Stereo Without Camera Models

Richard Hartley
Rajiv Gupta
Tom Chang

GE Corporate R&D
1 River Road, P.O. Box 8
Schenectady, NY 12301

Stereo Without Camera Models

1 Introduction

A typical system for the construction of 3-D models from stereo imagery operates in three phases. In the first phase a set of *match points* (i.e., pixels in the two views that are the images of the same point in the real world, also referred to as tie points), are established between the two images. In order for the matching procedure to succeed, several restrictions are placed on the imagery, principally to ensure that the corresponding areas and features in the two images are nearly identical. Without these restrictions, most area and feature based correlation techniques perform poorly while attempting to determine match points between images. In the second phase, the computed match points are used to derive the relative locations, orientations and other parameters of the cameras. This process usually requires iterative solution of a set of non-linear equations. With the information about the cameras known, one can analyze the disparity arising because of different elevations of various points and assign them a relative 3-D coordinate. In a third phase the actual locations of points are computed.

This paper describes a methodology for epipolar matching and stereo information extraction from a set of two or more images of a scene without placing any restrictions on the imagery, and at the same time, avoiding any assumptions about the camera model. We allow image pairs which may be only partially overlapping, scaled, rotated, taken from oblique viewing angles, or otherwise transformed with respect to each other. These factors generally result in imagery in which corresponding areas in different views do not look identical.

1.1 Camera Model

The general model of a perspective camera that will be used here is that represented by an arbitrary 3×4 matrix, P, known as the *camera matrix*. The camera matrix transforms points in 3-dimensional projective space to points in 2-dimensional projective space according to the equation

$$\tilde{u} = P\tilde{x}$$

where $\tilde{u} = (u, v, w)^T$ and $\tilde{x} = (x, y, z, 1)^T$. The camera matrix P is defined up to a scale factor only, and hence has 11 independent entries. This was the representation of the imaging process considered by Strat [**?**]. As shown by Strat, this model allows for the modeling of several parameters, in particular,

1. the location and orientation of the camera,

2. the principal point offsets in the image space, and

3. unequal scale factors in two directions parallel to the axes in image space.

This accounts for 10 of the total 11 entries in the camera matrix. It may be seen that if unequal stretches in two directions **not** aligned with the image axes are allowed, then further 11-th camera parameter may be defined. Thus, the imaging model considered here is quite general. In practical cases, the focal length (magnification) of the camera may not be known, and neither may be the principal point offsets. Strat [**?**] gives an example of an image where the camera parameters take on surprising values. Our purpose in treating general camera transforms is to avoid the necessity for arbitrary assumptions about the image.

The matrix P representing a general camera transformation may be factored into a product $P = KL$, where L is a 3×4 matrix representing the so-called "external parameters" of the camera and K is a 3×3 upper-triangular matrix representing so-called "internal parameters". The task of determining the internal parameters of the camera is known as "calibration". The matrix L of external parameters represents a simple pinhole camera model. If the camera is located at a point T with orientation represented by the rotation matrix R, then a simple pinhole camera model will take a point $(x, y, z, 1)^T$ to $(u, v, w)^T = R*(X-T)$. This is represented in matrix form as $(u, v, w)^T = (R|-RT)(x, y, z, 1)^T$. Matrix L is the matrix $(R| - RT)$. In general, therefore, the camera matrix will be written in the form $(M| - MT)$ where M is a non-singular 3×3 matrix and T is a vector representing the location of the camera.

In the usual approach to stereo, camera parameters for the two cameras involved are derived from a set of match points in the two images. Longuet-Higgins [**?**] gave a method for computing the relative locations and orientations of the cameras in the case where the internal parameters are all known. (More precisely, [**?**] assumes a mapping from object space coordinates to image space coordinates, which, if the internal camera model is known, reduces to the simple pinhole model.)

Out interest is in the case where the internal camera parameters are unknown. In this case, it is impossible to compute the camera model from a set of match points only. We prove there is an ambiguity represented by a general 3-dimensional projective transform of the cameras and the object-space points corresponding to the match points. The ambiguity may be resolved by the use of ground control points, or by placing restrictions on the camera model.

Previously known methods for solving the camera calibration and placement to take proper account of both ground-control points and image correspondences are unsatisfactory

in requiring either iterative methods or model restrictions. Purely non-iterative methods (e.g. those by Sutherland [**?**] or Longuet-Higgins [**?**]) are not able to handle ground-control and match-points simultaneously. In our approach, we avoid the actual computation of internal or external camera parameters as far as possible. In fact the geometry of the object-space points is determined without the need for the camera parameters to be computed, though they are easily obtained if needed. In fact, the method given in section 3 of this paper provides a non-iterative method for solving camera parameters given matched points and ground control points.

