MVA manuscript No.
(will be inserted by the editor)

The Agile Stereo Pair for Active Vision

Eric Samson[1], Denis Laurendeau[1], Marc Parizeau[1], Sylvain Comtois[1], Jean-François Allan[2], Clément Gosselin[2]

[1] Computer Vision and Systems Laboratory, Dept. of Electrical and Computer Engineering
[2] Robotics Laboratory, Dept. of Mechanical Engineering
Laval University, Québec, QC, Canada, G1K 7P4

Received: date / Revised version: date

Abstract This paper presents a new stereo sensor for active vision. Its cameras are mounted on two independent 2-DOF manipulators, themselves mounted on two translation stages. The system is designed for fast and accurate dynamical adjustments of gaze, vergence, and baseline. A complete description of its software and hardware components is given, including a detailed discussion of its calibration procedure. The performance of the sensor with respect to dynamical properties and measurement accuracy is also demonstrated through both simulations and experiments.

1 Introduction

Stereopsis is the task of combining the information contained in pictures of a scene obtained from two (or more) different viewpoints. Combining image information allows depth to be recovered by detecting the same features of a scene in both images and using projective geometry. Depth information, also called range data, is useful in numerous tasks such as scene modelling, object recognition, photogrammetry, collision avoidance in mobile robotics, etc. Stereopsis and, more generally stereovision, has been the object of intense research efforts in the fields of perception, psychology, computer vision and robotics. First because it is the approach that has been chosen by nature for depth measurement in many species of animals; stereovision is thus a challenging research field *per se*. Secondly, on a more practical side, stereovision does not require an external light source, such as a laser. It is thus well adapted to military applications or applications for which the use of a laser source could be harmful to humans or other living creatures. However, stereovision systems face a serious problem known as the matching problem: even though matching is performed so effortlessly by humans, it is a difficult and time consuming task for artificial vision systems to match scene features in a pair of images. Despite this fact, artificial stereovision systems have been widely used as visual input devices for autonomous vehicle guidance [BFZ93, DL01, LJM01, MLG00, Mor96].

In the early stages of computer vision research, some researchers realized that, for many tasks, the vision process could be greatly simplified or improved with the use of versatile sensors [Baj88, AWB88, KFS88, Bal91]. A versatile sensor is one that allows for reconfiguring itself dynamically in order to actively explore a scene, react to a change in the environment or track a moving target in real time. Such tasks are commonly referred to as *Active Vision*.

Even though it is not a mandatory requirement, a versatile system should also be biologically inspired and allow eye motions similar to those found in biological vision systems. This is because living creatures are particularly well adapted to the complex real-world environment. As a matter of fact, most authors attempt to design active vision systems that match as much as possible the characteristics of the human vision system. The most desirable characteristics are (i) flexibility, allowing the system to explore its surrounding environment in the most efficient way, (ii) good dynamical performances, for fast reaction to environment changes and high speed tracking, (iii) compactness, for easy integration in space-limited systems such as mobile robots, and (iv) high accuracy, to gather 3D information of the environment by stereopsis.

In this context, the KTH robotic head developed by Pahlavan *et al.* [PE92] is one of the first active vision systems providing a reasonable amount of performance and accuracy for some active vision tasks (references to earlier actives vision systems can be found in [PE92, MJT93, WFRL93]). Biologically inspired, KTH is aimed at replicating the performances of the human visual system in a compact design. With 13 DOF, it is flexible as well. Each of the two cameras can be oriented independently along 2 DOF (pan/tilt). The distance between them, called the baseline, can also be adjusted. The whole system is mounted on a 2 DOF neck performing pan/tilt move-

ments. On both cameras, the zoom, focus and aperture are adjustable. KTH has been tested with tracking algorithms involving neck/eyes coordination. With neck and eyes axes running at the same time, the cameras can reach 180°/s, which is far from the performances of the human eyes (capable of saccadic movements up to 900°/s [KSJ91]).

TRISH developed by Tsotsos *et al.* is another robotic head of interest [MJT93]. As the primary sensor of PLAYBOT, a robotic system providing help to disabled children, TRISH aimed at mimicking the human visual system capabilities and producing valuable 3D data by stereopsis [TVD+98]. In order to gain on accuracy, the number of DOF has been limited to 7. Each camera has independent control on tilt (±45°, 54°/s), pan (±80°, 54°/s), and torsion (also called swing). The head supporting the eyes allows for panning of the stereo pair. Torsion movement on each camera has been introduced to reduce the computation time of the stereo matching task. Nevertheless, Tsotsos *et al.* concluded that the mechanical accuracy needed to obtain valuable 3D data cannot be obtained at reasonable cost.

Despite the conclusion of Tsotsos *et al.*, many research groups pursued the accuracy objective in their design of a high performance flexible and compact active sensor. A good example is the Yorick series. They have 4 DOF : common pan and tilt axes and two independent vergence axes. The baseline varies from 11 cm to 55 cm, depending on the model. The most compact one can perform panning (±118°), tilt (±45°) and vergence (±14°) at 425°/s, 680°/s and 560°/s respectively [SMMB98]. Reducing the number of DOF leads to more compact design, higher speed and greater accuracy while being flexible enough for most active vision applications.

Another example of an active head that benefits from having a small number of DOF is TRICLOPS [WFRL93]. It has 4 DOF configured in the same way as those of the Yorick heads: common pan and tilt axes and two independent vergence axes. However, all of its DOF are directly driven. Direct drive systems avoid the use of gearbox and generally lead to greater accelerations and better accuracy. Accordingly, TRICLOPS have dynamical characteristics comparable to those of the human visual system. Pan (±96°), tilt (28° up; 65° down) and vergence (±44°) axes can reach 660°/s, 1000°/s and 1830°/s respectively. A special feature of TRICLOP is a fixed camera situated in-between its two mobiles cameras. With a larger field of view, it is intended to provide peripheral vision to the system.

The most recently reported active vision systems are CeDAR and BMC heads. The CeDAR head [TARZ00, DZ04] has 3 DOF: a common tilt axis and two independent vergence axes. As compared to the other active heads, CeDAR has the particularity that all of its DOF are cable driven by motors located on its base. As a result, the tilt motor does not have to carry the load of the motors of the vergence axes and the backlash problems related to the use of a gearbox transmission are avoided. This design allows for good dynamical characteristics: 600°/s on a 90° range for tilt axis and 800°/s on a 90° range for vergence axes.

Finally, the active stereo head of the Bio-Mimetic Control Research Center [NFM+04] is performance oriented. This heavy-weight 6 DOF (two pan/tilt cameras mounted on a pan/tilt neck) can perform high speed camera movements. The cameras can reach 30000°/s on both pan and tilt axes within a range of 60° while the neck axes can be driven at 600°/s within a range of 180° for the pan axis and 60° for the tilt axis. This active stereo head has been built to test tracking algorithms based on a fast vision control loop running at 1kHz.

This paper presents a new versatile stereo system, called the Agile Stereo Pair (ASP), for active vision applications. The ASP, which implements motion similar to the human eye, is characterized by its mechanical design based on two compact 2-DOF parallel orientation mechanisms. It has the advantage of being both compact and lightweight, while offering high dynamic performances and good accuracy. Moreover, it can be reproduced at low cost since it is composed of few mechanical parts, and high performances are achieved with low cost DC motors. The two orientation mechanisms are independent so they can be integrated in a system in any convenient way. In this paper, we present the ASP with both eyes mounted on accurate linear translation stages. This configuration allows for real-time baseline adjustment of the stereo head.

In addition to the detailed description of the ASP, the main contribution of this paper is to provide a calibration procedure for the new sensor along with experimental results on 3D measurements. Apart from trying to replicate the binocular configuration of the human vision system, the second camera of stereo systems has no other practical purpose than to compute depth measurements. It is thus worthwhile to note that very few 3D results are given in the literature for active vision systems. Shih et al. [SHL98] do provide a complete calibration procedure, but results are expressed in terms of the epipolar constraint. No metric measurements are given. In Crowley et al. [CBM93], a technique is described where extrinsic parameters are kept calibrated by tracking a group of points in the scene, but these points must remain in the field of view at all times. Some 3D reconstruction errors are given, but they do not give information about the sensor accuracy over its full operating range without the presence of these calibration points. In Beymer et al. [BF03], a stereo pair is used to position the gaze of a person on a monitor. Results which only provide final gaze positions in pixels do not give any indication on the 3D accuracy of the sensor. Others limit their results to mechanical angular precision. This lack of clear and comparable metric 3D measurements can be explained in two ways. First, many active systems are based on mechanisms that are simply too inaccurate to permit 3D

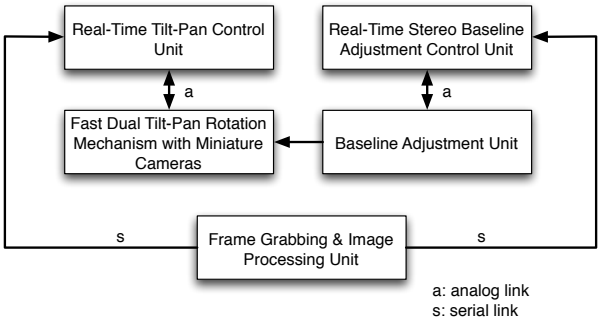


Fig. 1 Overview of the Agile Stereo Pair System.

computations. Second, existing metric calibration methods are not well adapted for active systems. In this paper we provide both a unified calibration procedure, and 3D metric results for the ASP.

The rest of the paper is structured as follows. Section 2 presents an overview of the complete system. Each system component is then described in the following sections. Section 3 presents the geometry of the ASP, while Section 4 describes the cameras that are used in the ASP. The procedure that was designed and implemented for calibrating the stereo pair is presented in Section 5. The dynamic performances of the ASP along with its measurement accuracy are presented in Section 6. Finally, Section 7 concludes with some discussion on future work and applications.

2 Overview of the Agile Stereo Pair

Figure 1 shows a diagram of the main building blocks of the Agile Stereo Pair system. The main component of the system is the Fast Dual Tilt-Pan Rotation Mechanism that is used for orienting a pair of miniature cameras. This mechanism is controlled by the Real-Time Tilt-Pan Control Unit which receives its commands through a serial link from the Frame Grabbing and Image Processing Unit. The Baseline Adjustment Unit allows real-time baseline adjustment. The Real-Time Stereo Baseline Adjustment Control Unit accepts commands from the Frame Grabbing and Image Processing Unit and sends the required signals to the Baseline Adjustment Unit. In addition to being the master component of the system, the Frame Grabbing and Image Processing Unit is used for calibrating the system and for computing the 3D coordinates from stereo images acquired by the miniature cameras. Except for the Tilt-Pan Rotation Mechanism, the system is composed of off-the-shelf components.

3 Geometry of the Agile Stereo Pair (ASP)

Figure 2a shows a schematic representation of the 2-DOF mechanism for the right eye of the Agile Stereo Pair (one half of the Fast Dual Tilt-Pan Rotation mechanism of Figure 1). It is based on a previous 3-DOF design (tilt, pan, torsion) developed by Gosselin *et al.* [GSP97]. The 2-DOF design consists of a 5R closed-loop spherical mechanism in which the axes of all 5 revolute joints intersect at one common point [GC99, Car97]. Moreover, the angle between any two neighbouring joints is equal to 90°. The mechanism has 2 DOF, thereby providing the ability to point the axis of the camera in any direction, within the physical system limits. Therefore, the camera undergoes pure rotations with respect to its principal point, located at the intersection of the 5 revolute joints of the mechanism. The angular span is ±40° in azimuth (pan) and ±40° in elevation (tilt)[1] with an angular resolution of 0.18° for each axis. Each axis is directly driven by a 24 volt DC motor. The angular resolution is limited only by the position encoders mounted on the actual setup since no gearbox is used between the axes and the motors. The mechanism for the left eye is a mirror image of the right eye allowing perfect enantiomorphism of the stereo pair.

