コンピュータビジョン 80－13
(1992. 11. 19)

Calibrating an Active Stereo Head to Object Centered Reference Frames

James L. Crowley, LIFIA(IMAG)
Philippe Bobet, LIFIA(IMAG)
Cordelia Schmid, University of Karlsruhe

LIFIA(IMAG)
46 Ave Félix Viallet
38031 GRENOBLE, France

10 October 1992

keywords: Stereo Calibration, Active Vision, Stereo Reconstruction

ABSTRACT

This paper presents a method for using objects in a scene to define the reference frame for 3-D reconstruction. We first present a simple technique to calibrate an orthographic projection from four non-coplanar reference points. We then show how the observation of two additional known scene points can provide the complete perspective projection. When used with a known object, this technique permits a calibration of the full projective transformation matrix. For an arbitrary non-coplanar set of four points, this calibration provides an affine basis for the reconstruction of local scene structure. When the four points define three orthogonal vectors, the basis is orthogonal, with a metric defined by the lengths of the three vectors.

We demonstrate this technique for the case of a cube. We present results in which five and a half points on the cube are sufficient to compute the projective transformation for an orthogonal basis by direct o bservation (without matrix inversion). We then present experiments with three techniques for reducing the imprecision due to errors in the positions of the reference point due to pixel quantization and noise. We provide experimental measurements of the stability of the stereo reconstruction as a function of the error in the observed pixel position of the reference points used for calibration.

1 Introduction

Efforts to implement 3D vision systems have led numerous groups to confront the problem of calibrating cameras. The most widely used camera model is the "thin-lens" or pin-hole model, modeled by a perspective transformation represented in homogeneous coordinates. Reconstruction techniques tend to be extremely sensitive to the coefficients of this transformation. Of particular diff iculty are are techniques which estimate distance to scene points and then attempt to reconstruct 3D shape using the so-called "intrinsic parameters" of the camera.

The intrinsic parameters are the parameters which are independent of camera position and orientation. They are typically listed as the "center" of the image, defined by the intersection of the optical axis with the retina, and the ratio of pixel size to focal length in the horizontal and vertical directions [Tsai 87]. Reconstruction using dep th is extremely sensitive to the precision of these parameters. This has led a number of investigators to develop techniques using estimation theory based on a large number of observations of a calibration pattern [Faugeras-Toscani 86] [Skordas-Puget 90]. Such techniques typically require careful set up and rather long computation times for precise location of the reference points.

It is often overlooked that the pin-hole model is only a rough approximation for the optics of a camera. For a real camera, there are typically a continuum of values for the intrinsic parameters providing reasonable approximations to the physical system. This continuum is extremely sensitive to the setting for focus and aperture and even to small perturbations in lense mounting due to vibration! Reconstruction techniques based explicit intrinsic camera parameters are extremely sensitive to the accuracy of these parameters. It is not surprising that most current 3-D vision systems only work for carefully set up laboratory demonstrations.

The techniques presented in this paper are the result of problems that we have encountered in the construction of a real-time active vision system. Our system employs a binocular camera head mounted on a robot arm which serves as a neck. The system uses dynamically controlled vergence to fixate on objects. It is designed to track and servo on 2-D forms, to interpret such forms as objects, and to maintain a dynamically changing model of the 3D form of a scene. Focus and convergence of stereo camer as are maintained by low level reflexes. Constantly changing these parameters has posed difficult problems for 3D techniques based on classical calibration of the intrinsic camera parameters. Cumbersome and time consuming set-up means that calibration can not be performed "on the fly" as the system operates.

Mohr and his collaborators [Mohr et. al. 91]have shown that the cross ratio can be used to construct a scene based reference frame, in which the objects the scene provide the reference coordinate syste m. Such an approach abandons the use of the camera intrinsic parameters, and measures the form of objects directly in a scene based reference frame. The idea of basing the reference frame on objects in the scene leads to an approach in which a 3D vision system automatically adapts to changes in camera optics and view position.

Koenderink and Van Doorn [Koenderink-Van Doorn 89] have observed that four fiducial marks ought to be sufficient to define a scene based reference frame for structure from motion or stereo. They have attempted to "stratify" the problem into a three stage process. They define an affine "projection" from an image reference to the image based on two views of four points. They then use this affine transformation to recover the position of points in the scene. In the second phase, they apply a Euclidean metric to the resulting structure by imposing a rigidity constraint. In the final phase, a third view removes possible ambiguous interpretations. Koenderink and Van Doorn argue that the most important and the most difficult part of the problem is the first phase, and that affine transformations provide a simple solution. They develop a mathematical model for recovering 3D form in an affine basis.

Sparr has shown how arbitrary reference points can be used to define an affine basis for reconstruction [Sparr 92]. In such a coordinate frame, the shape of a collection of points is represented in terms of "affine coordinates" provided by a sub-configuration of points. In the case of each of these techniques, the objects observed in the scene provide the reference frame in which objects are reconstructed. With such a technique, an object provides its own reference frame. In the techniques described below, we will first develop an affine basis and then show that an orthogonal basis is a special case which results when the reference points form a set of orthogonal vectors.

We have found that a robust 3D vision system may be constructed using the

objects in a scene to calibrate the cameras. With thi s technique, the cameras are calibrated by fixating on any known set of 6 points. Calibration is then updated continually by tracking the image position of points as optical parameters are adjusted or as the camera is moved.

We begin by showing how an orthographic transformation matrix from affine scene coordinates to image coordinates can be obtained from the observation of four non-coplanar points. These four points define a set of three basis vectors whose lengths become the units of measure in the system. If the three vectors are orthogonal, then the reference frame provides an orthogonal basis.

We show that the orthographic projection matrix can be completed to obtain the full perspective transformation by the observation of an additional two points whose position is known relative to the first four points. This permits us to use any known object containing six identifiable points to define a reference frame for 3-D reconstruction. If the size of the object is known, then the units of reference frame can be deduced. If a manipulator is available, it can be used to provide the reference frame, yielding a simple and reliable hand-arm calibration scheme. For known scene points, the calibration matrix may be computed without matrix inversion. Finally, we show how the reference frame may be transferred to an arbitrary set of four non-coplanar points. This permits us to calibrate to a known object and then "hop" the reference frame to unkown objects.

2 Calibrating to an Object Based Reference Frame

In this section we show how a 3-D object can provide a reference coordinate system for reconstruction. The first two sections present notations and mathematics which are basic to the rest of the paper. We begin with a brief review the use of homogeneous coordinates to model perspective projection and stereo reconstruction. We show how an orthographic projection matrix can be deduced by observation of four non-coplanar points. We then show how this transformation can be completed to form the perspective transformation by the observation of two additional fiducial marks whose position is known relative to the first four points. These techniques may be used to provide an affine basis for scene reconstruction. When the vectors are orthogonal, this basis is orthogonal.