1.2 Overview

In order to handle image pairs that may be locally distorted with respect to each other, an image to image transformation, designated by M_I, is used. For any given pixel $\tilde{u} = [u, v, 1]^T$, which is the image of a point P in the first image, M_I computes the vicinity in which the image of P would lie in the second image. This accelerates the search for match points. In addition, since areas in one image are transformed via M_I before comparing them with the areas in the second image, M_I also has the effect of undoing local distortions and registering the images with respect to each other, making feasible area-based correlation for finding the match points. As more and more match points are found, M_I can be made better and better in a bootstrapped fashion.

It has been known for some time that for two match points $\tilde{u}$ and $\tilde{u}'$, expressed in homogeneous coordinates, there exists a 3×3 matrix Q, also known as the essential matrix, such that $\tilde{u}'^T Q \tilde{u} = 0$ [**?**, **?**]. As shown in [**?**], $[r, s, t]^T = Q\tilde{u}$ is the equation of the epipolar line corresponding to $\tilde{u}$, in the second image. (The line $[r, s, t]^T$ in homogeneous coordinates corresponds to the line equation $ru + sv + t = 0$, in the image-space.)

A key contribution of this paper is to show that M_I and Q can be computed in conjunction thus making M_I respect the epipolar constraint. In other words, not only is the transformed point $M_I\tilde{u}$ close to its match point in the second image, it also lies on the epipolar line.

It should be emphasized that the epipolar constraint is enforced without computing the relative camera location, orientation, or any other camera parameters such as scale or principal point offset in the camera model. In fact, in [**?**] it is shown that the transformation Q contains all the information about relative camera parameters for *completely uncalibrated cameras* (i.e., cameras about which nothing is known) that can be derived from a set of match points.

Once a sufficient number of match points have been found, the analysis of their relative disparities to compute their relative 3-D location can start using the transformation Q which contains the relative information about the cameras. Unfortunately, as shown in Lemma 2, from a given set of correspondences between the two images, and the Q-matrix, it is *impossible* to determine uniquely all the camera parameters and the position of points in object-space that are compatible with the given data. However, we prove (see Theorem 1) that the various solutions (i.e., the 3-D location of points and the camera transformations matrices) that are compatible with the given set of match points are related with each other via a 3-dimensional projective transformation H. We show how one can compute two camera transformations P_1 and P_2, which are obviously not unique, from Q and use them to find a set of points locations in 3-D. Since both P_1 and P_2, and the set of points may be off by an unknown projective transformation H, ground control points are used to compute the absolute 3-D location of the points.

The techniques presented in this paper have been implemented and tested by augmenting the STEREOSYS testbed developed by Marsha Jo Hanna of SRI [**?**, **?**]. Our experiments reveal that these techniques result in faster processing and increased number of match points.

1.3 Notation

The symbol $\tilde{u}$ represents a column vector. We will use the letters u, v and w for homogeneous coordinates in image-space In particular, the symbol $\tilde{u}$ represents the column vector $(u, v, w)^T$. Object space points will also be represented by homogeneous coordinates x, y, z and t, or more often x, y, z and 1. The symbol $\tilde{x}$ will represent a point in three-dimensional projective plane represented in homogeneous coordinates. In general, unprimed image coordinates lie in the first or the reference image while the primed image coordinates lie in the target or the second image.

When a vector is represented by a single letter (for example a), it is assumed to be a column vector. The corresponding row vector is written a^T. On the other hand, (a_1, a_2, a_3) represents a row vector. The corresponding column vector is denoted by $(a_1, a_2, a_3)^T$.

Since all vectors are represented in homogeneous coordinates, their values may be multiplied by any arbitrary non-zero factor. The notation $\approx$ is used to indicate equality of vectors or matrices up to multiplication by a scale factor.

2 Processing Steps

As a preprocessing step, an image hierarchy or pyramid is constructed in order to accelerate the computation of match points. This is accomplished by successively reducing both images in the stereo pair to half their size (and resolution) via subsampling using gaussian convolution. The matching process begins at the bottom of the pyramid and works its way up to images with higher and higher resolution. During preprocessing, a set of "interesting points" are also computed in one of the images. The matching process attempts to match only the points in this set with their corresponding points in the other image. These preprocessing steps are largely the same as those in STEREOSYS testbed and the reader is referred to [?, ?, ?] for details.

2.1 Image to Image Transformation.

As mentioned earlier, our system can take as input oblique, rotated, and partially overlapping imagery. We overcome the problems arising because of these effects via a 2D perspective transformation M_I that maps a point $\tilde{u}$ in the first image, to the neighborhood of its corresponding match point $\tilde{u}'$ in the second image. The following observation establishes the existence of such a transformation.