Figure 2b shows the CAD model of the actual mechanical assembly, and Figure 2c shows an exploded view of the mechanism with all of its parts except the motors. Figure 2d shows the complete system assembly for the right eye (approximate dimensions of the bounding box are height = 20 cm, width = 20.5 cm, depth = 8 cm), and Figure 2e shows the complete stereo pair where each 2-DOF mechanism is mounted on an optional linear translation stage (Newport MTM250OCC1 stage) with 25 cm range (the Baseline Adjustment Unit in Figure 1). Mounting the left and right eye translation stages side by side allows a dynamic adjustment of the baseline in the 5 to 55 cm range.

An important feature of the 2-DOF mechanisms of the ASP is their parallel design. Parallel mechanisms are characterized by the fact that the end-effector (the camera holder in the present case) is connected to the base via one or more closed kinematic chain(s) and that all the actuators can be located on the base. As opposed to serial mechanisms, for which the actuator at one of the DOF joints must move the actuators of the following DOF up to the end effector, parallel mechanisms lead to very good dynamic properties since the inertia of the moving parts is considerably reduced. Moreover, closed kinematic chains lead to high stiffness mechanisms and great precision of movements. This is explained by the fact that the mechanism is constrained on two points for each kinematic chain. This results in an assembly that cannot deviate from its intended configurations, within the machining tolerances of the parts. Also, any misalignment in the assembly would result in a complete deadlock of the mechanism. For the particular case of

[1] The actual full angular span is ±45° but it is limited to ±40° by the control software in order to avoid the occurrence of overshoots that could damage the mechanisms.

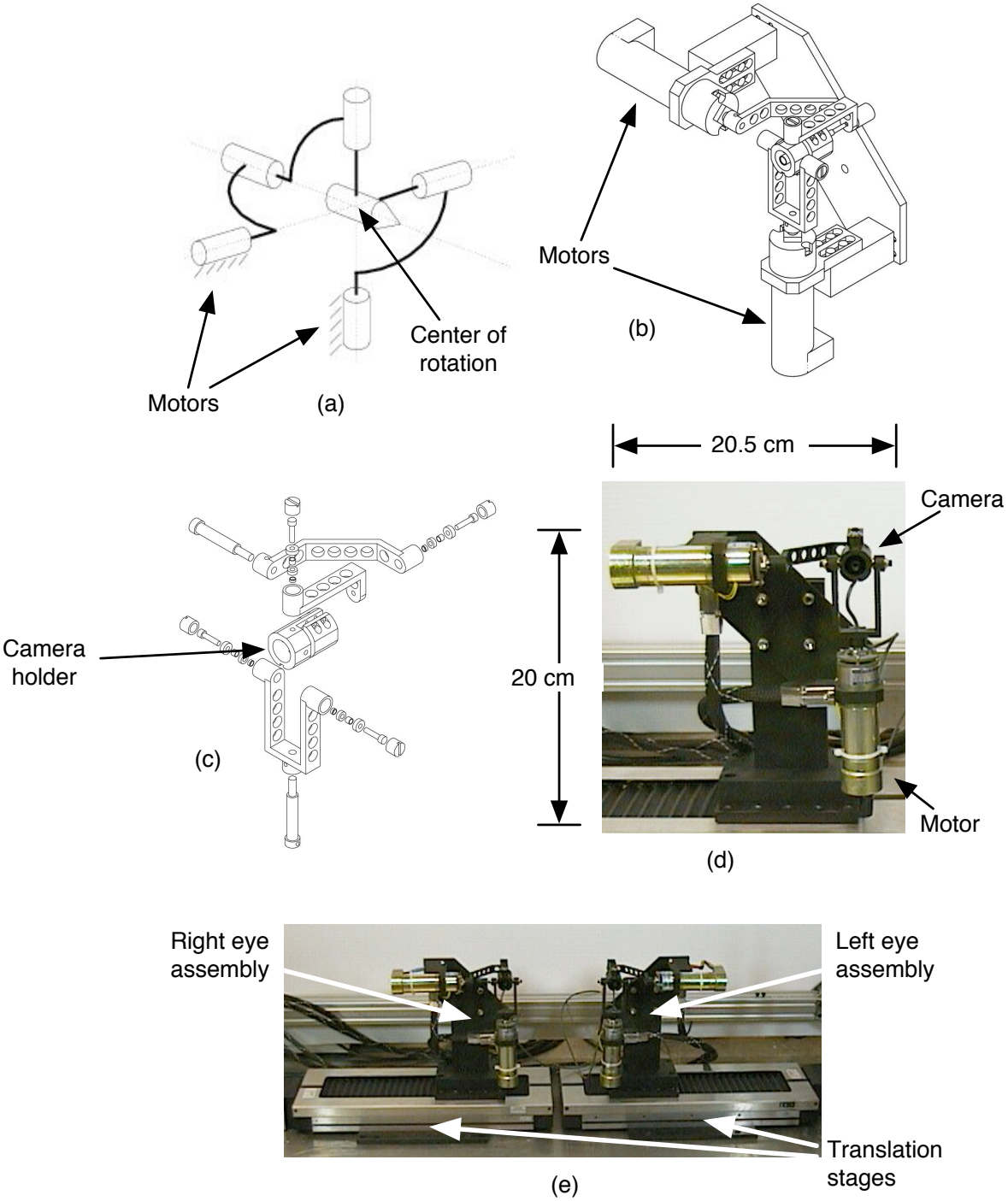


Fig. 2 Agile Stereo Pair: (a) Schematic model of the 2-DOF mechanism for the right eye of the ASP; (b) CAD model of the mechanism; (c) exploded view of the components; (d) photograph of the complete system for the right eye; (e) photograph of the stereo pair with baseline ajustment.

the 2-DOF mechanisms of the ASP and as mentioned above, all rotation axes of the joints on the unique closed kinematic chain are constrained to meet at a common point (the rotation center). From the properties of over-constrained closed kinematic chains, it can be assumed that pure rotation movements are performed whenever the system is able to move freely.

4 Sensing Devices of the ASP

The camera holder (see Figure 2c) accepts Toshiba SM-43H lipstick color video CCD-cameras (768 x 494 pixel matrix, 7 mm in diameter). This assembly is shown in Figure 3. The holder is designed to limit the perturbations caused by the camera cable on the motion of the axes. In the current design, the mass of the camera is

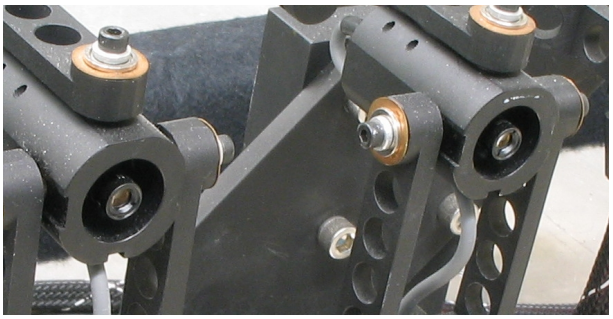


Fig. 3 Close up view of the two eyes showing the lipstick cameras.

negligible (9 g) with respect to the other components of the ASP. The NTSC video signal of each camera is fed to a Matrox Meteor II frame grabber mounted in the Frame Grabbing and Image Processing Unit (see Figure 1). The frame grabber produces 640 x 480 pixel color matrices (RGB).

5 Calibration of the ASP

As pointed out in the introduction, the main practical purpose of a second camera in an active vision system is to gather 3D information from the environment by stereopsis. 3D measurements can be obtained only if the relative pose (position and orientation) of the two cameras is known. With an active stereo sensor, this information is difficult to obtain since both cameras move independently. One possible solution is to use corresponding points in the two images, taken at a given time, to recover the epipolar geometry of the cameras. Provided that the constant intrinsic parameters of the cameras are known (either trough usual calibration or self-calibration techniques [Hem03, MF92]), this allows to recover the 3D structure of the scene up to an unknown scale factor [FL01]. The scale factor ambiguity comes from the fact that the epipolar geometry does not encode all of the information about the pose of the cameras. Another drawback of this approach is that it relies on the ability to find a sufficient large number of corresponding points between the two images, making it computationally intensive. Moreover, the quality of the result will depend on the distribution of the matched points in the scene, and on the accuracy of their localization in the images. Nevertheless, this approach has the advantage of providing a new (partial) calibration of the active sensor whenever it is needed using only the information contained in the images.

Because we seek full metric reconstruction of the scene, however, the exact pose of the cameras is required. With a carefully calibrated geometric model of the active sensor, the pose of the cameras can be computed in real-time from positional encoder readings. The accuracy of the resulting 3D measurements will then depend on the calibrated parameters, and on the precision of the mechanism. But since parallel mechanisms are capable of highly predictable rotational movements, the ASP opens the door to full metric measurements. In this section, we present a calibration procedure that was developed to investigate the possibility of producing valuable 3D data with active stereo sensors. The technique is designed to be the best compromise with respect to simplicity, robustness and accuracy. It is based on well-established, state-of-the-art calibration techniques for stereo vision. Even though it as been designed for the ASP, it can be adapted to other active sensors.

Sections 5.1 and 5.2 first present the geometric model adopted for describing the ASP. The detailed description of the calibration procedure is presented in Sections 5.3 to 5.6. Each step of the procedure is validated through experimental tests.

5.1 Geometric model

When the ASP is configured for dynamic baseline adjustment, its geometric model has 38 parameters. Table 1 lists the model elements that need to be calibrated along with the number of parameters defining each model element. Figure 4 shows a front view of the ASP with the coordinate reference frames describing the geometric model. Three reference frames are used for each eye. Frame O_{R_i} (where $i = r$ for the right eye and l for the left eye) is the base reference frame of each eye. It is a fixed cartesian reference frame with its origin being located at the intersection point between the two rotation axes of the mechanism. The X and Y axes of this frame correspond to the pan and tilt axes of the mechanism. The Z axis is directed toward the direction of observation. We refer to O_{R_i} as the robot reference frame. Frame O_{M_i}, called the manipulator reference frame, is a mobile cartesian reference frame that refers to the holder of the camera. It can be translated by the linear translation stage (in a direction parallel to vector $\mathbf{L}_i$) and rotated by the motors with respect to the tilt and pan axes of each mechanism. At startup, it is perfectly aligned with the robot frame O_{R_i}. Finally, O_{C_i}, the camera reference frame, is a cartesian frame attached to the camera according to the camera model described in Section 5.2.

In the geometric model, relations between reference frames are given using frame transformations. A frame transform is a 4x4 matrix (homogeneous coordinates) expressed as E_{AB} which gives the pose (position and orientation) of frame B with respect to frame A. Equivalently, we can use E_{AB} to transform the coordinates of points in frame B to their corresponding coordinates in frame A. Frame transforms are composed of a rotation followed by a translation (pre-multiplication):

$$E_{AB} = T_{AB} R_{AB} \tag{1}$$

Matrix $E_{R_i M_i}$, which defines the pose of the manipulator reference frame, O_{M_i}, with respect to the robot

Table 1 List of parameters used to describe the stereo pair.