2.1 The Transformation from Scene to Image

In homogeneous coordinates, a point in the scene is expressed as a vector:

$$ {}^{s}P = [x_s, y_s, z_s, 1]^T $$

The index "s" raised in front of the letter indicates a "scene" based coordinate system for this point. The origin and scale for such coordinates are arbitrary. A point in an image is expressed as a vector:

$$ {}^{i}P = [i, j, 1]^T $$

The projection of a point in the scene to a point in the image can be approximated by a three by four homogeneous transformation matrix, ${}^{i}_{s}\mathbf{M}$. This transformation models the perspective projection with the equation:

$$ {}^{i}P\, w = {}^{i}_{s}\mathbf{M}\, {}^{s}P $$

or

$$ \begin{bmatrix} w_i \\ w_j \\ w \end{bmatrix} = {}^{i}_{s}\mathbf{M} \begin{bmatrix} x_s \\ y_s \\ z_s \\ 1 \end{bmatrix} \qquad (2.1) $$

The variable w captures the amount of "fore-shortening" which occurs for the projection of point ${}^{s}P$. This notation permits the pixel coordinates of ${}^{i}P$ to be recovered as a ratio of polynomials of ${}^{s}P$. That is

$$ i = \frac{w_i}{w} = \frac{{}^{i}_{s}M_1 \bullet {}^{s}P}{{}^{i}_{s}M_3 \bullet {}^{s}P} \qquad j = \frac{w_j}{w} = \frac{{}^{i}_{s}M_2 \bullet {}^{s}P}{{}^{i}_{s}M_3 \bullet {}^{s}P} \qquad (2.2) $$

where ${}^{i}_{s}M_1, {}^{i}_{s}M_2$, and ${}^{i}_{s}M_3$ are the first, second and third rows of the matrix ${}^{i}_{s}\mathbf{M}$, and "•" is a scalar product.

Throughout this paper, we will use a notation for homogeneous transformations in which a preceding subscript represents the source coordinate frame and a superscript represents a destination coordinate frame. For example, ${}^{i}_{s}\mathbf{M}$ may be read as the transformation from s to i. Use of this notation makes the transformation of reference systems clear.

It is common to model cameras with a pin-hole model expressed mathematically as a projective transformation. It must be stressed that this is only an approximation. Real lenses and cameras do not have a unique projection p oint, nor a unique optical axis. One way to model such errors is by adding an unknown random vector, U_M, which accounts for the difference in pixel position.

$$ {}^{i}P\, w = {}^{i}_{s}\mathbf{M}\, {}^{s}P + U_M $$

When we evaluate the precision of an approximation ${}^{i}_{s}\mathbf{M}$ we employ an estimate of the covariance of U_M. For most of the analysis that follows we will assume that the expectation of this error vector is $\mathbf{0} = (0, 0, 0)^T$

$$ E\{U_M\} = \mathbf{0} $$

2.2. Computing 3-D Structure From Stereo Correspondences

The techniques for calibration described below were mainly developed to support stereo reconstruction. Thus a natural method for evaluating these techniques is to compare reconstructed scene points with their known values. In this section we recall the solution for scene reconstruction due to Faugeras and Toscani [Faugeras et. al. 86] which uses two equations from the left image and one equation from the right image to solve for 3D position. We extend this technique to provide a solution from all four equations.

Let ${}^{L}_{s}\mathbf{M}$ and ${}^{R}_{s}\mathbf{M}$ represent the transformations for the left and right cameras in a stereo pair. Let ${}^{L}_{s}M_1, {}^{L}_{s}M_2$, and ${}^{L}_{s}M_3$ represent the first, second third rows of the ${}^{L}_{s}\mathbf{M}$, and ${}^{R}_{s}M_1, {}^{R}_{s}M_2$ and ${}^{R}_{s}M_3$ represent the first, second third rows of the ${}^{R}_{s}\mathbf{M}$. Observation of a scene point, ${}^{s}P$, gives the image points ${}^{L}P = (i_L, j_L)$ and ${}^{R}P = (i_R, j_R)$. From equation 2.2 we can write.

$$ i_L = \frac{{}^{L}_{S}M_1 \bullet {}^{S}P}{{}^{L}_{S}M_3 \bullet {}^{S}P} \qquad i_R = \frac{{}^{R}_{S}M_1 \bullet {}^{S}P}{{}^{R}_{S}M_3 \bullet {}^{S}P} $$

$$ j_L = \frac{{}^{L}_{S}M_2 \bullet {}^{S}P}{{}^{L}_{S}M_3 \bullet {}^{S}P} \qquad j_R = \frac{{}^{R}_{S}M_2 \bullet {}^{S}P}{{}^{R}_{S}M_3 \bullet {}^{S}P} $$

With a minimum of algebra, these can be rewritten as

$$ ({}^{L}_{s}M_1 \bullet {}^{s}P) - i_L ({}^{L}_{s}M_3 \bullet {}^{s}P) = 0 \quad ({}^{R}_{s}M_1 \bullet {}^{s}P) - i_R ({}^{R}_{s}M_3 \bullet {}^{s}P) = 0 \qquad (2.3) $$

$$ ({}^{L}_{s}M_2 \bullet {}^{s}P) - j_L ({}^{L}_{s}M_3 \bullet {}^{s}P) = 0 \quad ({}^{R}_{s}M_2 \bullet {}^{s}P) - j_R ({}^{R}_{s}M_3 \bullet {}^{s}P) = 0 \qquad (2.4) $$

This provides us with a set of four equations for recovering the three unknowns of ${}^{s}P$. Each equation des cribes a plane in scene coordinates that passes through a column or row of the image, as illustrated in figure 2.1. The two equations from the left image describe a two planes which form a line projecting from the pixel (i_L, j_L) to the scene points. The equation containing i_R from the right image describes a vertical plane passing through i_R. The intersection of this plane with the line from the left image is the scene point which we wish to recover.

We can solve for a 3D point with two equations from the left camera and one from the right, or equally, using one equation from the left and two

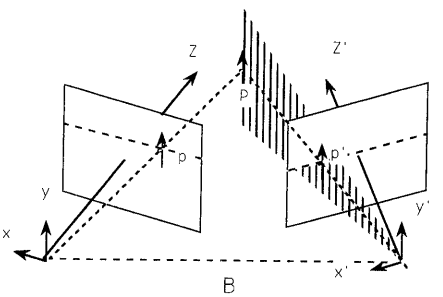

Figure 2.1 Computation of a Scene Point by Stereo Projection

from the right. When the projection matrices are exact, these points are identical. Unfortunately, because of errors in pixel position due to sampling and image noise, the projection of the rays from the left and right camera do not necessarily meet at a point.

Let us define point ${}^{s}P_{L}$ as the point obtained using two equations from the left camera and one equation from the right camera. Let us define, ${}^{s}P_{R}$, as the scene point obtained from two equations in the right and one in the left. A more precise scene point can be obtained from the mid-point of the 3-D segment which joins these two points:

$$ {}^{s}P = {}^{s}P_{L} + \frac{{}^{s}P_{L} - {}^{s}P_{R}}{2} $$

This is the technique which we will use to reconstruct 3-D points in the experiments described in sections 3 and 4 below.

Stereo reconstruction produces errors which are proportional to the distance from the origin. By placing the origin on the object to be observed, such error may be minimized. Computing the matrix ${}_{s}^{i}\mathbf{M}$ for a pair of cameras permits a very simply method to compute the position of points in the scene in a reference frame defined by the scene. Dynamically developing the transformations for the left and right images permit objects in the scene to be reconstructed independent of errors in the relative or absolute positions of the cameras.

2.3 Calibrating an Orthographic Projection from Scene Object to Image

Any four points in the scene which are not in the same plane can be used to define an affine basis. Such a basis can be used as a scene based coordinate system (or reference frame). One of the four points in this reference frame will be taken as the origin. Each of the other three points defines an axis, as shown in figure 2.2. On an arbitrary object, these axes are not necessarily orthogonal.