Observation 1 *Let $\tilde{u}_i = [u_i, v_i, 1]^T$ and $\tilde{u}'_i = [u'_i, v'_i, 1]^T$ be the images of points p_i, $i = 1 \ldots n$, in the given image pair. Each $\tilde{u}_i$ in the first image can be transformed to its corresponding match point $\tilde{u}'_i$ in the second image via a 2-dimensional perspective transformation if all p_is lie in a plane.*

Proof: For all match points $[u_i, v_i, 1]$ and $[u'_i, v'_i, 1]$ which are images of points p_i in a plane, we have to show the existence of a 3×3 matrix $M_I = [m_{ij}]$ such that

$$\begin{bmatrix} w'_i u'_i \\ w'_i v'_i \\ w'_i \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix} \begin{bmatrix} w_i u_i \\ w_i v_i \\ w_i \end{bmatrix}. \tag{1}$$

Without loss of generality, assume the all p_is lie in the X-Y plane (i.e. the plane $z = 0$) and the camera matrices are denoted by P_1 and P_2. The image coordinates of each $p_i = [x_i, y_i, 0]$, in the two images, are given by

$$[w_i u_i,\ w_i v_i,\ w_i]^T = P_1 [x_i,\ y_i,\ 0,\ 1]^T \tag{2}$$

$$[w'_i u'_i,\ w'_i v'_i,\ w'_i]^T = P_2[x_i,\ y_i,\ 0,\ 1]^T \tag{3}$$

Equations (2) and (3)) can be rewritten as

$$[w_i u_i,\ w_i v_i,\ w_i]^T = \hat{P}_1[x_i,\ y_i,\ 1]^T \tag{4}$$

$$[w'_i u'_i,\ w'_i v'_i,\ w'_i]^T = \hat{P}_2[x_i,\ y_i,\ 1]^T \tag{5}$$

where $\hat{P}_i$ is the 3×3 matrix formed by deleting the third column of P_i. From these relations it follows that

$$[w'_i u'_i,\ w'_i v'_i,\ w'_i]^T = \hat{P}_2 \hat{P}_1^{-1} [w_i u_i,\ w_i v_i,\ w_i]^T \tag{6}$$

as required. □

The problem of computing M_I can be stated as that of minimizing the sum of errors ϵ_i in the following, possibly overconstrained, system of equations.

$$[u'_i,\ v'_i,\ w'_i]^T = M_I[u_i,\ v_i,\ 1]^T + \epsilon_i. \tag{7}$$

Each match point leads to two equations:

$$\begin{aligned} m_{11}u_i + m_{12}v_i + m_{13} - u'_i(m_{31}u_i + m_{32}v_i + m_{33}) &= 0 \\ m_{21}u_i + m_{22}v_i + m_{23} - u'_i(m_{31}u_i + m_{32}v_i + m_{33}) &= 0 \end{aligned} \tag{8}$$

This system of equation can be cast as a minimum least-square error solution to $Ax = 0$, where x is a 9 dimensional vector containing the entries of M_I, and A is $2n \times 9$ matrix of known coefficients with n being the number of available tie points. Since the entries of M_I are only determined up to a constant multiplier, the constraint $||x|| = 1$ can be imposed to avoid the all-zero solution.

In practice, since p_is do not lie in a plane, there would be deviation in the values computed using M_I. Let $\mathcal{P}$ be the plane that fits p_is the best. M_I would account for all disparities arising because of the change in viewpoint except the deviations which are related to a point's distance from $\mathcal{P}$.

A rough, initial transformation is first computed based on user-provided tie points between the two images. As few as four tie points are sufficient to start the process. Using this rough transformation, unconstrained hierarchical matching, as described in [**?**, **?**], can proceed. Since the computation of M_I is rather fast, only requiring solution to a linear system of equations given in Eq. (7), M_I can be refined at any intermediate point during the hierarchical matching process as more match points become available.

2.2 Computation of Q

The essential matrix Q corresponding to the matching points was defined by Longuet-Higgins [**?**] to be a 3×3 matrix defined by the relation $\tilde{u}_i'^T Q \tilde{u}_i = 0$ for all i. In addition, it was shown in [**?**], and generalized in [**?**] for any arbitrary camera pairs, that Q can be factorized as RS, where R is a rotation matrix and S is a skew-symmetric matrix. In other words, for any stereo pair of images, one can find q_{ij} such that for all match point pairs $[u_k, v_k, 1]$ and $[u'_k, v'_k, 1]$,

$$\begin{bmatrix} u'_k & v'_k & 1 \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \begin{bmatrix} u_k \\ v_k \\ 1 \end{bmatrix} = 0. \tag{9}$$

In order to compute the entries q_{ij} with the help of match points computed earlier, Eq. (9) can be posed as a (possibly overconstrained) system of equations, one equation for each pair of match points. This system of equations has the form $Bx = 0$ where x is the 9 dimensional vector of q_{ij}s. Thus x (equivalently, Q) can be computed using linear, non-iterative techniques, if the constraint $||x|| = 1$ is imposed to avoid the trivial solution.

