT Media Laboratory [14], we rely on 'classical' photogrammetric calibration approaches determining lens distortion, interior and exterior orientation. The technique used for the complete calibration of the stereo rig is based on Zhang's approaches described in [15], [16].

Blob Detection. Based on the illumination hardware and markers we use, the resultant input image shows a dark background (no IR reflectivity) and a bright spot for every marker visible to the camera (see Fig. 3). Thus, simple thresholding is sufficient to provide a good segmentation of the input data. However, due to possible reflections from other objects in the scene (e.g. clothes, wrist watches, ...) an additional criterion to judge on the origin of the bright spot is needed. As the retroreflective spheres result in nearly round blobs using a measure of roundness [17] is suitable for this purpose. In the current application we constrain the area of blobs to be within a certain range deduced from the working volume and the size of the retroreflective spheres. Furthermore we constrain the roundness to allow for round or elliptical objects only.

Prediction. During the detection of blobs in a sequence of input images some knowledge about the movement of the markers can be obtained. This knowledge can be incorporated into the tracking algorithm using a prediction module. In the 2D blob tracking algorithm we use linear prediction in combination with a dynamic search frame size to allow for adaptation to different motions of the targets. Linear prediction is a straight forward and easy to apply prediction scheme which has shown to provide sufficient performance in combination with the dynamic adaptation of the frame size. However, other prediction schemes (e.g. Kalman filters) could be used at this point of the algorithm.

Epipolar Constraint. Once the intrinsic and extrinsic parameters of our stereo setup are computed, we can easily deduce the *epipolar constraint* by letting [13]

$$\mathbf{m}_r^t \mathbf{F} \mathbf{m}_l = 0, \tag{1}$$

where $\mathbf{m}_l$ and $\mathbf{m}_r$ are two corresponding points from the left and right images, respectively. The matrix $\mathbf{F}$, referred in the literature as the fundamental matrix, is the algebraic representation of the *epipolar geometry.*

In other words, a point in the left image (i.e. $\mathbf{m}_l$) generates, by means of the fundamental matrix, a line in the right image (i.e. $\mathbf{F}\mathbf{m}_l$) on which its corresponding point must lie. The search for correspondences is thus reduced from a region to a line. However under some circumstances, it could be possible that several points lie on the same epipolar line. In those cases, to avoid mismatches between stereo pair images, the order of points on the epipolar line is used to improve the matching process. In Fig. 3 the epipolar lines of the corresponding points from the left and right images are drawn to illustrate the matching technique used so far. We remark that the point without label in the left image is still tracked even if no correspondence was found in the right image. In this case, the system takes into account that either this point is an outlier or an occluded marker not detected by the right camera.

3D Reconstruction. By means of the intrinsic and the extrinsic parameters of the stereo rig, we can compute the 3×4 projective matrices $\mathbf{P}_l$ and $\mathbf{P}_r$ of the left and the right cameras [13]. Afterwards, for a given corresponding pair of points $\mathbf{m}_l = (u_l, v_l)^t$ and $\mathbf{m}_r = (u_r, v_r)^t$, we can recover the 3D position $\mathbf{M}_{C_l}$ (i.e. expressed with respect to the left camera coordinate system) of a target in the scene, by using the reconstruction method described in [18]. Basically we have

$$\begin{bmatrix} \mathbf{m}_l \\ 1 \end{bmatrix} = \mathbf{P}_l \begin{bmatrix} \mathbf{M}_{C_l} \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{m}_r \\ 1 \end{bmatrix} = \mathbf{P}_r \begin{bmatrix} \mathbf{M}_{C_l} \\ 1 \end{bmatrix}, \tag{2}$$

which may be written in the form (for a suitable 4×4 matrix $\mathbf{B}$)

$$\mathbf{B}\mathbf{M}_{C_l} = 0 \quad \text{with} \quad \mathbf{B} = \begin{bmatrix} \mathbf{p}_l^1 - u_l \mathbf{p}_l^3 \\ \mathbf{p}_l^2 - v_l \mathbf{p}_l^3 \\ \mathbf{p}_r^1 - u_r \mathbf{p}_r^3 \\ \mathbf{p}_r^2 - v_r \mathbf{p}_r^3 \end{bmatrix}, \tag{3}$$

where $\mathbf{p}_l^i$ and $\mathbf{p}_r^i$ are the i-th *row vectors* of the matrices $\mathbf{P}_l$ and $\mathbf{P}_r$, respectively. By setting the constraint $\|\mathbf{M}_{C_l}\| = 1$, the solution of (3) is simply the eigenvector of the matrix $\mathbf{B}^t\mathbf{B}$ associated to the smallest eigenvalue. We remark that this stage can be done by using the singular value decomposition method.