Model Element	Number of parameters	Number of instances	Total number of parameters
Intrinsic parameters of the camera model	7	2	14
Pose of the left eye w/r to the right eye ($E_{R_r R_l}$ transform)	6	1	6
Pose of the camera inside it's holder ($E_{M_i C_i}$ transform)	6	2	12
Orientation of the translation stage ($\mathbf{L}_i$ vector)	3	2	6
Total			38

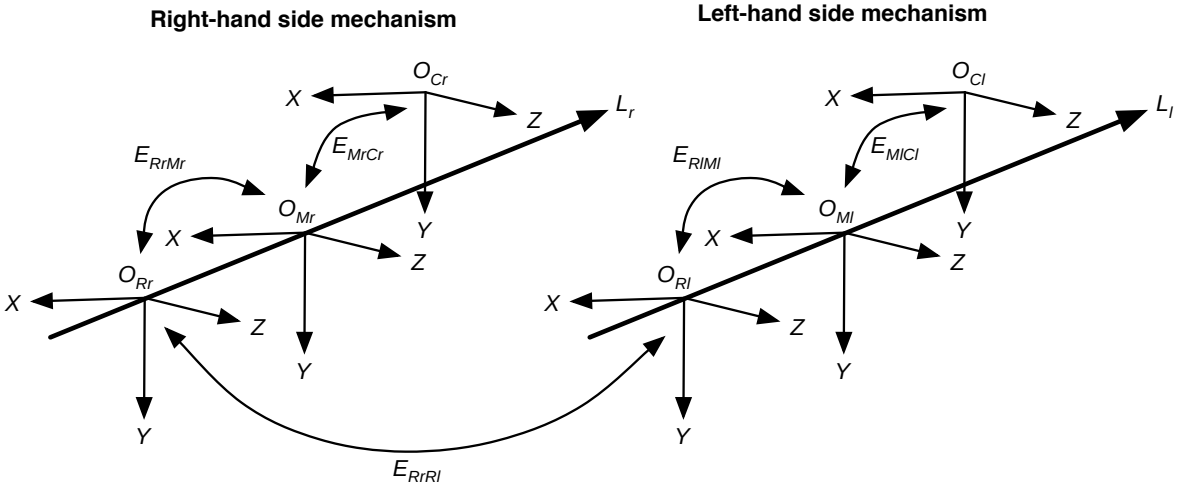


Fig. 4 Geometric model of the stereo pair.

reference frame, O_{R_i}, can be computed from the current state of the ASP, which is defined by the following parameters:

1. θ_1, the rotation angle of the motor controlling the azimuth (pan);
2. θ_2, the rotation angle of the motor controlling the elevation (tilt);
3. and α, the distance between the current position of the origin of the mechanism and its initial position at start up.

The orientation of the camera is defined by the pan (φ_1) and tilt (φ_2) angles as shown in Figure 5. They are related to the rotation angles of the motors according to the following equations derived from the geometry of the orientation mechanism [Car97]:

$$\varphi_1 = \theta_1 \tag{2}$$

$$\varphi_2 = \arctan\left[\frac{\tan\theta_2}{\cos\theta_1}\right] \tag{3}$$

Transformation $E_{R_i M_i}$ can be expressed as (in homogeneous coordinates):

$$E_{R_i M_i} = T_{R_i M_i} R_{R_i M_i, y} R_{R_i M_i, x} \tag{4}$$

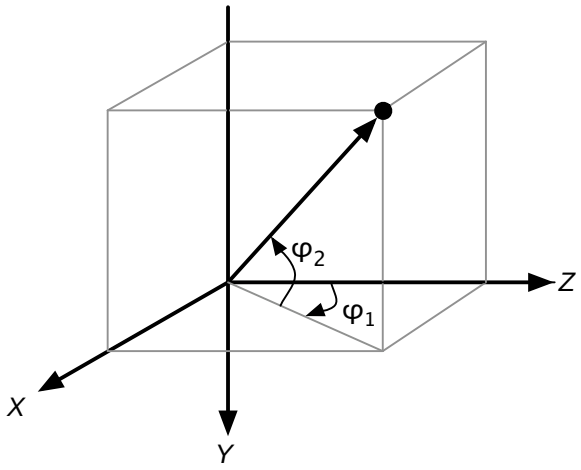


Fig. 5 Angles defining the camera orientation. They are related to the rotation angles of the actuators (θ_1 and θ_2) according to Equations 2 and 3.

where:

$$R_{R_i M_i, x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\varphi_2) & -\sin(\varphi_2) & 0 \\ 0 & \sin(\varphi_2) & -\cos(\varphi_2) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5}$$

$$R_{R_iM_i,y} = \begin{bmatrix} \cos(\varphi_1) & 0 & \sin(\varphi_1) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\varphi_1) & 0 & \cos(\varphi_1) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

$$T_{R_iM_i} = \begin{bmatrix} 1 & 0 & 0 & \alpha l_{x_i} \\ 0 & 1 & 0 & \alpha l_{y_i} \\ 0 & 0 & 1 & \alpha l_{z_i} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (7)$$

The vector $\mathbf{L}_i = [\, l_{x_i} \; l_{y_i} \; l_{z_i} \,]^T$ in equation 7 is obtained from the calibration procedure.

As it can be seen, all 3 DOF for each eye of the ASP are modeled by the $E_{R_iM_i}$ transform. As a result, only the left and right $E_{R_iM_i}$ transform matrices need to be updated while operating the ASP. The other transforms included in the model remain constant following calibration.

The transform $E_{R_rR_l}$ gives the relation between the left eye and the right eye. Its parameters are obtained from the calibration procedure as well.

The last transform included in the model, $E_{M_iC_i}$, should also be obtained from calibration. However, experiments have shown that classical calibration methods fail in achieving a satisfactory level of accuracy for this transform. Until an appropriate calibration technique is developed, the $E_{M_iC_i}$ transform is set to the identity matrix. The problem in calibrating $E_{M_iC_i}$ will be explained in Section 5.7 along with the justification for the choice of the identity matrix.

In order to compute 3D coordinates, transform $E_{C_rC_l}$ which links the left and right camera reference frames must be determined. However, because camera positions are not stationary, this transform cannot be established *a priori* but must rather be computed on the fly using the geometric model:

$$E_{C_rC_l} = E^{-1}_{M_rC_r} E^{-1}_{R_rM_r} E_{R_rR_l} E_{R_lM_l} E_{M_lC_l} \quad (8)$$

where $E_{R_iM_i}$ is built from Equation (4) using the position encoder as inputs, synchronized with image capture.

5.2 Camera model

The conventional pinhole camera model is used:

$$s\tilde{\mathbf{m}} = K \left[\, R \; \mathbf{t} \,\right] \tilde{\mathbf{M}} \quad (9)$$

where $\tilde{\mathbf{M}} = [\, X \; Y \; Z \; 1 \,]^T$ is a 3D point, $\tilde{\mathbf{m}} = [\, u \; v \; 1 \,]^T$ is its projection on the image plane, and s a scale factor. The "$\sim$" sign over a vector symbol such as $\mathbf{M}$ and $\mathbf{m}$ indicates that homogeneous coordinates are used.

The extrinsic parameters are defined by rotation matrix R and translation vector $\mathbf{t}$. They enable transformation of coordinates in the global reference frame, noted O_W, to coordinates in the camera reference frame, noted O_C. The origin of frame O_C is located at the pinhole's projection center. Its Z axis is the same as the optical axis and points in the direction of observation of the camera. The X and Y axes are aligned with the horizontal and vertical axes of the image plane.

The intrinsic parameter matrix K is given by:

$$K = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix}$$

where α and β correspond to scale factors for the X and Y axes, u_0 and v_0 are the pixel coordinates of the image plane principal point, and γ is related to the angle between the two axes of the image plane.

The radial distortion is modeled using two parameters, noted k_1 and k_2. Given the undistorted coordinates of a point in the normalized image plane (x, y), its distorted counterpart $(\breve{x}, \breve{y})$ can be computed using the following expression:

$$\breve{x} = x + x[k_1(x^2 + y^2) + k_2(x^2 + y^2)^2] \quad (10)$$

$$\breve{y} = y + y[k_1(x^2 + y^2) + k_2(x^2 + y^2)^2] \quad (11)$$

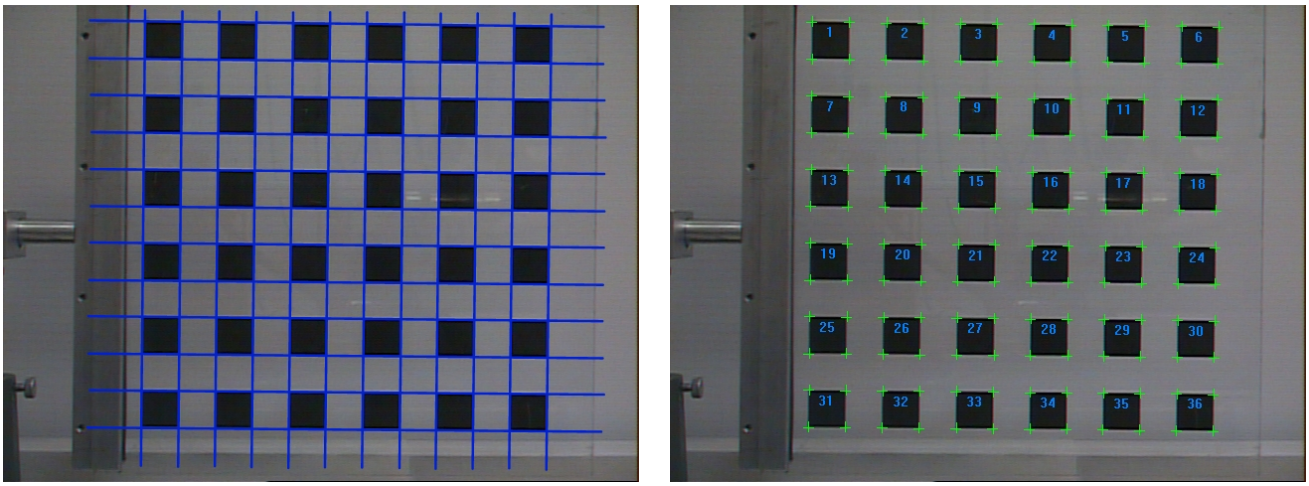
with the normalized image plane coordinates defined by:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (12)$$

5.3 Calibration target

Every calibration step for the ASP is conducted using a planar target composed of a 6×6 square grid printed on a glass substrate (see Figure 6). Two methods are used to extract characteristic points on this target, depending on whether or not the intrinsic parameters are known. When they are known, a sub-pixel detection of square contours is first carried out. Then, the coordinates of these contour points are corrected for distortion using the intrinsic parameters, and a line is fitted for each row and each column of contour points, as illustrated in Figure 6a. The calibration points of the target lie at the intersection of these lines. Finally, distortion is reapplied on these points so that they fit well with the square corners of the original image (Figure 6b).

With unknown intrinsic parameters, the procedure is slightly different. The line adjustment on a complete row or column may induce a large position error if the contour points are much affected by distortion. Thus, we limit the line adjustment to points that belong to a single square, and we process each square independently. The position of square corners are still determined from line intersections, but the precision is affected by the smaller number of points used in the estimation. It should be noted that these two methods can also be used for targets arranged as a checkerboard.



(a) (b)

Fig. 6 Features extraction for the calibration target: (a) lines fitted on rows and columns of edge points; and (b) final result for the detection of the corners of the squares.

5.4 Calibration of the intrinsic parameters

The seven intrinsic parameters of the camera model presented in Section 5.2 are estimated using the calibration method proposed by Zhang [Zha00]. This well-known approach uses 3 or more images of a planar calibration target observed from different vantage points.

To obtain better results, we use the Zhang method iteratively. For the first iteration, the characteristic points of the target cannot be extracted precisely because intrinsic parameters are unknown. We thus bootstrap the process with rough estimates and iterate with the intrinsic parameters found at the previous step, until the parameters stabilize. With low distortion lenses like those used in the ASP, this process requires about 4 or 5 iterations before convergence (see Figure 7).

5.5 Calibration of vector $\mathbf{L}_i$

Since the calibration of vector $\mathbf{L}_i$ is the same for the right ($\mathbf{L}_r$) and the left eye ($\mathbf{L}_l$), only the case of the right eye will be covered and indices identifying the camera will be omitted. Vector $\mathbf{L}$ is the direction of translation for the right eye of the ASP. Its components are obtained by observing its displacement with respect to a stationary planar target. The camera is brought to a hundred different positions on the linear stage. For each position of the camera, the frame transformation between the camera reference frame and the target reference frame is computed using the homography between points on the calibration target and their respective image on the image plane. The developments presented below enable the calculation of E_{C_jW}, the transformation between O_{C_j}, the reference frame of the camera at position j, and O_W, the reference frame of the calibration target.