A simple way to exploit this idea is to use any four non-coplanar points to define an orthographic projection from an affine reference frame in the scene to the image. Let us designate a point in the scene as the origin for a reference frame. By definition,

$$ {}^{s}R_{0} = [0, 0, 0, 1]^{T} $$

Three axes for an affine object-based reference frame may be defined by designating three additional scene points as:

$$ {}^{s}R_{1} = [1, 0, 0, 1]^{T} $$
$$ {}^{s}R_{2} = [0, 1, 0, 1]^{T} $$
$$ {}^{s}R_{3} = [0, 0, 1, 1]^{T} $$

The vector from the origin to each of these points defines an axis for mea suring distance. The length of each vector defines the unit distance along that vector. These three vectors are not required to be orthogonal. The four points may be used to define an affine basis by the addition of a constraint that the sum of the coefficients be constant [Sparr 92]. We note that when the points are the corners on a right parallelpiped (a box), then they can be used to define an orthogonal basis and the additional constraint is unnecessary.

Let the symbol » represent the composition of vectors as columns in a matrix. We can then represent our affine coordinate system by the matrix ${}^{s}R$.

$$ {}^{s}R = [\, {}^{s}R_{1} \cup {}^{s}R_{2} \cup {}^{s}R_{3} \cup {}^{s}R_{0}] = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} $$

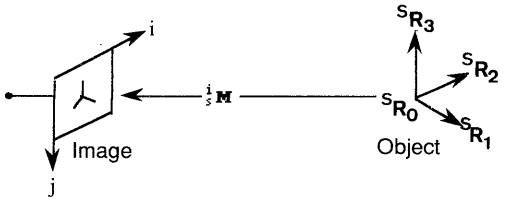

Figure 2.2 Any four points in the scene define a scene based coordinate system

The projection on these four points to the image can be written as four image points ${}^{i}P_{0}$, ${}^{i}P_{1}$, ${}^{i}P_{2}$, and ${}^{i}P_{3}$. These image points form an observation of the reference system, represented by the matrix ${}^{i}P$, where the term w_{o} has been set to 1.0.

$$ {}^{i}PW = [\, {}^{i}P_{1}w_{1} \cup {}^{i}P_{2}w_{2} \cup {}^{i}P_{3}w_{3} \cup {}^{i}P_{0}] $$
$$ = \begin{pmatrix} w_{1}i_{1} & w_{2}i_{2} & w_{3}i_{3} & i_{0} \\ w_{1}j_{1} & w_{2}j_{2} & w_{3}j_{3} & j_{0} \\ w_{1} & w_{2} & w_{3} & 1 \end{pmatrix} $$

W is a matrix whose diagonal elements are the vector $[w_{1}, w_{2}, w_{3}, 1]$. That is :

$$ W = \begin{pmatrix} w_{1} & 0 & 0 & 0 \\ 0 & w_{2} & 0 & 0 \\ 0 & 0 & w_{3} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} $$

This allows us to write a matrix expression.

$$ {}^{i}P\,W = {}_{s}^{i}\mathbf{M}\,{}^{s}R $$

The reference matrix ${}^{s}R$ has a simple inverse, which can be solved by hand.

$$ {}^{s}R^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{pmatrix} $$

Inverting this matrix allows us to write the expression:

$$ {}_{s}^{i}\mathbf{M} = ({}^{i}PW)\,{}^{s}R^{-1} = \begin{pmatrix} w_{1}i_{1} & w_{2}i_{2} & w_{3}i_{3} & i_{0} \\ w_{1}j_{1} & w_{2}j_{2} & w_{3}j_{3} & j_{0} \\ w_{1} & w_{2} & w_{3} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{pmatrix} \qquad (2.5) $$

or

$$ {}_{s}^{i}\mathbf{M} = \begin{pmatrix} w_{1}i_{1} - i_{0} & w_{2}i_{2} - i_{0} & w_{3}i_{3} - i_{0} & i_{0} \\ w_{1}j_{1} - j_{0} & w_{2}j_{2} - j_{0} & w_{3}j_{3} - j_{0} & j_{0} \\ w_{1} - 1 & w_{2} - 1 & w_{3} - 1 & 1 \end{pmatrix} \qquad (2.6) $$

Having performed the inversion of ${}^{s}R$ by hand, there is no need to compute an inverse when the system is calibrated.

The problem with equations 2.5 and 2.6 is the vector $\vec{w} = [w_{1}, w_{2}, w_{3}, 1]^{T}$. It is useful to consider the meaning of this vector. Each term "w" is a scale factor that describes the amount of "foreshortening" induced by perspecive along on of the reference vectors. The units of this forshortening are (1/ meters). Thus, if the scale facto r is defined to be 1.0 at the reference point R_{0}, then vectors emanating from reference point R_{1} will be "scaled" by a factor of w_{1}.

A simple solution is to employ the approximation $\vec{w} = [1\ 1\ 1\ 1]$, yielding an orthographic approximation to t he projective transformation. The magnitude of the error for such an approximation is proportional to the distance from the chosen origin, and inversely proportional to the focal length of the camera.

The orthographic approximation can provide a usable approximation for points near the reference object. For example, such an approximation was employed by artists before the effects of projection were discovered. In a case where the reference object is unknown, an approximate 3D reconstruction using orthogra phic projection can be constructed in terms of four non-coplanar reference points. As noted by Koenderink, such a reconstruction is qualitatively correct and may be sufficient for some applications. Koenderink (and others) presents a method to deduce the foreshorting by assuming rigidity and reconstructing from a second view point.

Alternatively, we may seek to determine the full perspective transformation by solving a set of linear equations to determine $\vec{w}$. Solving for this vector requires th ree additional constraints, or the observation of one and a half additional points whose position is known with respect to the first four points.

2.4 Obtaining the Perspective Projection by Observing a Known Object.

To obtain the perspective transformation from equation 2.6 we must solve for $\vec{w} = (w_1, w_2, w_3, 1)$. Solving for these three variables requires 3 independent equations, or the observation of the image coordinates for one and a half scene points. Let us define two known scene points as ${}^sR_4$ and ${}^sR_5$.

$$ {}^sR_4 = [x_4, y_4, z_4, 1]^T $$
$$ {}^sR_5 = [x_5, y_5, z_5, 1]^T $$

If we consider the method developped by Sparr [Sparr 92][1] , we can observe a relation which holds for four points within a plane. Let point ${}^sR_4$ be defined as the sum of two non-identical vectors ${}^sR_0{}^sR_1$ and ${}^sR_0{}^sR_3$, as shown in figure 3.1. In this case, If we observe the image position of these four points, we can write:

$$ \begin{pmatrix} i_1 & i_3 & i_4 & i_0 \\ j_1 & j_3 & j_4 & j_0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_3 \\ -w_4 \\ -1 \end{pmatrix} = 0 $$

This relation provides three equations for the three unkown values w_1, w_3 and w_4. The fact that w_4 depends on w_1 and w_3 and that the scale factors can be superimposed is easily demonstrated by considering the equation representing the third equation from this relation.

$$ w_4 = w_1 + w_3 - 1 $$

Equivalently, we can write a general relation using the positions of any reference points ${}^sR_4$ and ${}^sR_5$ provided that their position is known with respect to the four reference points. Using equation 2.6, we can write four equations with three unknowns.