Once the 3D positions of all tracked targets are computed, we build some workspace constraints (e.g. distances and

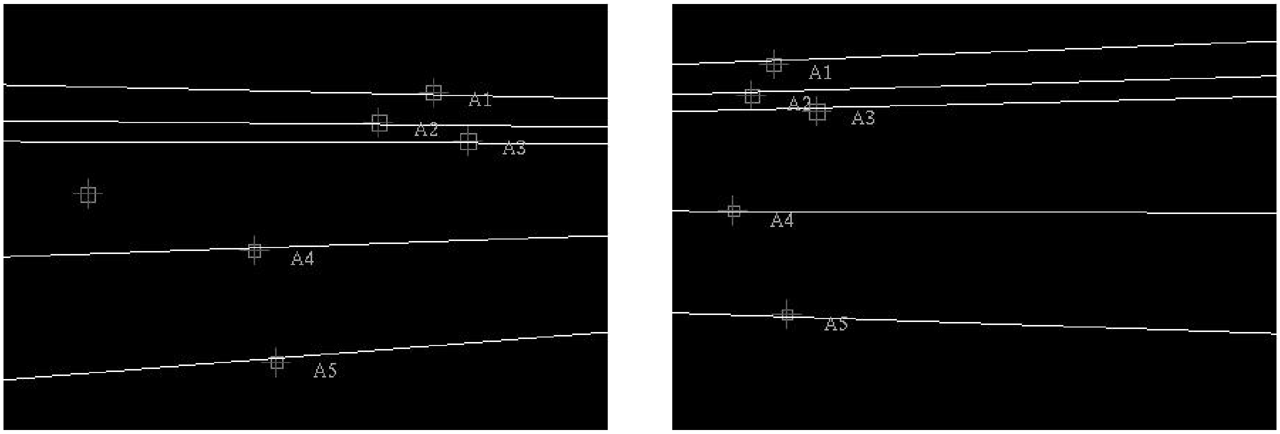


Fig. 3. Pair of stereo images during tracking process. Several retroreflective markers are tracked by the system. The crosses indicate the current blob position, while the frames quantify their size in the image. The labels A_i denote the corresponding points which minimize the epipolar constraint (white lines).

angles between points) in order to gather them into constellations. In this way, the system is able to identify which set of targets belongs to the appropriate 3D device.

III. EXPERIMENTAL RESULTS

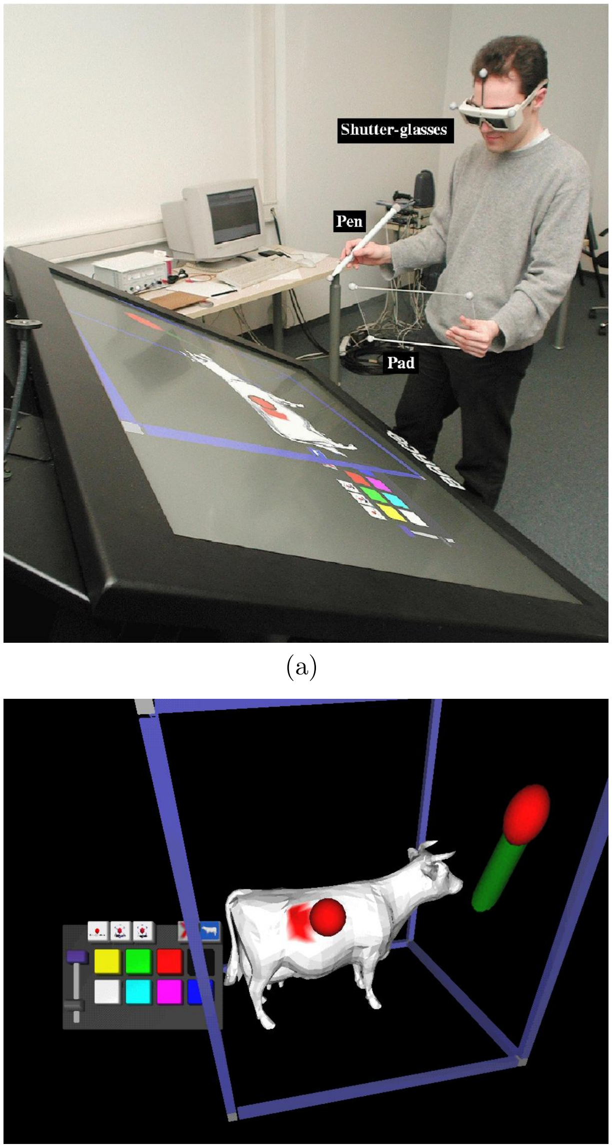
In our system, all objects of interest are marked using retroreflective spherical targets. The scene is illuminated by IR LED lighting and imaged using a calibrated stereo rig (see Fig. 1). Any object marked with one target (a point in 3D space) can thus be tracked with 3 DoF (object position). Two markers (a line in space) yield 5 DoF (position and orientation, just rotations around the line cannot be observed), and a minimum of three markers (a plane in space) allows to fully recover 6 DoF (position and orientation of any rigid object in 3D space). More than three markers can be used to increase robustness (with respect to occlusion and outliers) and accuracy.