By defining the reference frame of the calibration target in such a way that all of its points lie on the $Z = 0$ plane, the pinhole model (9) can be rewritten as follows:

$$s\tilde{\mathbf{m}} = K \begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t} \end{bmatrix} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{13}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the two first columns of the rotation matrix. Let:

$$H = K \begin{bmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \mathbf{t} \end{bmatrix}, \tag{14}$$

(13) becomes

$$s\tilde{\mathbf{m}} = H\tilde{\mathbf{M}}, \tag{15}$$

with $\tilde{\mathbf{M}} = [\, X \; Y \; 1 \,]^T$. The homography H is a 3×3 matrix defined up to a scale factor that establishes the relation between the points in the target plane and their projection in the image plane.

Several methods exist for estimating an homography. We use the one described in [Zha00] which is based on the following maximum likelihood criterion:

$$\min_H \sum_{k=1}^{N} \|\mathbf{m}_k - \hat{\mathbf{m}}_k\|^2 \tag{16}$$

where $\mathbf{m}_k = [\, u_k \; v_k \,]^T$ are the target points observed in the image, and $\hat{\mathbf{m}}_k = [\, \hat{u}_k \; \hat{v}_k \,]^T$ are the projected points using the homography H. Minimizing Equation (16) is a non-linear least-squares problem which can be solved using the Levenberg-Marquardt algorithm.

The minimization algorithm needs an initial estimate which can be found by solving a set of linear equations constructed from equation (15). Let $\mathbf{x} = [{\mathbf{h}_1}^T {\mathbf{h}_2}^T {\mathbf{h}_3}^T]^T$ where $\mathbf{h}_i$ is the i^{th} row-vector of H. Equation 15 can then be re-written as:

$$\begin{bmatrix} \tilde{\mathbf{M}}^T & \mathbf{0} & -u\tilde{\mathbf{M}}^T \\ \mathbf{0} & \tilde{\mathbf{M}}^T & -v\tilde{\mathbf{M}}^T \end{bmatrix} \mathbf{x} = \mathbf{0} \tag{17}$$

Each point of the target provides one such equation. With N points, we obtain a set of equations such as

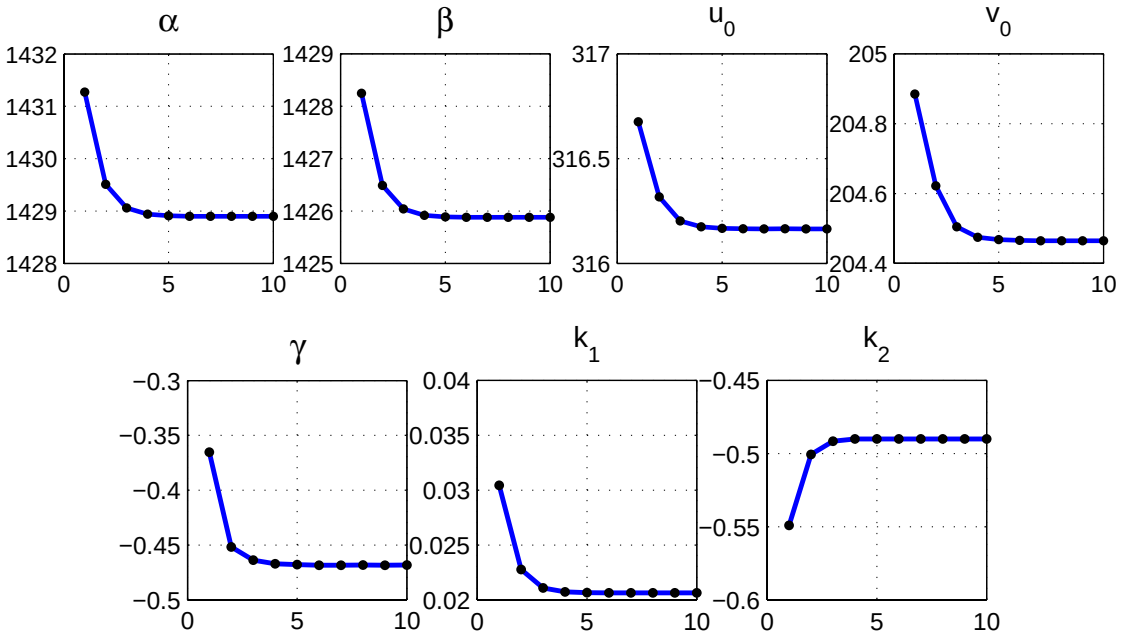


Fig. 7 Evolution of the intrinsic parameters for the Zhang calibration algorithm being used iteratively with the points detection method described in Section 5.3.

$A\mathbf{x} = \mathbf{0}$ where A is a $2N \times 9$ matrix. For $N \geq 4$, its solution is given by the eigenvector associated with the smallest eigenvalue of $A^T A$.

Once H is estimated, the next step is to extract the R and $\mathbf{t}$ components of E_{C_jW}. The following expression stems from Equation (14):

$$\mathbf{r}_1 = \lambda K^{-1}\mathbf{h_1} \tag{18}$$

$$\mathbf{r}_2 = \lambda K^{-1}\mathbf{h_2} \tag{19}$$

$$\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2 \tag{20}$$

$$\mathbf{t} = \lambda K^{-1}\mathbf{h}_3 \tag{21}$$

where

$$\lambda = \frac{1}{||K^{-1}\mathbf{h_1}||} = \frac{1}{||K^{-1}\mathbf{h_2}||} \tag{22}$$

and K is the matrix containing the intrinsic parameters estimated in Section 5.4.

Since $\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_3$ are estimated with real image data, matrix $[\,\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3\,]$ is not a pure rotation matrix. A pure rotation matrix can be derived from $[\,\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3\,]$ using the method described in [Zha00].

Until now, distortion has not been taken into account. Furthermore, with the correction leading to a pure rotation matrix, the optimization criterion (16) is no longer verified. To obtain better results, one can reinforce this criterion by optimizing parameters of E_{C_jW} while taking distortion into account. More formally, the following functional is to be minimized

$$\min_{\Omega} \sum_{k=1}^{N} \|\mathbf{m}_k - \hat{\mathbf{m}}_k\|^2 \tag{23}$$

where

$$\hat{\mathbf{m}}_k = \mathrm{Proj}(K,\, k_1,\, k_2,\, \Omega, \mathbf{M}_k)$$

are the projections of the target points using the complete camera model including distortion and

$$\Omega = \{\theta,\, \psi,\, \gamma,\, t_x,\, t_y,\, t_z\}$$

is a set of parameters describing the transform E_{C_jW} (three for rotation and three for translation). In order to reduce the number of steps for finding E_{C_jW}, one can skip the first minimization procedure. That is, computing R and $\mathbf{t}$ directly from the first estimation of H, obtained from the linear system, and then refining the result using the minimization functional of Equation (23).

The translation component of each transform E_{C_jW} gives the position of the target relative to the camera during its movement along the translation axis (see Figure 8). From the camera's vantage point, these positions appear as a cluster of points distributed along a straight line parallel to the translation axis. To find $\mathbf{L}$, the translation axis vector, a simple principal component analysis is required. In the geometric model of the ASP, the vector must be expressed with respect to the reference frame O_R. This is done by transforming the vector using the rotation component of E_{RM} and E_{MC}:

$$\mathbf{L}_R = R_{RM} R_{MC} \mathbf{L}_C \tag{24}$$

Calibration of the ASP translation axis is usually done in its initial position. Thus, $R_{RM} = I$.

To evaluate the precision of the estimated vectors, we have conducted the following experiment. The two

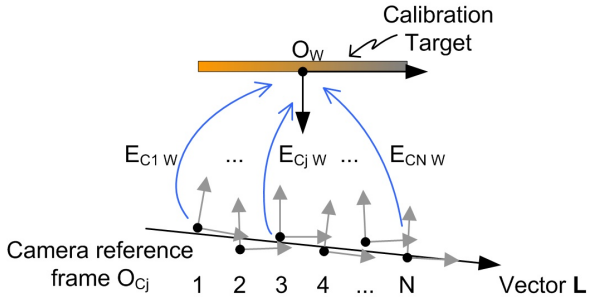


Fig. 8 Experimental procedure for estimating vector L.

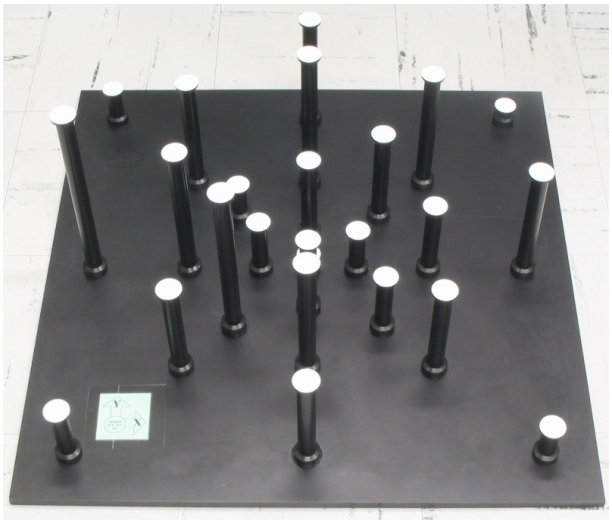


Fig. 9 High precision 3D calibration target used to assess the ASP measurement accuracy

eyes of the ASP were moved to four points so as to obtain baselines of 10, 15, 20, and 25 cm. The extrinsic parameters (E_{CrCl}) were determined for these four configurations using two methods. The first uses the geometric model of the ASP (equation 8). Its results thus depend on axes $\mathbf{L}_l$ and $\mathbf{L}_r$, as well as on the precision of the translation stages. The second method, which will serve as comparison, consists in obtaining each of the four transformations directly through calibration using the method that will be presented in Section 5.6.

Then, the two sets of extrinsic parameters obtained were used to measure an object of known dimensions. This object is the 3D calibration target shown in Figure 9. It is composed of 25 disks distributed within a volume of approximately $50 \times 50 \times 25$ cm. The positions of these disks have been measured using a high precision Coordinate Measuring Machine (CMM).

Figure 10 gives the RMS reconstruction error of this 3D target, when observed at a distance of approximately 1.43 m for each of the four considered configurations. These results show that the extrinsic parameters obtained through the two methods produce almost identical reconstruction errors. From this, we can conclude that $\mathbf{L}_r$ and $\mathbf{L}_l$ have been well estimated. Even with such satisfactory results, it remains that calculated values are

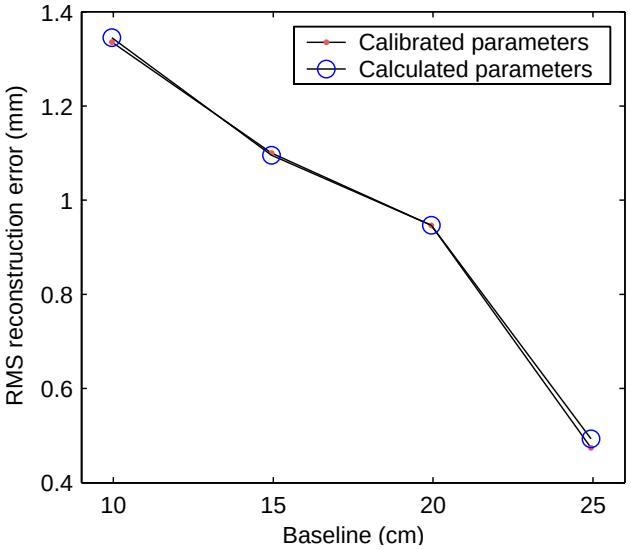


Fig. 10 Comparison of two sets of $E_{C_rC_l}$ transforms using the RMS reconstruction error of the 3D target. The $E_{C_rC_l}$ transforms obtained using the geometric model of the ASP lead to almost the same reconstruction error as $E_{C_rC_l}$ obtained from direct calibration. These results lead us to conclude that the elements of the ASP model, including the translation axes $\mathbf{L}_i$, have been successfully calibrated.

slightly less precise than the calibrated ones. The 20 cm baseline is a special case since this configuration is defined as the initial position of the ASP. As it will be explained in Section 5.6, the initial position is the one from which we take the calibrated version of $E_{C_rC_l}$ to estimate the parameters of $E_{R_rR_l}$ used in equation 8. The computed version $E_{C_rC_l}$ is thus the same as the calibrated one leading to the same reconstruction error for this configuration.