$$ i_4 = \frac{{}^i_sM_1 \cdot {}^sR_4}{{}^i_sM_3 \cdot {}^sR_4} = \frac{(w_1 i_1 - i_0)x_4 + (w_2 i_2 - i_0)y_4 + (w_3 i_3 - i_0)z_4 + i_0}{(w_1 - 1)x_4 + (w_2 - 1)y_4 + (w_3 - 1)z_4 + 1} \quad (2.7) $$

$$ j_4 = \frac{{}^i_sM_2 \cdot {}^sR_4}{{}^i_sM_3 \cdot {}^sR_4} = \frac{(w_1 j_1 - j_0)x_4 + (w_2 j_2 - j_0)y_4 + (w_3 j_3 - j_0)z_4 + j_0}{(w_1 - 1)x_4 + (w_2 - 1)y_4 + (w_3 - 1)z_4 + 1} \quad (2.8) $$

$$ i_5 = \frac{{}^i_sM_1 \cdot {}^sR_5}{{}^i_sM_3 \cdot {}^sR_5} = \frac{(w_1 i_1 - i_0)x_5 + (w_2 i_2 - i_0)y_5 + (w_3 i_3 - i_0)z_5 + i_0}{(w_1 - 1)x_5 + (w_2 - 1)y_5 + (w_3 - 1)z_5 + 1} \quad (2.9) $$

$$ j_5 = \frac{{}^i_sM_2 \cdot {}^sR_5}{{}^i_sM_3 \cdot {}^sR_5} = \frac{(w_1 j_1 - j_0)x_5 + (w_2 j_2 - j_0)y_5 + (w_3 j_3 - j_0)z_5 + j_0}{(w_1 - 1)x_5 + (w_2 - 1)y_5 + (w_3 - 1)z_5 + 1} \quad (2.10) $$

Provided that no five of our six points are coplanar, these four equations can be solved to obtain the values of $\vec{w} = (w_1, w_2, w_3, 1)$. The full projection matrix, ${}^i_sM$, can then be obtained from equation 2.6.

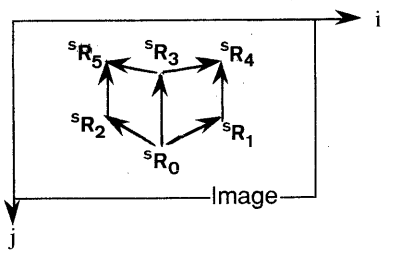

Figure 3.1 The full projective transform can be computed directly from the observation of 6 known scene points. Point ${}^sR_0$ defines the origin. Points ${}^sR_1$, ${}^sR_2$ and ${}^sR_3$ define the unit vectors of the three scene dimensions. Points ${}^sR_4$ and ${}^sR_5$ permit the full projective transformation to be recovered.

When the positions of the points ${}^sR_4$ and ${}^sR_5$ are known in advance, the solution can be structured to yield the full perspective transformation by direct observation, without matrix inversion. To illustrate this, let us consider the problem of calibrating ${}^i_sM$ by observation of 6 vertices of cube.

3 Calibration by Direct Observation of a Cube

A direct solution for calibrating the projective form of the matrix ${}^i_sM$ is possible when the reference points are known in advance. This solution can be had without matrix inversion. Let us illustrate the technique by deriving the equations for calibrating the matrix ${}^i_sM$ from the observation of 6 points on a cube.

3.1 Derivation of Solution

Consider a reference frame defined by six points on a cube, as shown in figure 3.1. Point ${}^sR_0$ defines the origin. Points ${}^sR_1$, ${}^sR_2$ and ${}^sR_3$ define the unit vectors for the X, Y and Z axes. Points ${}^sR_4$ and ${}^sR_5$ permit the full projective transformation to be recovered. Points ${}^sR_0$, ${}^sR_1$, ${}^sR_2$ and ${}^sR_3$ are defined above as:

$$ {}^sR_0 = [0, 0, 0, 1]^T $$
$$ {}^sR_1 = [1, 0, 0, 1]^T $$
$$ {}^sR_2 = [0, 1, 0, 1]^T $$
$$ {}^sR_3 = [0, 0, 1, 1]^T $$

Points ${}^sR_4$ and ${}^sR_5$ are given by:

$$ {}^sR_4 = [1, 0, 1, 1]^T $$
$$ {}^sR_5 = [0, 1, 1, 1]^T $$

Substituting ${}^sR_4$ and ${}^sR_5$ into equations 2.7 through 2.9 gives:

$$ (i_4 - i_1)w_1 + (i_4 - i_3)w_3 - (i_4 - i_0) = 0 \quad (3.1) $$
$$ (j_4 - j_1)w_1 + (j_4 - j_3)w_3 - (j_4 - j_0) = 0 \quad (3.2) $$
$$ (i_5 - i_2)w_2 + (i_5 - i_3)w_3 - (i_5 - i_0) = 0 \quad (3.3) $$
$$ (j_5 - j_2)w_2 + (j_5 - j_3)w_3 - (j_5 - j_0) = 0 \quad (3.4) $$

The coefficients w_1 and w_3 can be had fromequations 3.1 and 3.2, that is from observation of ${}^sR_4$. The coefficients w_2 and w_3 can be had from equations 3.3 and 3.4, that is from observation of ${}^sR_5$.

From the point ${}^sR_4$ we obtain:

$$ w_1 = \frac{(i_4 - i_0)(j_4 - j_3) - (i_4 - i_3)(j_4 - j_0)}{(i_4 - i_1)(j_4 - j_3) - (i_4 - i_3)(j_4 - j_1)} \quad (3.5) $$

$$ w_3 = \frac{(i_4 - i_0)(j_4 - j_1) - (i_4 - i_1)(j_4 - j_0)}{(i_4 - i_3)(j_4 - j_1) - (i_4 - i_1)(j_4 - j_3)} \quad (3.6) $$

While, from the point ${}^sR_5$ we obtain:

$$ w_2 = \frac{(i_5 - i_0)(j_5 - j_3) - (i_5 - i_3)(j_5 - j_0)}{(i_5 - i_2)(j_5 - j_3) - (i_5 - i_3)(j_5 - j_2)} \quad (3.7) $$

$$ w_3 = \frac{(i_5 - i_0)(j_5 - j_2) - (i_5 - i_2)(j_5 - j_0)}{(i_5 - i_3)(j_5 - j_2) - (i_5 - i_2)(j_5 - j_3)} \quad (3.8) $$

The fact that the equations are over-constrained poses a small problem. If the image position of points ${}^sR_4$ and ${}^sR_5$ are not perfectly measured, the resulting solution for w_1, w_2 and w_3 will not be consistent. However, it is inevitable that the position of the reference points will be corrupted by small random variations in position, if for no other reason, because of image sampling. If we simply compute w_1 and w_3 from ${}^sR_4$ and then compute w_2 from ${}^sR_5$, this inconsistency can yield an imprecise solution for the position of 3-D points. We can turn this problem to our advantage by exploiting the redundancy of the last half of a point to correct for random errors in the image position of the reference points.

3.2 Correcting for Pixel Errors in the Observed Reference Points

The classic method for minimizing the inconsistency in reference point

[1] This techniques has been pointed out by Kalle Åaström of Lund Institute of Technology

position is to compute a mean-squared solution. Faugeras and Toscani [Toscani-Faugeras 87] present a direct method to minimize the sum of the error between the projection of calibration points and and their observation. From equation 2.3, for each calibration point ${}^{s}R_k$ and its image projection ${}^{i}P_k$, we can write:

$$({}^{i}_{s}M_1 \bullet {}^{s}R_k) - i_k ({}^{i}_{s}M_3 \bullet {}^{s}R_k) = 0 \qquad ({}^{i}_{s}M_2 \bullet {}^{s}R_k) - j_k ({}^{i}_{s}M_3 \bullet {}^{s}R_k) = 0$$

For N non-coplanar calibration points we can write a linear system of 2N equations of the form:

$$\mathbf{A}\ {}^{i}_{s}\mathbf{M} = 0.$$

where the rank of **A** is 11. The problem is to find a matrix ${}^{i}_{s}M$ which best minimizes a criterion equation

$$\mathbf{C} = \| \mathbf{A}\ {}^{i}_{s}\mathbf{M} \|$$

We use Lagrange multipliers to obtain a least squares value for ${}^{i}_{s}\mathbf{M}$ which minimizes **C**. We will refer to this below as the "mean square technique", denoted "msq" in the tables of experimental results below.