Within the spatial volume defined by the stereo setup and by the illumination, the system fulfills all requirements enumerated in section I. It provides real-time simultaneous tracking with 3 to 6 DoF per object (depending on the number of targets attached to it), accuracy and jitter are satisfactory, tracking is robust unless visual (self-) occlusion of targets occurs, and full mobility of the user within the tracking volume - as long as the user faces the display, which is required for normal operation anyway - is achieved at minimal extra effort (no cables, no electronic devices, no extra weight, just a few retroreflective targets). We demonstrate the capabilities of our tracking system in Studierstube on the Virtual Table [11] with one user wearing tracked 6DOF shutter-glasses and interacting via 5DOF pen and 6DOF pad (see Fig. 4a). The size of the table is roughly 1.6m (horizontal) × 1.2m (vertical / depth, depending on the tilt angle of the table). Fig. 4a shows a typical situation: the shutter-glasses and the Pad are tracked with 6 DoF, while the pen is tracked with 5 DoF. Fig. 4b shows the 3D models of those devices as viewed through a graphical interface using the information delivered by the optical tracking system. After simultaneous grabbing of stereo pair images at 30 Hz, our tracking software performs the following steps: extraction of bright blobs, establishment of stereo correspondences, measurement of 3D positions of all markers, determination of the transformation from stereo rig to virtual table coordinates, calculation up to 6 DoF for each of the 3 interaction devices (shutter-glasses, Pen, Pad). Search for bright blobs in small windows of the next stereo images.

IV. CONCLUSION

A real-time optical tracking system based on a calibrated stereo setup was presented in this paper which fulfills all requirements enumerated in section I. Up to 25 retroreflective markers can be tracked in the scene at 30 Hz. To our knowledge there is currently no other VR / AR tracking system of similar performance. We achieve very good spatial accuracy in all 6 DoF due to an accurate calibration of the system. A balanced hard- and software concept (IR lighting, synchronous stereo grabbing at 30 Hz with progressive scan CCD cameras, retroreflective spherical targets, and efficient blob tracking software) leads to a sufficient temporal resolution to meet the hard real-time constraint of VR / AR applications.

For the VR / AR experiment of one user working in front of a virtual table (see section III), where magnetic tracking systems are not well suited due to strong distortions of the electromagnetic field, our tracking setup proves its superiority and gives the user unconstrained mobility (neither cables nor other communication channels are required). Moreover, a set of novel 3D interaction devices



(a)

(b)

Fig. 4. (a) Person in interaction with a virtual table: shutter-glasses (tracked with 6 DoF), transparent Pad (6 DoF), and Pen (5 DoF). (b) 3D graphical model of the working scene.

consisting of shutter-glasses (tracked with 6 DoF), transparent Pad (6 DoF), and a pen (5 DoF) has been devised. They are lightweight, easy to use, and enhance the interaction capabilities of the user with respect to VR / AR applications.