5.6 Calibration of Transformation $E_{R_rR_l}$

The calibration of $E_{R_rR_l}$ is equivalent to the well-known problem of calibrating the extrinsic parameters of a standard static stereo pair. Indeed, this transform can be determined by calibrating $E_{C_rC_l}$ when the ASP is in a given position. Knowing $E_{C_rC_l}$, $E_{R_rR_l}$ can be found easily from the geometric model of the ASP:

$$E_{R_rR_l} = E_{R_rM_r} E_{M_rC_r} E_{C_rC_l} E^{-1}_{M_lC_l} E^{-1}_{R_lM_l} \qquad (25)$$

By choosing to calibrate $E_{C_rC_l}$ with the ASP in its initial position, equation (25) is simplified since in this case, $E_{R_rM_r} = E_{R_lM_l} = I$:

$$E_{R_rR_l} = E_{M_rC_r} E_{C_rC_l} E^{-1}_{M_lC_l} \qquad (26)$$

Our approach for calibrating $E_{C_rC_l}$, inspired from [GOD00], uses several views of the planar target. The objective is to find the extrinsic parameters that minimize the reprojection error of the target points in the

images. Again, this is a non-linear optimization problem that can be solved using the Levenberg-Marquardt algorithm. With N views of a target of M points, the objective function to be minimized is:

$$\min_{\Omega} \sum_{i=1}^{N} \sum_{j=1}^{M} \|\mathbf{m}_{ij_r} - \hat{\mathbf{m}}_{ij_r}\|^2 + \|\mathbf{m}_{ij_l} - \hat{\mathbf{m}}_{ij_l}\|^2 \qquad (27)$$

where:

$$\hat{\mathbf{m}}_{ij_r} = \text{Proj}(K_r,\, k_{1_r},\, k_{2_r},\, E_{C_rW_i}, \mathbf{M}_{ij})$$

and:

$$\hat{\mathbf{m}}_{ij_l} = \text{Proj}(K_l,\, k_{1_l},\, k_{2_l},\, E_{C_rC_l}, E_{C_rW_i}, \mathbf{M}_{ij})$$

are respectively the projection of point j on view i for the right and left images. These projections are applied using the camera model described in Section 5.2. For the camera on the right, the $E_{C_rW_i}$ transforms are used to bring the points from the target reference frame to the camera reference frame. For the camera on the left, this operation is achieved through $E_{C_lW_i}$, which are obtained by combining the $E_{C_rW_i}$ transforms with $E_{C_rC_l}$:

$$E_{C_lW_i} = E_{C_rC_l}^{-1} E_{C_rW_i}$$

In this way, the reprojection error can be linked to the parameters to be estimated. With known intrinsic parameters, the set Ω of parameters that require optimization include those of transform $E_{C_rC_l}$ and those of the N $E_{C_rW_i}$ transforms.

The procedure requires a first estimation of parameters. For $E_{C_rW_i}$ which provides the target's pose in the reference frame of the camera, the technique described in Section 5.5 is used. The same approach is used for $E_{C_lW_1}$ which provides the camera-target relationship for the first left image. Combining this transform with $E_{C_rW_1}$, we obtain the first estimation for $E_{C_rC_l}$:

$$E_{C_rC_l} = E_{C_rW_1} E_{C_lW_1}^{-1}$$

The plots of Figure 11 illustrate the effect of the optimization procedure on the reprojection error. The data shows the RMS reprojection error for a given image with respect to its view number:

$$\xi_{\text{RMS}} = \left[\frac{1}{M} \sum_{j=1}^{M} \|\mathbf{m}_{ij} - \hat{\mathbf{m}}_{ij}\|^2 \right]^{1/2}$$

Figure 11a gives the reprojection error before optimization. It should be noted that the error for left images is significantly larger than those of the right images. This can be explained by the fact that transforms $E_{C_rW_i}$, now used to compute the re-projections in the right images, have been estimated by a procedure whose objective was to minimize the re-projection errors (see Section 5.5). On the other hand, the left image reprojections are computed from a combination of $E_{C_rW_i}$ and

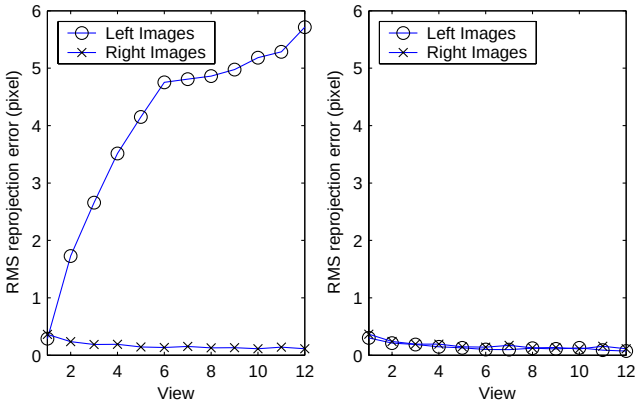


Fig. 11 RMS reprojection error before (a) and after (b) the optimisation procedure

$E_{C_rC_l}$. Since $E_{C_rC_l}$ was determined from the first view as a first approximation, only the corresponding left error is comparable to observed errors in the right images. This illustrates the well known fact that extrinsic parameters obtained from a single view are valid only for stereo measurements that are made in that view plane. As illustrated in Figure 11b, the optimization process compensates for this limitation by adjusting $E_{C_rC_l}$ so that it is valid for all views. Consequently, the extrinsic parameters become valid within the volume that contains the complete set of calibration views.

5.7 Calibration of Transformation $E_{M_iC_i}$

As it has been pointed out at the beginning of the section, $E_{M_iC_i}$ cannot be estimated using classical calibration methods. Actually, the only way to evaluate the position of O_{C_i} with respect to O_{M_i} is by moving the camera using the orientation mechanism. By analyzing the spherical trajectory of the camera principal point (O_{C_i}) around the fixed rotation center of the manipulator (O_{M_i}), one could estimate $E_{M_iC_i}$. But tracking O_{C_i} in its trajectory around O_{M_i} depends on the ability to estimate its pose with respect to a global reference frame. Unless a very large calibration target covering the entire field of view of the ASP can be built, this cannot be achieved using conventional methods. This problem is the counterpart of one of the ASP main advantages: its extended field of view.

In order to circumvent the problem of calibrating the pose of the camera with respect to the orientation mechanism, most active systems have been designed in such a way that the cameras can be positioned in their holder so that their center of projection is almost aligned with the center of rotation of the devices [CBM93, FC93, MJT93, WFRL93, US92, SMMB98]. The ASP is not an exception, its camera holder allows such an alignment. With a perfect alignment of camera and manipulator reference frames, $E_{M_iC_i}$ would be the identity matrix. Most

researchers that used this approach do not include the $E_{M_iC_i}$ transform in their geometric model, thus assuming perfect alignment. But in practice, this assumption may not be true.

In the context of a stereo sensor, even a slight non compensated misalignment of the cameras may lead to significant errors on computed 3D data. Although this fact has not been properly proven in the literature (very few experimental results on 3D measurements are available), it can be deduced from related works on active sensors. For example, research on self-calibration techniques based on rotating cameras demonstrated that better results are obtained if the translational offset between the camera and the rotation axes is considered [JD04, WKSX04, HM03]. This is particularly true if observed points are close to the sensor [HM03], which is the case for stereo vision.

Some calibration methods have been proposed [Li98, DJK02]. Results are however not satisfactory. As explained before, the problem lies in the fact that reference points used for calibration rapidly exit the field of view of the rotating cameras. Wada *et al.* [WM96] proposed a technique that uses a laser beam for the alignment of a rotating camera. Based on rotations, which are inherently limited by the field of view of the camera, this technique is not accurate enough for stereo vision. To our knowledge, the only calibration procedure that successfully estimated $E_{M_iC_i}$ is the one presented by Shih *et al.* [SHL98]. They addressed the problem of the range of rotation using a calibration target mounted on a long range, accurate translation stage. This setup is the equivalent of the large calibration target mentioned before, which is both expensive and cumbersome. Thus, in the absence of a satisfactory solution to this problem, we move on to evaluate the performance of the ASP based on the assumption that $E_{M_iC_i}$ is the identity matrix.

6 Performances of the ASP

In this section, the performances of the ASP is evaluated with respect to both dynamical properties and 3D measurement accuracy. Dynamical performances are first given in Section 6.1. Then, Section 6.2 presents the maximum achievable theoretical accuracy of the ASP for 3D measurements, and these predictions are validated experimentally in Section 6.3. Finally, from a functional point of view, Section 6.4 describes a target tracking experiment that demonstrates the dynamic capabilities of the ASP.

6.1 Dynamic performances

The ability of an active vision system to rapidly change its gazing direction, for example to track a moving target, depends on its dynamical characteristics. These capabilities are performed with two kinds of movements, smooth pursuit and saccades. Smooth pursuit is intended to keep a moving target at the center of the field of view of the cameras while saccades are used to rapidly change the focus of attention from one target to another. Saccadic movements require that the eyes start from rest at a given position and stop at another position as quickly as possible. The movement must be performed in a short period of time in order to miss as few video frames as possible (images taken during saccadic movements are blurred due to camera motion).

Table 2 Dynamical properties of the ASP according to its different degrees of freedom

Parameters	Span	Speed	Acceleration
pan	$\pm 40^\circ$	$1950^\circ/\text{sec}$	$78000^\circ/\text{sec}^2$
tilt	$\pm 40^\circ$	$1350^\circ/\text{sec}$	$40000^\circ/\text{sec}^2$
baseline	50 cm	8 cm/sec	

The dynamic characteristics of the ASP are summarized in Table 2 which gives maximum position, velocity, and acceleration values for its different degrees of freedom. As shown in this table, low inertia due to the parallel design of the orienting devices leads to high velocity and acceleration. Performances for the baseline adjustment come from manufacturer's specifications.

To determine the dynamic characteristics related to the orientation mechanism, the system was directed to move each axis from one end of its angular range to the other as quickly as possible. To measure speed and acceleration, the angular position was sampled at 1 ms intervals. The position was then differentiated using a 3-point scheme and low-pass filtered by a mean of 7-point averaging filter (our procedure is similar to the one used in [TARZ00]). Results for the pan axis are shown in Figure 12. The system started from rest at 10 ms and settled down at the end of its course 87 ms later. Tilt axis executed its 80° angular motion in 130 ms. Tilt axis carries more load, which explains the longer execution time. The experiment was performed on each axis separately since they have different dynamic properties. Moving both axes simultaneously however has no effect on performances.

Considering the slower axis, a maximum of 4 video frames are lost during a full range saccade. Obviously, shorter saccades will execute more quickly. For example, a 15° movement, which would bring to the center an object located at the field of view border, is performed in 60 ms. When a succession of saccades is required, a pause between them must be added in order to let the system analyse the new situation and make a decision about the next destination. Depending on the task, a period of time corresponding to 4 video frames may be enough, as suggested in [BDZ+97]. In these conditions, the ASP is able to perform 3 to 5 saccadic movements per second. This is a little more than the capability of

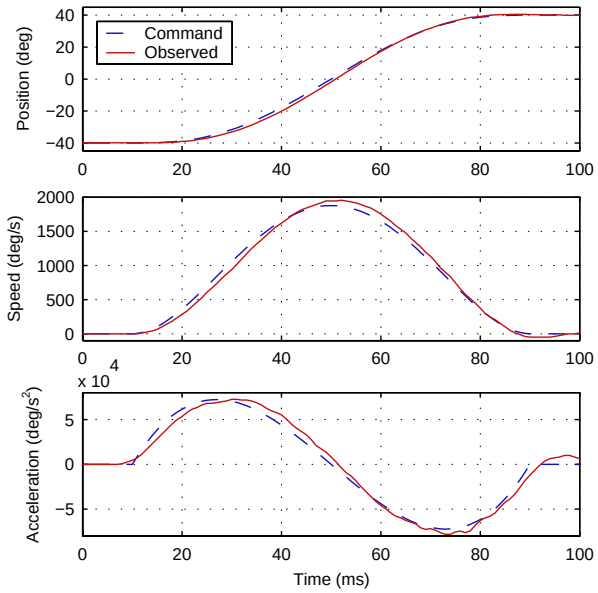


Fig. 12 While performing an 80° panning motion, the eyes of the ASP reach a speed of 1950°/sec and an acceleration of 78000°/sec^2. The whole stop-to-stop movement, was executed in 87 ms.

the human eye which performs 3 to 4 saccades per second [Rod98].