It can be shown that the Q matrix computed above *must* have two non-zero singular values that are equal, and the third singular value equal to zero, if the two cameras have magnifications equal to unity and zero principal point offsets (i.e., the K matrix describing the internal calibration is an identity matrix) [**?**]. The non-zero singular values of Q need not be equal for cameras about which the assumptions concerning the principal point offsets and magnifications do no hold. In practice, because of the round-off errors and other approximations involved, a computed Q would rarely satisfy this constrain. However, one can use this result to improve the value of Q as follows. Let $Q = UDV^T$ be the singular value decomposition of Q with $D = diagonal(\lambda_1, \lambda_2, \lambda_3)$, and $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Replace D with $D' = diagonal(\frac{\lambda_1+\lambda_2}{2}, \frac{\lambda_1+\lambda_2}{2}, 0)$ (in general, $D' = diagonal(\lambda_1, \lambda_2, 0)$) and recompute $Q = UD'V^T$. It is shown in [**?**] that this gives the best approximation.

For any point $x = [u, v, 1]^T$ in the first image, $[r, s, t]^T = Qx$ represents the equation of the epipolar line on which the match point of x would lie in the second image.

2.3 Recomputation of M_I

It is well known that the search for a match point can be made more efficient by constraining it to the epipolar line in the second image. Epipolar searching can be performed by transforming a point $\tilde{u}$ to $M_I\tilde{u}$ and searching along the line given by $Q\tilde{u}$. However, a difficulty

arises at this point. Even though both M_I and Q are computed using the same set of tie points, there is no guarantee that $M_I\tilde{u}$ actually lies on the line $Q\tilde{u}$. We now describe a method for recomputing M_I so that it satisfies the epipolar constraint.

The problem of recomputing such an M_I can be stated as follows: Find M_I and Q such that:

1. For all $\tilde{u}_i$, the corresponding $\tilde{u}'_i$ in the second image lies on the epipolar line given by Q, i.e., $\forall i, \tilde{u}_i'^T Q\tilde{u}_i = 0$.
2. M_I transforms all $\tilde{u}_i$s so that the overall error between $\tilde{u}'_i$ and $M_I\tilde{u}_i$ is as small as possible, i.e., find M_I such that $\epsilon_i = (\tilde{u}'_i - M_I\tilde{u}_i)$ results in minimum $\sum \epsilon_i.\epsilon_i$.
3. For all points $\tilde{u}$ in the first image, $M_I\tilde{u}$ lies on the epipolar line $Q\tilde{u}$. Note that this is a condition on all $\tilde{u}$ and not just those that are included in the match points obtained so far.

In the previous section, we showed how to compute a Q that satisfies the first condition for all match points. So it only remains to compute an M_I that is compatible with this Q.

The minimization problem in (2.) can once again be stated as that of solving an over-constrained system of equations, $M_I\tilde{u}_i = \tilde{u}'_i$ with minimum least-squared error (see Eq. (8)). In other words, we have to solve equations of the form $Ax = 0$, where x contains the m_{ij}s, under the constraint given in (3.) above.

The third condition, viz., for all points $\tilde{u}$ in the first image, $M_I\tilde{u}$ must lie on the epipolar line $Q\tilde{u}$, dictates that $(M_I u)^T Q u = 0$, $\forall u$. Equivalently,

$$\forall u, \quad u^T (M_I^T Q) u = 0. \tag{10}$$

It can be shown that Eq. (10) can be satisfied if and only if $M_I^T Q$ is skew-symmetric. This leads of six linear constraints, given below, on the entries of M_I.

$$\begin{aligned}
m_{11}q_{11} + m_{21}q_{21} + m_{31}q_{31} &= 0 \\
m_{12}q_{12} + m_{22}q_{22} + m_{32}q_{32} &= 0 \\
m_{13}q_{13} + m_{23}q_{23} + m_{33}q_{33} &= 0 \\
m_{11}q_{12} + m_{21}q_{22} + m_{31}q_{32} &= -(m_{12}q_{11} + m_{22}q_{21} + m_{32}q_{31}) \\
m_{11}q_{13} + m_{21}q_{23} + m_{31}q_{33} &= -(m_{13}q_{11} + m_{23}q_{21} + m_{33}q_{31}) \\
m_{12}q_{13} + m_{22}q_{23} + m_{32}q_{33} &= -(m_{13}q_{12} + m_{23}q_{22} + m_{33}q_{32})
\end{aligned} \tag{11}$$

Here the q_{ij} are known and the m_{ij} are to be determined. Thus the overall problem becomes: Solve $Ax = 0$ subject to the condition $Bx = 0$, where B is a 6×9 coefficient matrix given by Eq. (11) and $x = [m_{11}, m_{12}, \ldots, m_{33}]$ is the vector of unknowns. In order to avoid the trivial solution $x = [0, \ldots, 0]$, we impose the further constraint that $||x|| = 1$.