As an alternative, it is possible to obtain a direct solution by constraining one of the points ${}^{s}R_4$ or ${}^{s}R_5$ to be consistent with the other five points. For example, we can use the value of w_3 computed from ${}^{s}R_4$ to compute a correction for ${}^{s}R_5$. We can then use the corrected value to compute w_2.

To illustrate, let us compute a correct i_5^* for the value of i_5. From equation 3.8, we can write:

$$w_3 = \frac{(i_5^* - i_0)(j_5 - j_2) - (i_5^* - i_2)(j_5 - j_0)}{(i_5^* - i_3)(j_5 - j_2) - (i_5^* - i_2)(j_5 - j_3)}$$

Solving for i_5^* gives:

$$i_5^* = \frac{i_0(j_5 - j_2) - i_2(j_5 - j_0) - i_3(j_5 - j_2)w_3 + i_2(j_5 - j_3)w_3}{-j_2 + j_0 + j_2 w_3 - j_3 w_3} \qquad (3.9)$$

we then use this value in equation 3.7.

In the following two sections we will compare the precision obtained from direct solution using 5 and 1/2 points to precision obtained from each of these two techniques for exploiting the redundancy in the sixth point. We will first use artificial data to explore the sensitivity of these three techniques to the standard deviation of the error in pixel position as well as the sensitivity to the size of the reference object in the image. We will show results with real data measured on a cube and on sugar box.

3.3 An Example of a Calculation

In this section we present an example of the calibration using a aluminium cube with a side of 20cm. This example illustrates the method used in the experiments in the following sections.

In our experiments, images of the cube are projected on the work-station screen and the pixel coordinates of the vertices ${}^{i}P_0$, ${}^{i}P_1$, ${}^{i}P_2$, ${}^{i}P_3$, ${}^{i}P_4$, and ${}^{i}P_5$ were selected with the mouse. The image size is 512 by 512 pixels. A standard left handed image coordinate system is used in which the origin is the upper left hand corner, positive i (columns) is to the left, and positive j (rows) is down. Reference points were indicated by pointing with a mouse, a technique which can sometimes result in an error of one or two pixels.

The images used for this example are shown in figure 3.2.

For the left image, the vertices of the cube were detected at:

${}^{L}P_0 = (228, 481)$ ${}^{L}P_1 = (347, 351)$ ${}^{L}P_2 = (77, 374)$
${}^{L}P_3 = (229, 223)$ ${}^{L}P_4 = (354, 107)$ ${}^{L}P_5 = (69, 125)$

Equations 3.5 through 3.7 give a solution for wDBA8()[4]: of:

$$\vec{w} = (0.917610, 0.858158, 1.052614, 1)$$

By the direct method we then obtain

$${}^{L}_{S}\mathbf{M} = \begin{pmatrix} 147.589396 & -146.422112 & -11.048572 & 228.000000 \\ -101.081043 & -84.764543 & -269.732889 & 481.000000 \\ 0.082390 & 0.059453 & -0.052614 & 1.000000 \end{pmatrix}$$

Correcting by computing i_5^* with equation 3.9 gives a corrected solution for the matrix as:

$${}^{L}_{S}\mathbf{M} = \begin{pmatrix} 147.589396 & -146.622393 & -11.048572 & 228.000000 \\ -101.081043 & -85.737338 & -269.732889 & 481.000000 \\ 0.082390 & 0.056852 & -0.052614 & 1.000000 \end{pmatrix}$$

Using the least squares technique, the matrix for the left image ${}^{L}_{s}\mathbf{M}$ is computed as:

$${}^{L}_{S}\mathbf{M} = \begin{pmatrix} 148.016122 & -146.716244 & -12.239302 & 228.149911 \\ -100.417731 & -85.159763 & -270.607106 & 481.003325 \\ 0.084301 & 0.058403 & -0.056504 & 1.000000 \end{pmatrix}$$

For the right image, the vertices of the cube were detected at:

${}^{R}P_0 = (212, 464)$ ${}^{R}P_1 = (343, 332)$ ${}^{R}P_2 = (74, 360)$
${}^{R}P_3 = (197, 208)$ ${}^{R}P_4 = (337, 88)$ ${}^{R}P_5 = (52, 116)$

With the direct method, from equation 3.5 through 3.7 we obtain

$${}^{R}_{S}\mathbf{M} = \begin{pmatrix} 158.141055 & -132.711116 & -26.996216 & 212.000000 \\ -105.729358 & -78.270296 & -268.666055 & 464.000000 \\ 0.079128 & 0.071471 & -0.060894 & 1.000000 \end{pmatrix}$$

Correcting by computing i_5^* with equation 3.9 gives a corrected solution for the matrix as:

$${}^{R}_{S}\mathbf{M} = \begin{pmatrix} 158.141055 & -132.661599 & -26.996216 & 212.000000 \\ -105.729358 & -78.029403 & -268.666055 & 464.000000 \\ 0.079128 & 0.072141 & -0.060894 & 1.000000 \end{pmatrix}$$

Using the least squares technique, the matrix for the right image ${}^{R}_{s}\mathbf{M}$ is computed as:

$${}^{R}_{S}\mathbf{M} = \begin{pmatrix} 158.066763 & -132.620333 & -26.745194 & 211.958839 \\ -105.863649 & -78.136621 & -268.493161 & 464.002612 \\ 0.078734 & 0.071856 & -0.060038 & 1.000000 \end{pmatrix}$$

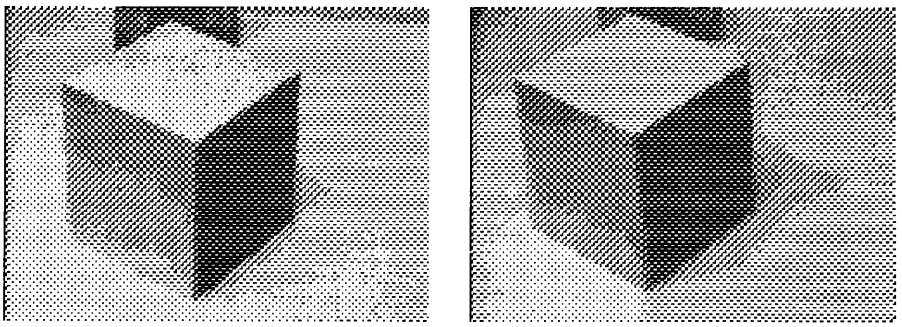

Figure 3.2 Stereo Images of a 20 cm Calibration Cube at a distance of 1.2 meters.