6.2 Theoretical measurement accuracy of the ASP

The accuracy of the ASP depends on several factors. Even though some of these factors are specific to the ASP, most of them are also common with static stereo pairs: baseline, distance, focal length, pixel size, quality of calibration parameters (intrinsic and $E_{C_rC_l}$), and accuracy of target points. The factors that are specific to the ASP are: orientation of cameras, accuracy of position encoders, and quality of calibrated parameters ($\mathbf{L}_i$ axes and $E_{M_iC_i}$).

The plots in Figure 13 show the maximum achievable theoretical accuracy of the ASP with respect to the three dynamic parameters (pan, tilt and baseline), and the distance of scene points. The default parameter values are: baseline = 20 cm, pan = 0°, tilt = 0°, and distance = 1.5 m. These theoretical assessments are also based on the following intrinsic parameters: focal length of cameras = 8 mm, sensor pixel size = 5.7 μm (square), and accuracy of target points = ±0.1 pixels. Moreover, perfect lenses without distortion are considered.

The current prototype is equipped with 11-bit encoders for measuring pan and tilt angles which enable an accuracy of ±0.09° for both orientations. Comparing this value with a one pixel arc-length, then the encoder error corresponds to ±2.2 image pixels. The encoders thus constitute the prevalent source of error in the system. We are currently contemplating the possibility of replacing these encoders with compact 17-bit models that would induce errors of only ±0.034 pixels. The right hand plots in Figure 13 give the accuracy that could be obtained by an ASP with such 17-bit encoders.

6.3 Experimental assessment of the ASP accuracy

In this section, we provide an experimental evaluation of the accuracy of the ASP by observing an object whose dimensions are known with high precision. The complete procedure is illustrated by Figure 14.

The object used in this experiment is the 3D calibration target that was presented in Section 5.5. The disks on this target are segmented using a sub-pixel edge detection algorithm. Ellipses are fitted on detected edge points and the center of these ellipses is then used to compute the 3D coordinates of the target's disks. The cluster of points thus obtained is then registered with the target's model in order to compute the RMS reconstruction error. More formally:

$$\text{RMS Error} = \left[\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{M}_i - \hat{\mathbf{M}}_i\|^2\right]^{1/2}$$

where N is the number of target disks observed in both images, $\mathbf{M}_i$ gives the coordinates of the target point i according to the target model, and $\hat{\mathbf{M}}_i$ provides the coordinates of the target point i measured with the ASP, after registration.

This procedure was repeated for several configurations: 7 pan orientations, 7 tilt orientations, and 4 baselines. For each configuration, the target was moved so that the majority of the disks were visible in both images. Moreover, the distance between the center of the ASP (halfway between O_{R_r} and O_{R_l}) and the center of the target was kept constant at about 1.43m.

The experimental results are shown in Figure 15. They show that the reconstruction error is less than the theoretical maximum error presented in Section 6.2. There are two reasons for this phenomenon. The first is the fact that each target reconstruction is built from a single pair of images. In other words, the target points are reconstructed by the ASP for a single configuration. The encoder reading error is thus the same for all points. Therefore, this error has a global impact on the cluster of points instead of causing an independent positioning error on each point. For example, it can produce a scale factor or a translation. The second factor is the registration process that adjusts the measured points on the target model. This is achieved by minimizing the reconstruction error. Any systematic positioning error is thus eliminated. These two factors reduce the error for this experiment well below the theoretical limitations induced by the position encoders of the ASP.

To validate this last assertion, the 3D target measurement experiment was reproduced in simulation. The experimental procedure is the same as above, including the

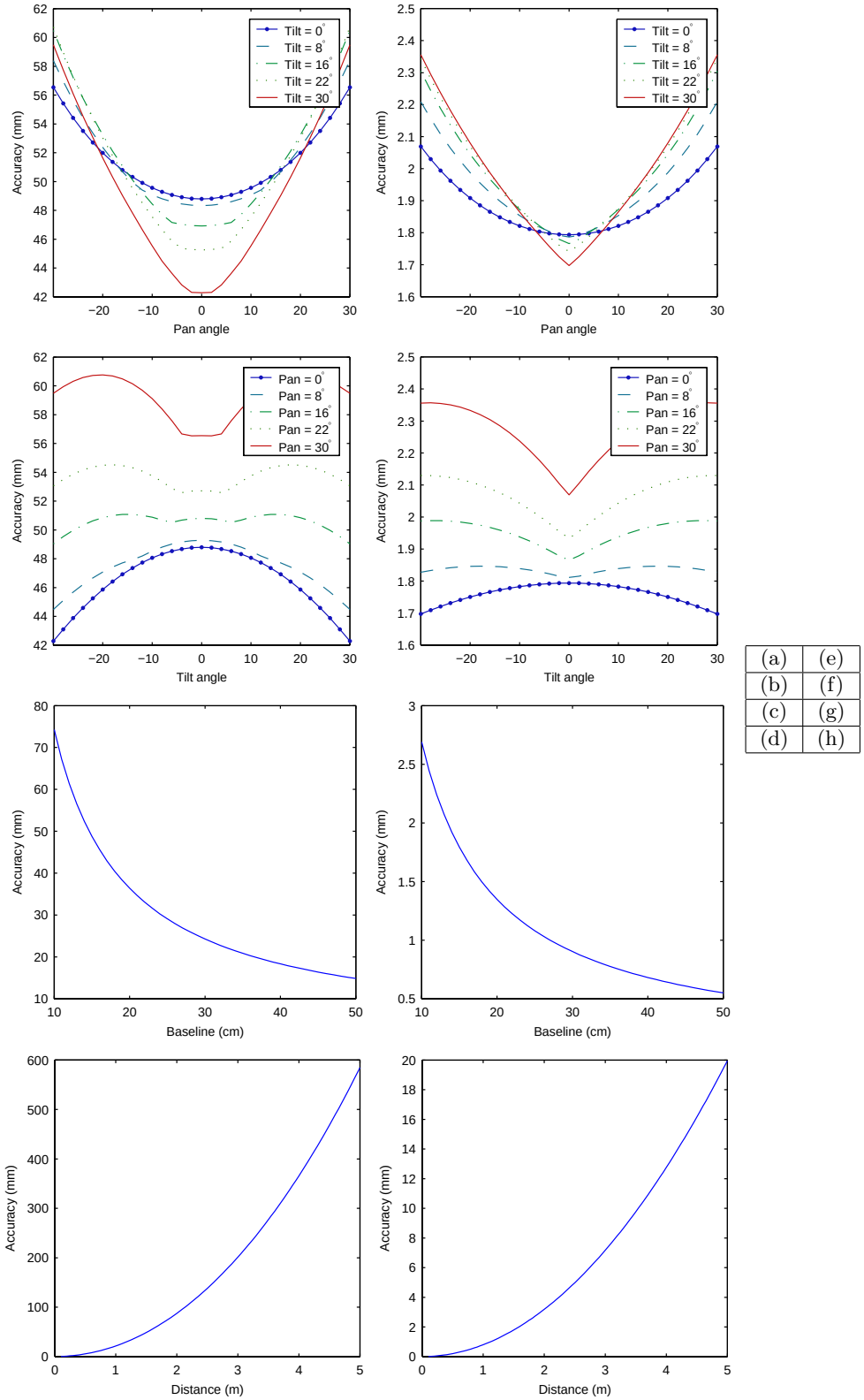


Fig. 13 Theoretical measurement accuracy of the ASP using 11bits encoders (a-d) and 17-bit encoders (e-h). The accuracy is given as a function of pan angle (a,e), tilt angle (b,f), baseline (c,g), and distance (d,h). When not specified, variables are set to Pan = 0°, Tilt = 0°, Distance = 1.5m, Baseline = 20cm.

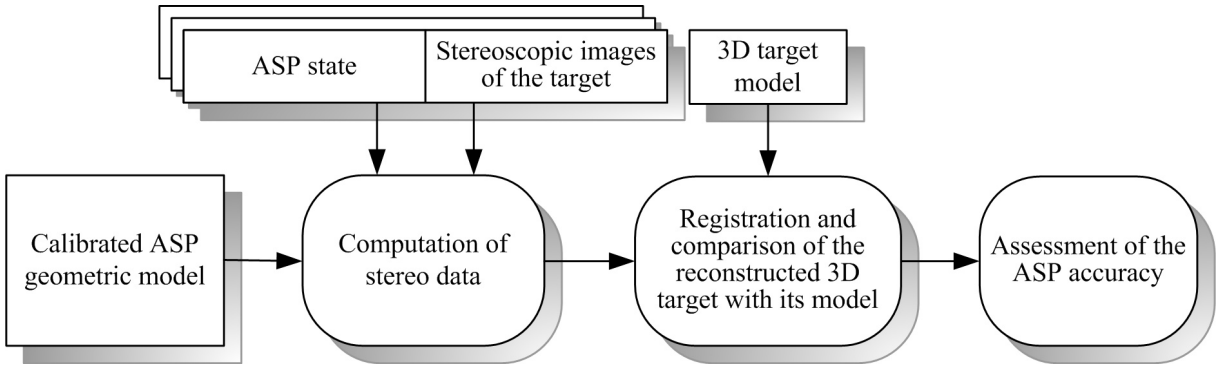


Fig. 14 Experimental procedure for evaluating the accuracy of the ASP.

assumptions for the simulation of Section 6.2. Two error sources are considered. The first is the disk positioning errors in the images which was simulated by adding a random Gaussian noise $N(0, \sigma)$ with $sigma = 0.1$ pixel. The second error source is the encoders read error. For each of the studied configurations of the ASP (7 pan orientations, 7 tilt orientations, and 4 baselines), its impact has been evaluated by performing several 3D reconstructions according to various combinations of error on the four encoders. The combination leading to the worst reconstruction error is kept so the results of these simulations give the maximum reconstruction error of the target. Plots (d) to (f) in Figure 15 show these results. It should be observed that they are consistent with those of the real experiment, which demonstrates that the deviation from the theoretical limits of the ASP is indeed related to the experimental protocol.

6.4 Stereo tracking with real-time depth estimation

A simple tracking experiment was designed to demonstrate the dynamic features of the ASP. A red Light Emitting Diode (LED) was mounted at the tip of a wooden stick and moved in front of the ASP. The 3D coordinates of the LED were computed by stereo in real-time and the LED was tracked by each camera of the ASP in such a way as to keep it near the center of the images.

The tracking algorithm currently implemented is quite basic but very successful at keeping the tracked object (the LED) near the center of the image and producing smooth motions. The reaction time of the whole system is 61 ms. That is, if a stationary object starts moving, the eyes will begin their pursuit 61 ms after that event. As a comparison, the human visual system needs 180 ms to 200 ms to react to a given event [Rod98]. In order to avoid jagged movements in smooth pursuit, the tracking algorithm needs information about the speed of the moving target. So, it does not instruct the eyes to catch the target at the very moment that the motion is detected. Instead, it actually follows the exact trajectory of the object with a 111 ms lag. The lag could be reduced with the addition of a predictive scheme to the tracking algorithm.

Figure 16 shows the field of view of the stereo pair as well as the computer interface of the ASP. As the LED is moving around, the system computes its 3D coordinates with respect to the ASP fixed reference frame O_{R_r}. Its path is displayed in a 3D window in real time. Figure 17 shows two examples of trajectories captured by the system : the word "hello" (Figure 17a) and a spiral (Figures 17b and 17c). Both paths are clearly visible.