The problem of minimizing $||Ax||$ subject to $Bx = 0$ and $||x|| = 1$ is a fairly standard problem in linear algebra and there are several ways that it may be done. We describe one way based on the singular value decomposition. In the particular case that we consider here, A represents a set of equations in 9 unknowns, the entries of the matrix M_I, whereas B is a set of 6 equations in the 9 unknowns. As shown below, not all the constraints in Eq. 11 are independent.

Observation 2 *The coefficient matrix B in the constraint equations $Bx = 0$ given by Eq. (11) has rank 5.*

Proof: Note that B is entirely made up of the elements of Q. The result follows from the fact that Q has rank 2. This can be used to show that there is one redundant equation. The details are left to the reader. □

Let $B = UDV^T$ be the singular value decomposition of B, which yields $UDV^T x = 0$ as the set of constraints. There will be one singular value equal to zero, since B has rank 5. We assume that the last element in the diagonal of D is zero. Now, write $x' = V^T x$ (equivalently, $x = Vx'$). Our new task then is to minimize $||AVx'||$ subject to $UDx' = 0$ and $||x'|| = 1$ (since $||x|| = ||x'||$). Now, the solution to $UDx' = 0$ is the same as the solution to $Dx' = 0$ and is equal to $x'_1 = x'_2 = \ldots = x'_5 = 0$. Setting the first five entries of x' to zero, we form a new problem : Minimize $||A'x''||$ subject to $||x''|| = 1$, where A' is the matrix formed from AV by deleting the first 5 columns, and x'' is a vector of length 4.

To solve the problem : Minimize $||A'x''||$ subject to $||x''|| = 1$, once more we take the Singular Value Decomposition of $A' = U'D'V^T$. The solution is then given by $x'' = V(0, 0, 0, 1)^T$ assuming that the last diagonal entry of D' is the smallest.

Once x'' is found, we extend it to a 9-vector x' by adding 5 leading zeros. Then finally, the solution to the original problem is $x = Vx'$. The entries of x can be written as a 3×3 matrix to yield M_I. Q and M_I, computed as above, have the desired property that $M_I\tilde{u}$ lies on the line $Q\tilde{u}$ and epipolar matching can proceed.

2.4 Relative Placement of Cameras and Points

For cameras with known internal calibration Q can be separated into a product $RS(T)$ [**?**], where R is a rotation matrix and $S(T)$, for some $T = (t_x, t_y, t_z)^T$ is a skew symmetric matrix of the form

$$S = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix} \tag{12}$$

It is also possible to accomplish such a factorization for completely uncalibrated cameras [**?**]. In this case, S is still skew-symmetric, however, R is a non-singular (not necessarily a rotation) matrix. The procedure for factorization of Q into R and S matrices is beyond the scope of this paper; the reader is referred to [**?**] for details.

Because of the form of S, it is convenient to use the following notation. If $T = (t_x, t_y, t_z)^T$ is a column vector, then by $S(T)$ is meant the skew-symmetric matrix:

$$\begin{pmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{pmatrix}$$

Matrix $S(T)$ is a singular matrix of rank 2, unless $T = 0$. Furthermore, the null-space of $S(T)$ is generated by the vector T. This means that $T^T.S(T) = S(T).T = 0$ and that any other vector annihilated by $S(T)$ is a scalar multiple of T.

We are interested in computing $X = [x, y, z]^T$ for each pair of match points $\tilde{u} = [u, v, 1]^T$ and $\tilde{u}' = [u', v', 1]^T$. There are two cases. If the factorization of Q into *unique* R and S is known because of some prior knowledge about the cameras — for example, the focal lengths of the two cameras may be known a priori — then the computation of $\tilde{x}$, as described in [**?**], is relatively straightforward. In the absence of any prior information, for completely uncalibrated cameras, it will be shown that an infinite number of solutions exist. Fortunately, it is still possible to find the actual 3-D points if a minimum of 8 ground control points are given.

We consider a general pair of camera matrices represented by $P_1 = (M_1 \mid -M_1T_1)$ and $P_2 = (M_2 \mid -M_2T_2)$. (By completely general we mean that M_1 and M_2 are not restricted to be pure rotations.) We will determine the form of the matrix Q in terms of P_1 and P_2.