As a check, we indicated the image positions of the point ${}^{s}\mathbf{R}_6 = [1, 1, 1, 1]^T$ and construct the 3D position o f this point by a stereo solution. Clicking on the corner corresponding to point 6 in the left and right images gives:

$${}^{L}\mathbf{P}_6 = (200, 23) \qquad {}^{R}\mathbf{P}_6 = (193, 11)$$

Solving for the 3-D position with the stereo technique using all four equations as described above gives

method	X	Y	Z	Dist
direct	1.004628	1.005564	0.997816	0.007559
corrected	1.014828	1.016062	0.987449	0.025206
msq	0.992905	0.993915	1.004042	0.010183

These reconstructed points are expressed in units defined by the side of the cube. One multi plies by 20 to obtain centimeters. We can observe that for this example, the direct calculation gives an error of about 0.7%, while the matrices which use a correction based on i_5^* , has about 2.5% error. The mean square technique gives about 1% error. The error is due to both the sampling interval of the pixels and imprecision of our mouse clicks. Although the direct solution happened to perform best in this example, we will see in the experiments presented in the next section that the error is a random fu nction. The mean square technique tends to give an error with the lowest average value. Such a tendency is made evident by a systematic exploration of the precision of the three techniques.

3.4 Experiments with Sensitivity to Pixel Noise

In order to measure the precision of the recovered projection matrix ${}_{s}^{i}\mathbf{M}$, we define an ideal projection matrix, ${}_{s}^{i}\mathbf{M}^*$. We compute a corrupted observation of the calibration points by projecting the calibration points with the ideal matrix ${}_{s}^{i}\mathbf{M}^*$ and adding a random Gaussian variable, U_M , with a known standard deviation.

$${}^{i}\mathbf{P}_k w = {}_{s}^{i}\mathbf{M}^* \, {}^{s}\mathbf{R}_k + U_M$$

We then use the corrupted points ${}^{i}\mathbf{P}_k$ to solve for ${}_{s}^{i}\mathbf{M}$. To measure the sensitivity of the solution to the pixel position of the calibration points, we choose a scene point and compute

$$E_k = \| {}_{s}^{i}\mathbf{M}^* \, {}^{s}\mathbf{P}_k - {}_{s}^{i}\mathbf{M} \, {}^{s}\mathbf{P}_k \|.$$

In our first experiment, we compute the average value for this measure as a function of the standard deviation of the pixel noise, U_M, for the six points used for calibration. The row labeled "direct" is a direct solution ${}_{s}^{i}\mathbf{M}$ using 5 and 1/2 points. The line labeled "corrected" uses the correction of the point i_5^* as described above. The line "msq" is computed using the least squares method. These average of the error for the six calibration points is presented in table 3.1.

M*P - MP	0.0125	0.250	0.500	1.0	2.0	4.0
direct	0.0424	0.0847	0.1695	0.3388	0.6769	1.3522
corrected	0.0406	0.0813	0.1627	0.3253	0.6503	1.2998
msq	0.0209	0.0418	0.0837	0.1675	0.3355	0.6771

Table 3.1 Average error E (in pixels)for calibration points as a function of standard deviation of pixel error.

As a second test, we computed the same measure for the scene point (1, 1, 1) as a function of the standard deviation of pixel noise as shown in table 3.2. In both experiments we can observe that the magnitude of the error in the projection is proportional to the standard deviation of the Gaussian noise added to the pixels from which the matrix ${}_{s}^{i}\mathbf{M}$ is derived. We can also observe that the direct solution and the correction technique give similar values, while a mean square solution give a systematically better solution. For the average error of the calibration points, mean-square is nearly twice as precise. For the cube vertex point (1,1,1) the mean square gives about 2/3 of the error of the other two techniques.

M*P - MP	0.0125	0.250	0.500	1.0	2.0	4.0
direct	0.2133	0.4267	0.8535	1.7069	3.4141	6.8323
corrected	0.2270	0.4541	0.9082	1.8163	3.6328	7.2675
msq	0.1715	0.3429	0.6853	1.3682	2.7272	5.4302

Table 3.2 Average error E (in pixels)for the point (1,1,1) as a function of standard deviation of pixel error of calibration points.

Another measure of precision is to model a stereo pair of cameras and then compute a stereo solution using the corrupted projection matrices to recover the 3-D position of known scene points. To perform such an experiment, we simulated our nominal experimental set up composed a pair of cameras with a base line of 20cm, a focal length of 25mm and images with 512 x 512 pixels. The cameras are simulated to be looking at a cube 20 cm on each side at a distance of 1.2 meters. Three dimensional points were computed using all four stereo equations, as presented in the second technique in section 2.2.

Table 3.3 shows the average error of reconstruction for the six calibration points as a function of the pixel noise. The units are measured in units of the side of a cube. To obtain distance in cm one multiplies by 20. To obtain pe rcentage, one multiplies by 100.

3-D Dist	0.0125	0.250	0.500	1.0	2.0	4.0
direct	0.0027	0.0055	0.0111	0.0224	0.0450	0.0918
corrected	0.0029	0.0058	0.0116	0.0233	0.0469	0.0955
msq	0.0019	0.0038	0.0076	0.0152	0.0303	0.0609

Table 3.3 Average 3-D distancefor calibration points as a function of standard deviation of pixel error.

As a second test, we compute the same measure for the scene point (1, 1, 1), as shown in table 3.4. We can notice that once again the direct and corrected method give very similar results, and that the mean-square technique is more than twice as precise.

distance	0.0125	0.250	0.500	1.0	2.0	4.0
direct	0.0193	0.0388	0.0786	0.1613	0.3397	0.8395
corrected	0.0192	0.0386	0.0780	0.1596	0.3350	0.8264
msq	0.0092	0.0183	0.0364	0.0718	0.1395	0.3031

Table 3.4 3-D error for scene point (1,1,1) as a function of standard deviation of pixel error of calibration points.

Another question which one might ask is, what is the influence of the size of the cube in the image on the error of reconstruction. Equation 2.6 shows that the coefficients are calculated from the lengths of the vectors in the image. Thus, the larger the distance between the image of the calibration points, the less sensitive the coefficients are to an error in image position. None-the-less, one should ask: how sensitive is the 3-D reconstruction to the length of this vector?

Using a simulated cube, and the mean square correction method, we computed calibration matrices for a 20cm cube at distances of 100 cm to 200 c m in steps of 20 cm. At 100 cm, the cube fills the image. At 200 cm the cube is the size of a quarter of the image. For each pair of calibration matrices, we computed the stereo solutions for image projects at scene points (1,1,1). We used calibration matrices computed from pixel positions corrupted by Gaussian noise of standard deviation 0.125, 0.25, 0.5, 1, 2, 4 and 8. For each point we performed a stereo reconstruction 100 times and computed the average error (table 3.5) and the maximum error (table 3.6). The stereo solutions are computed, as above, using all four equations.

At a distance of 100 cm, the sides of the cube project to vectors of nearly the entire image. Interesting, in table 3.5, we see that in this case, the percentage of error in reconstruction is almost exactly proportional to the standard deviation of the pixel noise. That is, for a pixel error of 0.5 pixels the reconstruction error is 0.53%, for a pixel error of 1.0 the reconstruction error is 1.07%. The error percentages doubles when the cube occupies half the image at 140 cm, and double again when the cube reaches a quarter of the image at 200 cm.

dist	0.125	0.25	0.50	1.00	2.00	4.00	8.00
100	0.0013	0.0026	0.0053	0.0107	0.0214	0.0428	0.0864
120	0.0019	0.0038	0.0076	0.0152	0.0303	0.0609	0.1244
140	0.0025	0.0051	0.0102	0.0204	0.0410	0.0826	0.1727
160	0.0033	0.0066	0.0132	0.0265	0.0532	0.1082	0.2476
180	0.0041	0.0083	0.0167	0.0335	0.0675	0.1440	0.4539
200	0.0051	0.0103	0.0206	0.0412	0.0832	0.1745	0.6485

Table 3.5 The average 3-D error as a function of distance of the calibr ation cube from the camera (rows) and as a function of pixel noise (columns). Errors are expressed in units of the length of the side of the calibration cube (20cm). Projection was computed using the mean square technique. Scene points were computed using all four stereo equations.