As expected, and observed qualitatively by the increased jaggedness of the 3D path, the accuracy of depth measurements decreases as the distance between the LED and the stereo pair increases (along the positive Z axis). This noise is mainly due to position encoders. As opposed to the experiment of Section 6.3, each point on the path has been observed with the ASP in a different configuration since the system is in a tracking mode. The encoder reading error thus varies randomly from point to point. When brought together in one of the fixed reference frames of the ASP to form a path such as the ones of Figure 17, the observed noise corresponds to the absolute ASP accuracy as defined in Section 6.2. If the tracking function is turned off, the cameras remain fixed and the observed trajectories are similar to the one presented in Figure 18. However, we then lose the principal interest of the ASP, that is, its enlarged field of view granted by the mobility of its cameras. As a matter of fact, this is the reason why this last spiral has a more elongated form than the previous one: it had to be drawn in a much smaller volume because of a smaller common field of view between the cameras of the stereo pair.

7 Conclusion and future work

This paper presented a new active stereo sensor called the Agile Stereo Pair (ASP). The ASP is characterized by its mechanical design: two 2 DOF parallel orientation mechanisms that allow the cameras to be oriented independently in a fast and accurate fashion. High performances are achieved with low-end DC motors. Both eyes are modular so they can easily be integrated in dif-

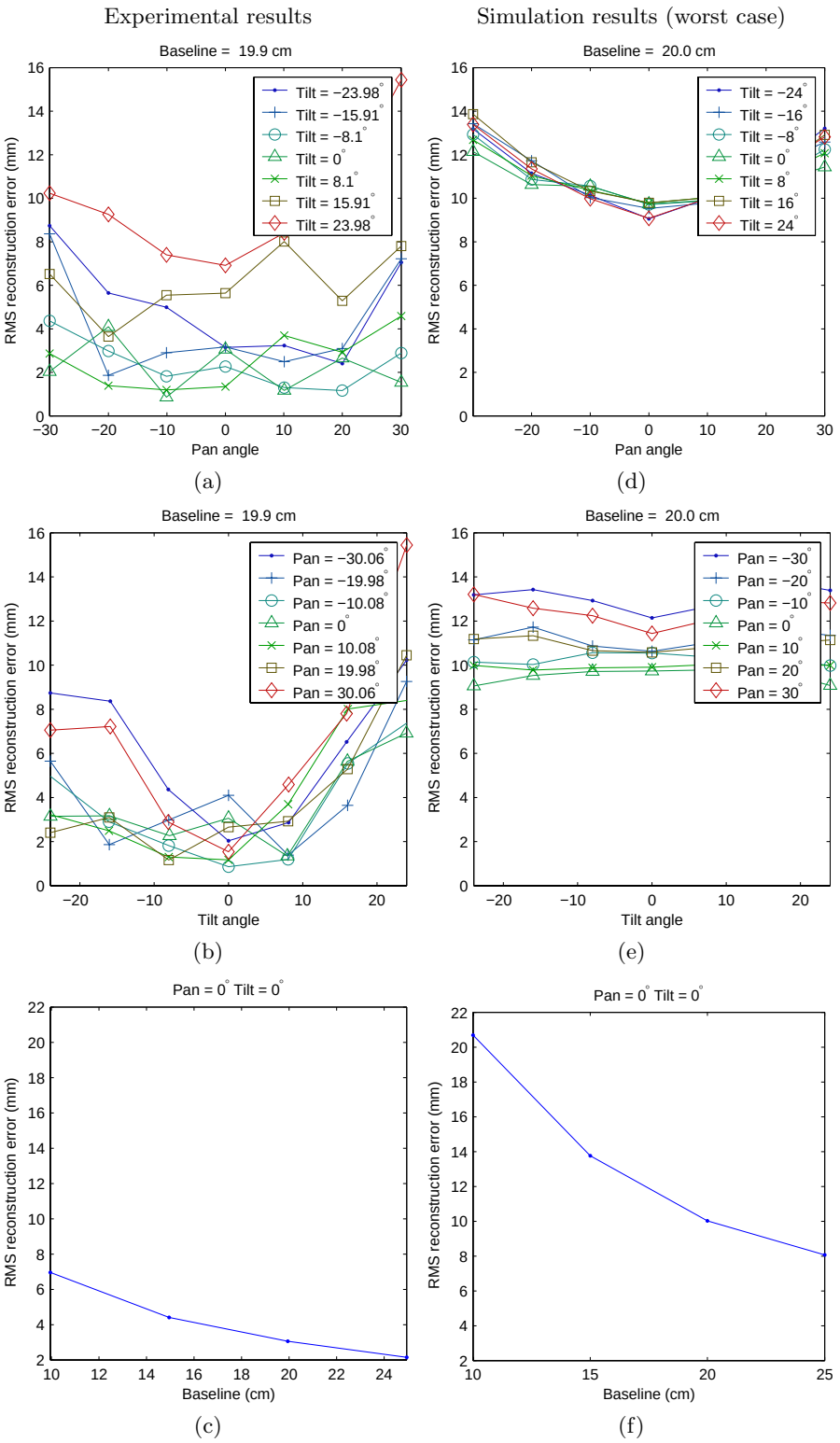


Fig. 15 RMS reconstruction error of a 3D calibration target observed at a distance of 1.43m as a function of pan angle (a,d), tilt angle (b,e) and baseline (c,f). Plots (a) to (c) present experimental results whereas plots (d) to (f) present the results of a simulation of the same experiment.

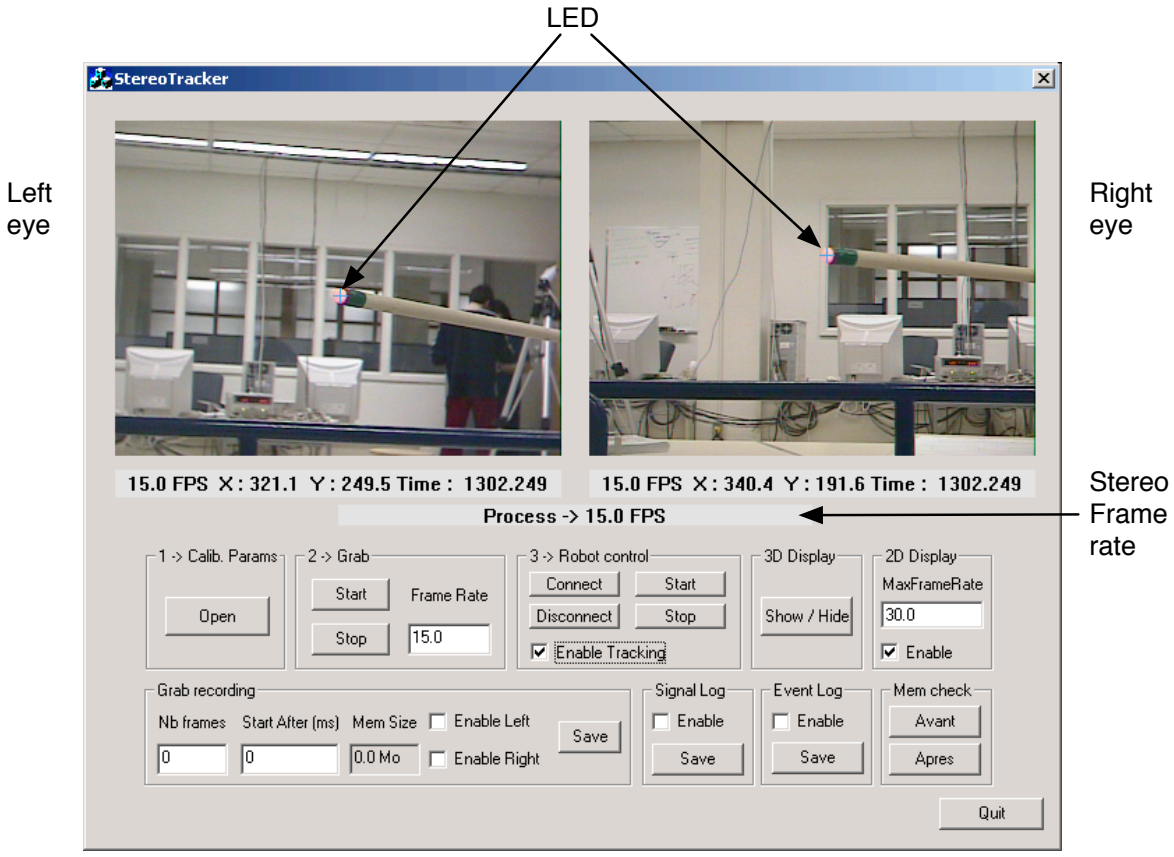


Fig. 16 Field of view of the stereo pair as displayed in real-time on the ASP computer interface.

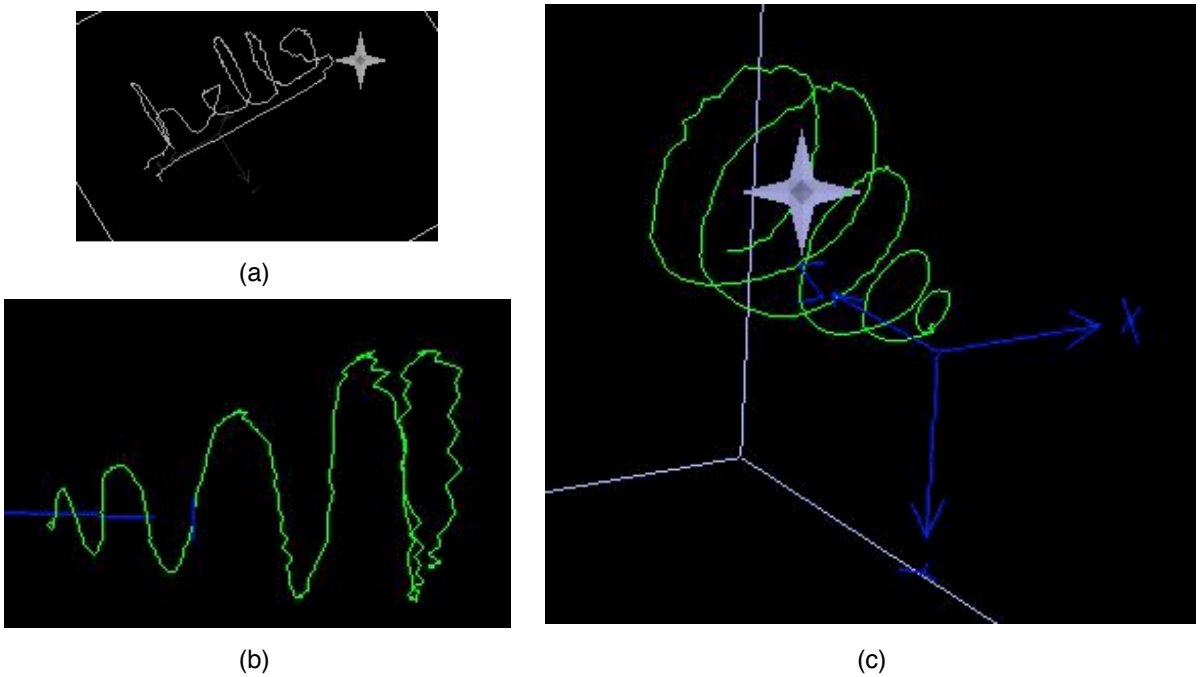


Fig. 17 Examples of trajectories captured by the ASP in the simple stereo experiment: (a) word "hello"; (b) side view of a spiral drawn into the air; (c) 3D view of the same spiral.