Lemma 1 *The essential matrix corresponding to the pair of camera matrices* $(M_1 \mid -M_1T_1)$ *and* $(M_2 \mid -M_2T_2)$ *is given by*

$$Q \approx M_2^* M_1^T S(M_1(T_2 - T_1)).$$

Here A^* represents the adjoint of a matrix A, that is, the matrix of cofactors. If A is an invertible matrix, then $A^* \approx (A^T)^{-1}$.

As is indicated by the previous lemma, an essential matrix Q factors into a product $Q = RS$, where R is a non-singular matrix and S is skew-symmetric. The next lemma shows to what extent this factorization is unique.

Lemma 2 *Let the 3×3 matrix Q factor in two different ways as $Q \approx R_1 S_1 \approx R_2 S_2$ where each S_i is a non-zero skew-symmetric matrix and each R_i is non-singular. Then $S_2 \approx S_1$. Furthermore, if $S_i = S(\tilde{t})$ then $R_2 = R_1 + \tilde{a}\tilde{t}^T$ for some vector $\tilde{a}$.*

Proof: Since R_1 and R_2 are non-singular, it follows that $Q\tilde{t} = 0$ if and only if $S_i\tilde{t} = 0$. From this it follows that the null-spaces of the matrices S_1 and S_2 are equal, and so $S_1 \approx S_2$. Matrices R_1 and R_2 must both be solutions of the linear equation $Q \approx RS$. Consequently, they differ by the value $\tilde{a}.\tilde{t}^T$ as required. (Notice that $\tilde{a}.\tilde{t}^T$, the product of column $\tilde{a}$ and row $\tilde{t}^T$, is a 3×3 matrix.) □

We now prove a theorem which indicates when two pairs of camera matrices correspond to the same essential matrix.

Theorem 1 *Let $\{P_1, P_2\}$ and $\{P_1', P_2'\}$ be two pairs of camera transforms. Then $\{P_1, P_2\}$ and $\{P_1', P_2'\}$ correspond to the same essential matrix Q if and only if there exists a 4×4 non-singular matrix H such that $P_1 H = P_1'$ and $P_2 H = P_2'$.*

Proof : First we prove the **if** part of this theorem. To this purpose, let $\{\tilde{x}_i\}$ be a set of at least 8 points in 3-dimensional space and let $\{\tilde{u}_i\}$ and $\{\tilde{u}_i'\}$ be the corresponding image-space points as imaged by the two camera P_1 and P_2. By the definition of the essential matrix Q satisfies the condition

$$\tilde{u}_i^T Q \tilde{u}_i' = 0$$

for all i. We may assume that the points $\{\tilde{x}_i\}$ have been chosen in such a way that the matrix Q is uniquely defined up to scale by the above equation. The point configurations that defeat this definition of the Q matrix are discussed in [**?**]. Suppose now that there exists a 4×4 matrix H taking P_1 to P_1' and P_2 to P_2' in the sense specified by the hypotheses of the theorem. For each i let $x_i' = H^{-1} x_i$. Then we see that

$$P_1' x_i' = P_1 H H^{-1} x_i = P_1 x_i = u_i, \text{ and,}$$

$$P_2' x_i' = P_2 H H^{-1} x_i = P_2 x_i = u_i'$$

In other words, the image points $\{\tilde{u}_i\}$ and $\{\tilde{u}_i'\}$ are a matching point set with respect to the cameras P_1' and P_2'. Thus the essential matrix for this pair of cameras is defined by the same relationship $\tilde{u}_i^T Q \tilde{u}_i = 0$ that defines the essential matrix of the pair P_1 and P_2. Consequently, the two camera pairs have the same essential matrix.

Now, we turn to the **only if** part of the theorem and assume that two pairs of cameras have the same essential matrix, Q. First, we consider the camera pair $\{(M_1 \mid -M_1 T_1), (M_2 \mid -M_2 T_2)\}$. It is easily seen that the 4×4 matrix

$$\begin{pmatrix} M_1^{-1} & T_1 \\ 0 & 1 \end{pmatrix}$$

transforms this pair to the camera pair $\{(I \mid 0), (M_2 M_1^{-1} \mid -M_2(T_2 - T_1))\}$ where I and 0 are identity matrix and zero column vector respectively. Furthermore by the **if** part of this theorem (or as verified directly using Lemma 1), this new camera pair has the same Q-matrix as the original.