The maximum errors for the same 100 runs are shown in table 3.6.

An interesting effect was observed during the experiment with the maximum error. A sort of singularity was detected for noise of standard deviation 6 pixels when the cube was 200 cm from the camera. For such noise, the maximum reconstruction error reached as high as 318.839 times the side of the cube, or 6.367 meters! As can be seen from the table, the error dropped as we passed beyond this (supposed) singularity. We have no explanation for this effect.

dist	0.125	0.25	0.50	1.00	2.00	4.00	8.00
100	0.0065	0.0130	0.0260	0.0515	0.1008	0.2042	0.4745
120	0.0092	0.0183	0.0364	0.0718	0.1395	0.3031	0.7649
140	0.0123	0.0245	0.0485	0.0953	0.1893	0.4345	1.9043
160	0.0158	0.0315	0.0623	0.1216	0.2529	0.6112	6.3583
180	0.0198	0.0394	0.0776	0.1505	0.3306	3.5540	20.8398
200	0.0243	0.0481	0.0945	0.1860	0.4257	1.8469	42.2503

Table 3.6 The maximum reconstruction error, due to pixel noise, for the corners of the cube. Calibration was computed using the mean square correction (msq). Rows indicate distance of the cube from the camera (rows). Columns give standard deviation of pixel po sition noise. Errors are expressed as a fraction of the length of the side of the calibration cube (20cm).

3.5 Experimental Precision with Real Images

In our active vision system we dynamically calibrate our stereo cameras by observation of an aluminium cube with a size of 20cm. In order to perform an evaluation of our technique, we set up an interactive program in which points in images are indicated by pointing with a mouse. This experiment was performed with live images produced by a Pulnix TM 560 camera equipped with a Cosmicar 25 mm F1.8 lens fed CCI R video signals to an Imaging Technologies FG100 Digitizer. Images were acquired with a resolution of 512 by 512.

Our 20 cm aluminium cube was painted such that two of its faces are white, two are gray and two are black. The cube was placed on a white table-cloth with a black face to the left, gray face to the right and the white face up. Stereo images of this cube from a distance of 1.2 meters were shown in figure 3.2. These images are the first pair from a sequence which are used in chapter 5 below.

Using the image in figure 3.2, we computed calibration matrices for the left and right cameras. We then computed a stereo reconstruction for the point (1, 1, 1) after having corrected using the three techniques described above (direct, corrected and msq). The results are presented in table 3.5. Naturally, the direct method reconstructs each corner at its exact position. The corrected method involved correcting point P5 to obtain a coherent matrix. Thus, reconstruction with the real P_5 yields a small error. The msq error distributes this error over all of the points.

Point	Real Position	direct	corrected	msq
P0	(0, 0, 0)	0.0000	0.0000	0.0208
P1	(1, 0, 0)	0.0000	0.0000	0.0187
P2	(1, 0, 0)	0.0000	0.0000	0.0034
P3	(1, 0, 0)	0.0000	0.0000	0.0198
P4	(1, 0, 0)	0.0000	0.0000	0.0180
P5	(1, 0, 0)	0.0000	0.2422	0.0033

Table 3.7 Distance between the real and reconstructed 3-D positions for the six calibration points using the three techniques. All units are in terms of the side of the cube.

We then placed a box of sugar next to the calibration cube, as shown in figure 2.6 and reconstructed the corners of the box using the matrices determined by the three techniques. The six visible corners of the sugar box are listed as points S_0 through S_5. The 3-D error, measured as a percentage of the side of the cube, are shown for each of the 6 points.

Point	Real Position	direct	corrected	msq
S_0	(0, –0.325, 0)	**0.0175**	0.0320	0.0185
S_1	(4.75, –0.325, 0)	0.1065	0.1327	**0.0948**
S_2	(0, 0, 0)	0.0131	**0.0130**	0.0523
S_3	(0, –0.325, 0.95)	**0.0240**	0.0310	0.0549
S_4	(0.475, –0.325, 0.95)	0.1082	0.0762	**0.0372**
S_5	(0, 0, 0.95)	0.0750	0.0762	**0.0549**

Table 3.6 Errors for reconstructed corners of sugar box using three techniques. All distances are in units defined by the side of the calibration cube (20cm). The most precise values are indicated in bold.

The first thing that we can observe is that no one technique produces the best result for all six corner s. The mean square solution produces the smallest error for three of the corners, the direct method for two of the corners, and correcting the 5th point in one of the corners. The largest error was on the order of 13% for the the corrected method, while the smallest was on the order of 1% for the direct method. None-the-less, our conclusion from these and many other experiments is that computing the calibration matrix using the mean-square technique gives a slight improvement in precision at a slight increase in computational cost. The direct method provides a 3D solution which is less precise but easier to program.

4 Transferring the Scene Coordinates to a New Reference Frame.

In a continuously operating vision system, it is convenient to be able to transfer the reference frame to any object in the scene. The simplest way to do this is to empty the 3-D model, chose a new reference object and then reconstruct the model with respect to this new object. However, in some cases it may be desirable to preserve the 3D model, and transform it to the new reference frame. This section concerns a method to obtain the four by four homogeneous coordinate expression for a transformation of the 3D model to a new reference frame. This technique may be used when both the new and the old reference objects are simultaneously in the field of view of the two stereo cameras. This transformation makes it possible to "hop" the coordinate system of the 3D model from one object in the field of view to another.

The projective transformation has the form of a 3 by 4 homogeneous matrix. The 3 dimensional side produces points in image coordinates while the 4 dimensional side refers to scene coordinates. A four by four correction matrix provides a transformation to the scene based reference frame. Such a transformation may be used to change the scene based reference frame.

Suppose that we have calibrated to a known reference object, and that we now wish to transfer the coordinates to a new, perhaps unknown object. Such a transformati on may be achieved using the observed scene position of four points. That is, any four points in the scene can be used to define a new affine reference frame, without re-calibration. In particular, it is possible to "hop" the reference frame from the original calibration object to a sequence of other objects reconstructed by stereo.

Let us designate the the original reference frame as "O" and the new reference frame as "S". To transform our reference, we need the homogeneous transformation from O to S: ${}_{O}^{S}\mathbf{T}$. Let ${}_{O}^{1}\mathbf{M}$ and ${}_{O}^{8}\mathbf{M}$ be the results of calibrating to the reference frame "O" as described in section 2 above. We note that post-multiplying our calibration matrices by a homogeneous matrix ${}_{S}^{O}\mathbf{T}$ transforms the calibration matrices to the new reference frame.