In the future, we plan to mark the four corners of the virtual table to allow the online determination of the coordinate transform between the stereo rig and the virtual table. In this way, the tracking system can continue smoothly even after changes in the overall system setup (e.g. moving the stereo rig, changing the tilt angle of the table). In addition, the proper concept of our blobs extraction / prediction algorithm (i.e. tracking within dynamic frames) gives us the possibility to utilize dedicated hardware as CMOS cameras. This will improve the capabilities of our tracking system by increasing significantly the working fame rate.

ACKNOWLEDGMENT

We would like to thank Markus Brandner for his help on the conception of the optical tracking algorithm, Harald Ganster for discussions and remarks about the pose estimation of the 3D devices in the scene, and Rainer Splechtna and Jan Prikryl for support concerning the VR / AR application with the virtual table. This work was supported by VRVis competence center for Virtual Reality and Visualization.

References

[1] E. Foxlin, M. Harrington, and G. Pfeiffer, "Constellation^{TM}: A wide-range wireless motion-tracking system for augmented reality and virtual set applications," in *Proc. SIGGRAPH*, 1998, pp. 371–378.

[2] J.P. Rolland, L.D. Davis, and Y. Baillot, "A survey of tracking technology for virtual environments," in *Augmented Reality and Wearable Computers*. Ed. Bardfield and Caudell, 2000.

[3] S. You, U. Neumann, and R. Azuma, "Orientation tracking for outdoor augemented reality registration," *IEEE Computer Graphics and Applications*, pp. 36–41, 1999.

[4] T. Auer and A. Pinz, "The integration of optical and magnetic tracking for multi-user augmented reality," *Computers & Graphics*, vol. 23, no. 6, pp. 805–808, 1999.

[5] R. Azuma, J.W. Lee, B. Jinag, J. Park, S. You, and U. Neumann, "Tracking in unprepared enviroments for augmented realitiy systems," *Computers & Graphics*, vol. 23, pp. 787–793, 1999.

[6] H. Kato and M. Billinghurst, "Marker tracking and hmd calibration for a video-based augemented reality conferencing system," in *Proc. IWAR*, 1999.

[7] G. Welch, G. Bishop, L. Vicci, S. Brumback, K. Keller, and D. Colucci, "The hiball tracker: High-performance wide-area tracking for virtual and augmented environments," in *Proc. VRST*. 1999, pp. 1–10, ACM.

[8] K. Dorfmüller and H. Wirth, "Real-time hand and head tracking for virtual environments using infrared beacons," in *Proceedings CAPTECH'98*. 1998, vol. 1537 of *LNCS*, pp. 113–127, Springer.

[9] K. Dorfmüller, "Robust tracking for augmented reality using retroreflective markers," *Computers & Graphics*, vol. 23, no. 6, pp. 795–800, 1999.

[10] Z. Szalavári and M. Gervautz, "The personal interaction panel - A two-handed interface for augmented reality," *Computer Graphics Forum*, vol. 16, no. 3, pp. 335–346, 1997.

[11] D. Schmalstieg, L.M. Encarnaçao, and Z. Szalavári, "Using transparent props for interaction with the virtual table," in *Proceedings SIGGRAPH Symposium on Interactive 3D Graphics '99*, 1999, pp. 147–154.

[12] D. Schmalstieg, A. Fuhrmann, G. Hesina, Z. Szalavari, and W. Purgathofer L. M. Encarnao, M. Gervautz, "The studierstube augmented reality project," Tech. Rep. TR-186-2-00-22, Vienna University of Technology, December 2000, Submitted for publication.

[13] O. Faugeras, *Three-Dimensional Computer Vision: a Geometric Viewpoint*, MIT Press, 1993.

[14] A. Azarbayejani and A. Pentland, "Camera self-calibration from one point correspondence," Tech. Rep. 341, MIT Media Lab, 1995.

[15] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientation," in *Proceedings IEEE International Conference on Computer Vision*, Corfu, Greece, September 1999, pp. 666–673.

[16] Z. Zhang, "Motion and structure from two perspective views: From essential parameters to Euclidean motion via fundamental matrix.," *Journal of the Optical Society of America*, vol. 14, no. 11, pp. 2938–2950, 1997.

[17] J. C. Russ, *The Image Processing Handbook*, CRC Press, 2^{nd} edition, 1995.

[18] C. Rothwell, O. Faugeras, and G. Csurka, "A comparison of projective reconstruction methods for pairs of views," *Computer Vision and Image Understanding*, vol. 68, no. 1, pp. 37–58, October 1997.