Applying this transformation to each of the camera pairs

$$\{(M_1 \mid -M_1 T_1), (M_2 \mid -M_2 T_2)\} \text{and} \{(M_1' \mid -M_1' T_1'), (M_2' \mid -M_2' T_2')\}$$

we see that there is 4×4 matrix transforming one pair to the other if and only if there is such a matrix transforming

$$\{(I \mid 0), (M_2 M_1^{-1} \mid -M_2(T_2 - T_1))\} \text{to} \{(I \mid 0), (M_2' M_1'^{-1} \mid -M_2'(T_2' - T_1'))\}$$

Thus, we are reduced to proving the theorem for the case where the first cameras, P_1 and P_1' of each pair are both equal to $(I \mid 0)$. Thus, let $\{(I \mid 0), (M \mid -MT)\}$ and $\{(I \mid 0), (M' \mid -M'T')\}$ be two pairs of cameras corresponding to the same essential matrix. According to Lemma 1, the Q-matrices corresponding to the two pairs are $M^* S(T)$ and $M'^* S(T')$ respectively, and these must be equal (up to scale). According to lemma 2, $T \approx T'$. Further,

$$M'^* = M^* + \tilde{a} T^T$$

for some vector $\tilde{a}$. Taking the transpose of this last relation yields

$$M'^{-1} = M^{-1} + T \tilde{a}^T \tag{13}$$

At this point we need to interpolate a lemma.

Lemma 3 *For any column vector $\tilde{t}$ and row vector $\tilde{a}T$, if $I - \tilde{t}\tilde{a}^T$ is invertible then*

$$(I - \tilde{t}\tilde{a}^T)^{-1} = I + k.\tilde{t}\tilde{a}^T$$

where $k = 1/(1 - \tilde{a}^T\tilde{t})$.

Proof : The proof is done by simply multiplying out the two matrices and observing that the product is the identity. One might ask what happens if $\tilde{a}^T\tilde{t} = 1$ in which case k is undefined. The answer is that in that case, $I - \tilde{t}\tilde{a}^T$ is singular, contrary to hypothesis. Details are left to the reader. □

Now we may continue with the proof of the theorem. Referring back to Eq. 13, it follows that

$$\begin{aligned} M' &= (M^{-1} + T\tilde{a}^T)^{-1} \\ &= (M^{-1}(I + MT\tilde{a}^T))^{-1} \\ &= (I - k.MT\tilde{a}^T)M \\ &= M - k.MT(\tilde{a}^T M) \end{aligned}$$

and

$$\begin{aligned} M'T &= MT - k.MT(\tilde{a}^T MT) \\ &= k'.MT \end{aligned} \tag{14}$$

where $k' = 1 - k.\tilde{a}^T MT$. Since T' is a constant multiple of T according to Lemma 2, $M'T' = k''MT$. From these results, it follows that

$$(M' \mid -M'T') = (M \mid -MT) \begin{pmatrix} I & 0 \\ k.\tilde{a}^T M & k'' \end{pmatrix}$$

This completes the proof of the theorem. □

Choosing a Factorization. Given a set of image correspondences $\tilde{u}_i \leftrightarrow \tilde{u}'_i$ defining an essential matrix Q, the previous theorem shows that one cannot unambiguously determine the position of the cameras, or the corresponding object-space points from Q. Since Q contains all the information that is available from the point correspondences, it follows that the position of the cameras and the object points can be determined only up to a 3-dimensional projective transform as specified by the matrix H. In order to determine the positions of the

object-space points $\{x_i\}$ unambiguously, it is necessary for some ground-control points to be specified. Our strategy, therefore, is to select *any* pair of camera placements consistent with the essential matrix, Q. Later, a 3-dimensional projective transform will be carried out to translate to an absolute coordinate system.

The first task is to determine a pair of camera matrices corresponding to a given essential matrix, Q. To this purpose, suppose that the singular value decomposition [**?**] of Q is given by $Q = UDV^T$, where D is the diagonal matrix $D = \text{diagonal}(r, s, 0)$. In a practical case, the smallest singular value of Q will not be exactly equal to 0 because of numerical inaccuracies. However, setting the smallest singular value to 0 gives the matrix closest to Q in Euclidean norm that has the required rank 2. The following factorization of Q may now be verified by inspection.

$$Q = RS \ ; \quad R = Udiag(r, s, \gamma)EV^T \ ; \quad S = VZV^T$$

where

$$E = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ ; Z = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and γ is any non-zero number, but is best chosen to lie between r and s so that the condition number [**?**] of R is as good as possible. From Lemma 1 it follows that the pair of camera matrices

$$P_1 = (I \mid 0),$$
$$P_2 = (Udiag(r, s, \gamma)EV^T \mid U(0, 0, \gamma)^T)$$

correspond to the given essential matrix, Q. It is in no way intended that this should represent the true placement of the cameras, but it is related to the true camera placement by a 3-dimensional projective transformation.