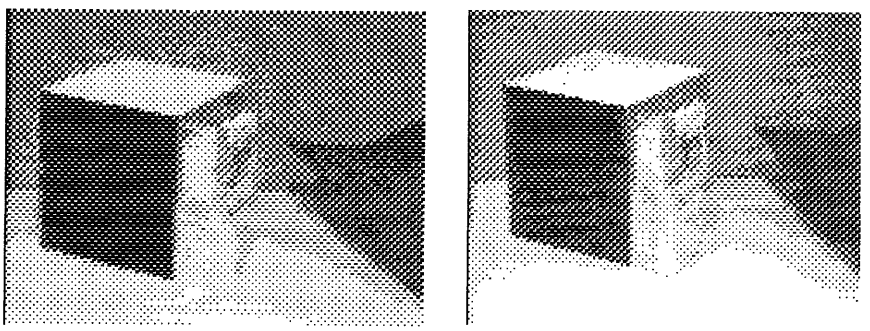

Figure 3.3 Left and right images of cube and sugar box used for table 3.6.

$${}^{L}_{s}\mathbf{M} = {}^{L}_{o}\mathbf{M}\ {}^{o}_{s}\mathbf{T} \qquad \text{and} \qquad {}^{R}_{s}\mathbf{M} = {}^{R}_{o}\mathbf{M}\ {}^{o}_{s}\mathbf{T}$$

Let us call these four reference points expressed in the original coordinate system: ${}^{o}R_0$, ${}^{o}R_1$, ${}^{o}R_2$ and ${}^{o}R_3$. To obtain ${}^{o}_{s}\mathbf{T}$ we compose a matrix with the 3D position of these reference points.

$${}^{o}R = [\ {}^{o}R_1 \cup {}^{o}R_2 \cup {}^{o}R_3 \cup {}^{o}R_0\ ]$$

We note that in the new reference frame, these same points will represent the origin and the three unit vectors, represented by

$${}^{s}R = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

To calculate ${}^{o}_{s}\mathbf{T}$ we write

$${}^{o}R = {}^{o}_{s}\mathbf{T}\ {}^{s}R$$

and then note that ${}^{s}R$ has a trivial inverse (as in chapter 2.) Thus ${}^{o}_{s}\mathbf{T}$ is given by

$${}^{o}_{s}\mathbf{T} = {}^{o}R\ {}^{s}R^{-1} = {}^{o}R \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} x_1 - x_0 & x_2 - x_0 & x_3 - x_0 & x_0 \\ y_1 - y_0 & y_2 - y_0 & y_3 - y_0 & y_0 \\ z_1 - z_0 & z_2 - z_0 & z_3 - z_0 & z_0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad (4.1)$$

Thus the affine reference frame can be moved to any set of four points whose position is known (or observed) in the original reference frame by a direct formula.

5 Discussion and Conclusions

The reliable operation of a 3D vision system depends on accurate calibration. Calibration procedures which require time consuming and cumbersome set-up a re of little use when the optical parameters of the lenses are continually changing. In this paper we have presented the foundations for a technique in which camera calibration is determined and maintained using objects in the scene. These techniques permit objects in the scene to serve as the reference frame in which the scene is reconstructed. Because the object is reconstructed in its own reference frame, information about the shape of an object can be registered and fused without knowledge of the camera positions relative to the object.

After some definitions, we presented a technique in which six points in the scene, for which at most four are in the same plane, can be used to compute the transformation from scene to image. While most of the experimental results have been presented using a cube, the same techniques can be used with any object for which the six points can be unambiguously determined. When the reference points are taken from the corners of a right parallelpiped (such as a box), the resul ting basis is orthogonal. Otherwise, the result is an affine 3D basis. In either case, the coordinate system is invariant to camera viewing angle and may be used for a reconstruction which is intrinsic to the shape of the reference object.

Finally we have shown how tracking four points for which the scene position is known can be used to hop the coordinate reference frame from one object to another. We have also shown how four reconstructed points can be used to keep the reference frame locked onto an object as the head (or object) is moved.

This paper is concerned with the mathematics of calibration and reconstruction. The precision of stereo reconstruction is very sensitive to the precision of the image location of the points on which calibration is based. This dependence is highly non-linear. We can draw several conclusions from this dependence.

The precision of reconstruction is very dependent on the focal length of lenses and the size of images. Many of the experiments for this paper were initially performed with 256 x 256 images, formed by digitizing only the even lines from the image. This was done to avoid interlacing problems during motion. Changing to 512 x 512 images (with the same lenses) improved the 3-D precision by well over a factor of 4.A reduction in the size of image must be accompanied by a increase in the focal length of the lenses, in order to maintain a similar reconstruction precision.

In this paper, we have not addressed the problem of locating the reference points. Yet the cri tical dependence of 3-D precision on image location shows that this is a fundamental problem for which a satisfactory solution still does not exist. What we can offer to this problem is the criteria for evaluating image analysis techniques for 3D reconstruction. The mathematics of reconstruction show that an image description algorithm must locate image features with both high precision and with stability. Advances in precision have generally resulted in reduction in stability. Real time active vision requires both.

A final conclusion involves calibration. The current wisdom argues for an initial calibration phase using a complex set up involving many reference points. The argument is that additional reference points permit improvement in precision through use of statistical methods. In a continuously operating vision system, calibration matrices must be continuously corrected for effects due to focus, aperture, vergence and camera zoom, as well as vibrations that can change the lense mounting. Thus, a more precise reconstruction of the scene requires continually updating the calibration.

Acknowledgements

The ideas developed in this paper first evolved in discussion with a number of people. The philosophical contribution of Mohr is sufficiently important that we have even offered to name him as a co-author. Listening to presentations by Jan Koenderinck and by Gunnar Sparr has also provided key inspirations. Without Sparr's explanations concerning affine spaces, we would have believed that the techniques only applies to box-shaped objects. Discussion with Thierry Vieville have also helped us to organise and better understand certain aspects of the technique.

Bibliography

[Anon 93] "Maintaining Stereo Calibration by Tracking Image Points", Companion paper submitted to ICCV-93.

[Ayache 89] Ayache, N. "Construction et Fusion de Représentations Visuelles 3D", Thèse de Doctorat d'Etat, Universtié Paris-Sud, centre d'Orsay, 1988

[Crowley 91] Crowley, J. L. "Towards Continuously Operating Integrated Vision Systems for Robotics Applications", SCIA-91, Seventh Scandinavian Conference on Image Analysis, Aalborg, DK, August 91.

[Faugeras-Toscani 86] Faugeras, O. D. ,G. Toscani, "The Calibration Problem for Stereo. Computer Vision and Pattern Recognition, pp 15-20, Miami Beach, Florida, USA, June 1986.

[Koenderink-Van Doorn 89] J. Koenderink and A. J. Van Doorn, "Affine Structure from Motion", Technical Report, Universtiy of Utrecht, Oct. 1989.

[Mohr et al. 91] R. Mohr, L. Morin and E. Grosso, "Relative Positioning with Poorly Calibrated Cameras", LIFIA-IMAG Technical Report RT 64, April 1991.

[Mohr et al. 92] R. Mohr, L. Morin and E. Grosso, "An Egomotion Algorithm Based on the ...", The Second European Conference on Computer Vision (ECCV-2), St. Margherita, Italy, May 1992.

[Puget-Skordas 90] P. Puget et T. Skordas, "Calibrating a Mobile Camera", Image and Vision Computing, Vol 8 No. 4, November 1990.

[Sparr 92] G. Sparr, "Depth Computations from Polyhedral Images", The Second European Conference on Computer Vision (ECCV-2), St. Margherita, Italy, May 1992.

[Tsai 87] R. Y. Tsai, "A Versatile Camera Calibration Technique for High Accuracy 3D Machine Vision Metrology Using off the Shelf TV Cameras and Lenses", IEEE Journal of Robotics and Automation, Vol 3 No. 4, August 1987.