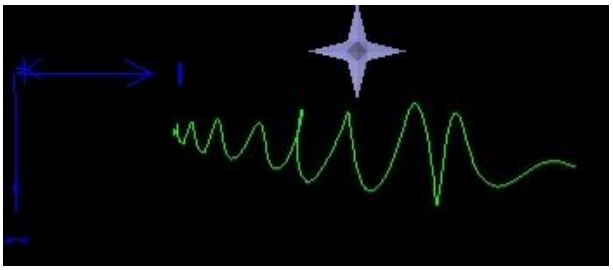


Fig. 18 Side view of a spiral pattern with the tracking option of the ASP turned off.

ferent system configurations. In this paper, we presented the ASP with both eyes mounted on two accurate translation stages allowing for real time baseline adjustment. With this added feature, the ASP has a total of 6 DOF.

A geometric model describing the functional characteristics of the ASP was proposed along with a calibration procedure. Each step of the procedure was fully detailed and its validity was demonstrated through the presentation of results (intrinsic and extrinsic parameters) and a simple experiment (translation vectors). The experiment on translation vectors showed that the dynamic baseline adjustment feature can be used without significant loss in accuracy. Finally, the performance of the whole system was demonstrated in simulation and with experiments. The results demonstrate that the ASP behaves as predicted by the simulations and that it can perform high speed saccadic movements as well as smooth pursuit motions.

According to the results of 3D measurement experiments, it is clear that the resolution of the currently used encoders is a limitation to the accuracy of the system. In these circumstances, the full potential of the parallel mechanism cannot be exploited. The next step in the ASP development will thus be to replace these encoders with higher resolution ones that are now becoming available. With new encoders, however, the problem of calibrating the $E_{M_iC_i}$ transform may be of concern. As explained in section 5.7, any misalignment of the cameras in their holders, which is compensated for by this transform, could be an important source of error in 3D measurements.

Up to now, we have not targeted the ASP to any particular active vision application. Our focus was on the development of a versatile sensor able to perform accurate measurements. As we believe that the ASP offers very good characteristics, there is a broad range of possibilities. With its ability to actively change its focus of attention in real-time while continuously acquiring 3D data, the ASP could be used in applications such as autonomous mobile robot guidance [DM98], 3D scene modeling [AA93, KB98, LSHT02], remote sensing/operation, industrial inspection, visually guided manipulation, surveillance, face tracking [DZ04], and human gaze sensing [AZ02, BF03]. Also, any application could benefit from dynamic baseline adjustment since it allows for multi-scale acquisitions. Multi-scale acquisitions are a good compromise between the large common field of view conferred by short baselines and high accuracy measurements delivered by long baselines. That is, dynamic baseline allows for fast and coarse scanning of the environment followed by refined acquisitions on area of interest. In addition, some researchers have used dynamic baseline to relieve ambiguities arising in stereo matching algorithms [OK93, JKKH01]. Finally, as the ASP is able to mimic human eye movements, it can help social interactions between humans and robots [DRH+97, BEFS01].

Acknowledgments

This research was supported by NSERC-Canada and FQRNT-Québec grants to D. Laurendeau, M. Parizeau and C. Gosselin. The authors also express their gratitude to A. Schwerdtfeger for proofreading this manuscript.

References

[AA93] A. Lynn Abbott and Narendra Ahuja. Active stereo: Integrating disparity, vergence, focus, aperture, and calibration for surface estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1007–1029, October 1993.

[AWB88] J.Y. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1:333–356, 1988.

[AZ02] Rowel Atienza and Alexander Zelinsky. Active gaze tracking for human-robot interaction. In *Fourth IEEE International Conference on Multimodal Interfaces*, pages 261–266, October 2002.

[Baj88] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE, Special issue on Computer Vision*, 76(8):966–1005, 1988.

[Bal91] D. Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[BDZ+97] Andrew Brooks, Glenn Dickins, Alexander Zelinsky, Jon Kieffer, and Samer Abdallah. A high-performance camera platform for real-time active vision. In *In Proceedings of the First International Conference on Field and Service Robotics, Canberra, Australia*, pages 559–564, 1997.

[BEFS01] Cynthia Breazeal, Aaron Edsinger, Paul Fitzpatrick, and Brian Scassellati. Active vision for sociable robots. *IEEE Transactions on Man and Cybernetics Systems, Part A*, 31(5):443–453, September 2001.

[BF03] David Beymer and Myron Flickner. Eye gaze tracking using an active stereo head. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, volume 2, pages 451–458, June 2003.

[BFZ93] M. Buffa, O. Faugeras, and Z. Zhang. A stereovision-based navigation system for a mobile robot. Technical Report 1895, INRIA, France, Mai 1993.

[Car97] François Caron. Analyse et conception d'un manipulateur parallèle sphérique à deux degrés de liberté pour l'orientation d'une caméra. Master's thesis, Laval University, Quebec, QC, Canada, G1K 7P4, August 1997.

[CBM93] James L. Crowley, Philippe Bobet, and Mouafak Mesrabi. Layered control of a binocular camera head. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(1):109–122, 1993.

[DJK02] Guilherme Nelson DeSouza, Andrew H. Jones, and Avinash C. Kak. An world-independent approach for the calibration of mobile robotics active stereo heads. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation Washington, DC*, pages 3336–3341, May 2002.

[DL01] C. Dima and S. Lacroix. Using multiple disparity hypotheses for improved indoor stereo. Technical Report 01458, LAAS, France, October 2001.

[DM98] Andrew J Davison and David W Murray. Mobile robot localisation using active vision. In *Proceedings of the 5th European Conference on Computer Vision, Freiburg, Germany*, volume 1407, pages II:809–825. Springer, 1998.

[DRH+97] J. Demiristt, S. Rougeaux, G. M. Hayes, L. Berthouze, and Y. Kuniyoshi. Deferred imitation of human head movements by an active stereo vision head. In *Proceedings of the 6th IEEE International Workshop on Robot and Human Communication*, pages 88–93, 1997.

[DZ04] Andrew Dankers and Alexander Zelinsky. Cedar: A real-world vision system. mechanism, control and visual processing. *Machine Vision and Applications*, 16:47–58, 2004.

[FC93] Nicolas J. Ferrier and James J. Clark. The harvard binocular head. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(1):9–31, 1993.

[FL01] Olivier Faugeras and Quang-Tuan Luong. *The geometry of Multiple Images : The Laws That Govern the Formation of Multiple Images of a Scene and Some of Their Applications*. The MIT Press, Cambridge, Massachusetts, 1 edition, 2001.

[GC99] Clément M. Gosselin and François Caron. Two degree-of-freedom spherical orienting device. US Patent 5966991, October 1999.

[GOD00] Dorian Garcia, Jean-José Orteu, and Michel Devy. Accurate calibration of a stereovision sensor: Comparison of different approaches. In *5th Workshop on Vision, Modeling, and Visualization, Saarbrücken (Germany), 22-24 November*, 2000. Rapport LAAS No00301.

[GSP97] Clement M. Gosselin and Eric St-Pierre. Development and experimentation of a fast three-degree-of-freedom camera-orienting device. *The International Journal of Robotics Research*, 16(5):619–630, October 1997.

[Hem03] Elsayed E. Hemayed. A survey of camera self-calibration. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)*, pages 351–357, July 2003.

[HM03] Eric Hayman and David W. Murray. The effects of translational misalignment when self-calibrating rotating and zooming cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):1015–1020, August 2003.

[JD04] Qiang Ji and Songtao Dai. Self-calibration of a rotating camera with a translational offset. *IEEE Transactions on Robotics and Automation*, 20(1):1–14, February 2004.

[JKKH01] Jeonghee Jeon, Kyungsu Kim, Choongwon Kim, and Yo-Sung Ho. A robust stereo-matching algorithm using multiple-baseline cameras. In *PACRIM. 2001, IEEE Pacific Rim Conference on Communications, Computers and signal Processing*, volume 1, pages 263–266, August 2001.

[KB98] William N. Klarquist and Alan Conrad Bovik. Fovea: A foveated vergent active stereo vision system for dynamic three-dimensional scene recovery. *IEEE Transactions on robotics and Automation*, 14(5):755–770, October 1998.

[KFS88] E. Krotkov, F. Fuma, and J. Summers. An agile stereo camera system for flexible image acquisition. *IEEE Journal of Robotics and Automation*, 4(1):108 – 113, February 1988.

[KSJ91] Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell. *Principles of neural science*. New York : Elsevier, 3 edition, 1991.

[Li98] Mengxiang Li. Kinematic calibration of an active head-eye system. *IEEE Transactions on Robotics and Automation*, 14(1):153–158, February 1998.

[LJM01] S. Lacroix, I.K. Jung, and A. Mallet. Digital elevation map building with low altitude stereo imagery. In *9th Symposium on Intelligent Robotic Systems*, pages 207–216, Toulouse (France), July 2001.

[LSHT02] Chung-Yi Lin, Sheng-Wen Shih, Yi-Ping Hung, and Gregory Y. Tang. A new approach to automatic reconstruction of a 3-d world using active stereo vision. *Computer Vision and Image Understanding*, 85(2):117–143, February 2002.

[MF92] S.J. Maybank and O.D. Faugeras. A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123–151, August 1992.

[MJT93] E. Milios, M. Jenkin, and J. Tsotsos. Design and performance of trish, a binocular robot head with torsional eye movements. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(1):51–68, February 1993.

[MLG00] A. Mallet, S. Lacroix, and L. Gallo. Postion estimation in outdoor environments using pixel tracking and stereovision. In *International Conference on Robotics and Automation*, volume 4, pages 3519–3524, San Francisco, CA (USA), April 2000. IEEE.

[Mor96] Hans P. Moravec. Robot spatial perception by stereoscopic vision and 3d evidence grids. Technical Report CMU-RI-TR-96-34, Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, September 1996.

[NFM+04] Y. Nakabo, N. Fujikawa, T. Mukai, Y. Takeuchi, and N. Ohnishi. High-speed and bio-mimetic control of a stereo head system. In *SICE Annual*

Conference in Sapporo, pages 2371–2376, August 2004.

[OK93] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363, April 1993.

[PE92] Kourosh Pahlavan and Jan-Olof Eklundh. A head-eye system: Analysis and design. *CVGIP: Image Understanding*, 56(1):41–56, July 1992.

[Rod98] Robert W. Rodieck. *The First Steps in Seeing*. Sinauer Associates, 1 edition, 1998.

[SHL98] Sheng-Wen Shih, Yi-Ping Hung, and Wei-Song Lin. Calibration of an active binocular head. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 28(4):426–442, July 1998.

[SMMB98] P.M. Sharkey, D.W. Murray, P.F. McLauchlan, and J.P. Brooker. Hardware development of the yorick series of active vision systems. *Microprocessors and Microsystems*, 21(6):363–375, 1998.

[TARZ00] Harley Truong, Samir Abdallah, Sebastien Rougeaux, and Alexander Zelinsky. A novel mechanism for stereo active vision. In *Conference on Robotics and Automation (ACRA2000), Melbourne Australia*, August 2000.

[TVD+98] J.K. Tsotsos, G. Verghese, S. Dickenson, M. Jenkin, A. Jepson, E. Milios, F. Nuflo, S. Stevenson, M. Black, D. Metaxas, S. Culhane, Y. Ye, and R. Mann. Playbot: A visually-guided robot to assist physically disabled children in play. *Image and Vision Computing journal*, 16:275–292, April 1998.

[US92] Colin W. Urquhart and J. Paul Siebert. Development of a precision active stereo system. In *Proceedings of the 1992 IEEE International Symposium on Intelligent Control, Glasgow, UK*, pages 354–359, 1992.

[WFRL93] Albert J. Wavering, John C. Fiala, Karen J. Roberts, and Ronald Lumia. Triclops: a high-performance trinocular active vision system. In *Proceedings of the IEEE International Conference on Robotics and Automation, Atlanta, GA*, volume 3, pages 410–417, May 1993.

[WKSX04] Lei Wang, S.B. Kang, Heung-Yeung Shum, and G. Xu. Error analysis of pure rotation-based self-calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):275–280, February 2004.

[WM96] T. Wada and T. Matsuyama. Appearance sphere: background model for pan-tilt-zoom camera. In *Proceedings of 13th International Conference on Pattern Recognition, Vienna, Austria*, pages 718–722, 1996.

[Zha00] Zhenyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, November 2000.