Computation of 3-D Points. The point in the object space that projects on to $\tilde{u}_i$ and $\tilde{u}'_i$ in the two images, according to P_1 and P_2, can be computed as follows. The equations of the rays originating at the focal point of the two cameras and passing through the two match points are given by

$$[w_i u_i, \ w_i v_i, \ w_i]^T = P_1[x_i, \ y_i, \ z_i, \ 1]^T$$
$$[w'_i u'_i, \ w'_i v'_i, \ w'_i]^T = P_2[x_i, \ y_i, \ z_i, \ 1]^T$$

The values of u_i, v_i, u'_i, v'_i, P_1 and P_2 are known. Thus we have and overconstrained system of equations in 5 unknowns and the values $\tilde{x}_i = (x_i, y_i, z_i)$ that minimizes the error can be computed. This will correspond to the point of intersection of these two rays, if they do intersect in space, or the point midway between the points of their closest approach.

2.5 Relative to Absolute Transformation

Since the relative 3-D points computed above may be off by a perspective transformation, ground control points are needed to transform the relative coordinates to absolute coordinates in some user-specified coordinate system. In order to determine absolute placements of the cameras, it is necessary to have at least 8 ground control points to resolve the ambiguity in camera placements derived from the match point data. The method that is used here may be regarded in some ways as a generalization of the method of Sutherland [**?**] to more than one camera. Suppose that we have n cameras represented by matrices $P_1, P_2, \ldots, P_n$ and a set of ground control points $\{x_i\}$, where ground control point x_i is visible in camera $P_{\sigma(i)}$, the corresponding image-coordinates being $\tilde{u}_i$. It is assumed that there is a 4×4 non-invertible matrix H that transforms each P_i to its true placement. This leads to a set of equations

$$\begin{pmatrix} w_i u_i \\ w_i v_i \\ w_i \end{pmatrix} = P_{\sigma(i)} H \begin{pmatrix} x_i \\ y_i \\ z_i \\ 1 \end{pmatrix}$$

The only unknowns in this set of equations are the entries of the matrix H and the values w_i, the above equations may be written as a set of equations

$$\begin{aligned} w_i u_i &= A_i(h_{11}, h_{12}, \ldots, h_{44}) \\ w_i v_i &= B_i(h_{11}, h_{12}, \ldots, h_{44}) \\ w_i &= C_i(h_{11}, h_{12}, \ldots, h_{44}) \end{aligned}$$

where A, B and C are linear expressions in the entries h_{jk} of H. Since the w_i are unknown values, it is possible to eliminate them from the above equations by writing

$$\begin{aligned} C_i(h_{11}, \ldots, h_{44}) u_i &= A_i(h_{11}, \ldots, h_{44}) \\ C_i(h_{11}, \ldots, h_{44}) v_i &= B_i(h_{11}, \ldots, h_{44}) \end{aligned}$$

This gives a set of linear equations in the entries h_{jk} of H, which can be solved to find the matrix H. The solution will be determined only up to a scale factor, corresponding to the fact that H is itself only determined up to a scale factor.

We can now compute the 3-D points by applying the inverse transformation, H^{-1} to the points $\tilde{x}_i$ computed earlier. Points $\tilde{x}_i$ may not have any physical meaning except that they give rise to the known match points when viewed through cameras with $P_1 = (I \mid 0)$ and $P_2 = (Udiag(r, s, \gamma)EV^T \mid U(0, 0, \gamma)^T)$. However, by virtue of Theorem 1 and the fact that ground control points are used in the computation of H, $H^{-1}\tilde{x}_i$ does correspond to the actual 3-D point responsible for the i-th match point.

3 Experimental Validation

The methodology described in this paper has been implemented by augmenting the STEREOSYS testbed with appropriate routines. Fig. **??** shows a stereo image pair showing two overlapping views of a portion of Malibu, California, and the associated image hierarchy (the highest resolution images in the hierarchy, which are 1K×1K in size, are not shown in the figure). The result after epipolar matching using the M_I and Q transformations is shown.

In the figure, the green squares represent the successful matches. These are obtained by starting with an interesting point $\tilde{u}$ in the first image, computing its approximate location in the second image using M_I, and searching along the line given by $Q\tilde{u}$. Once a match $\tilde{u}'$ is found in the second image, the processing is reversed and a match point $\tilde{u}''$, corresponding to $\tilde{u}'$, is found in the first image. If $\tilde{u}$ and $\tilde{u}''$ are close to each other, and the matched points exhibit high correlation, the matched pair is accepted. The red squares in the figure are the results of unsuccessful matches. Typically, matching is unsuccessful because for an interesting point in the first image, the corresponding match point lies outside the second image.

As can be seen, the two images are translated with respect to each other and are only partially overlapping. However, because of the image to image transformation, translations, rotations, and several other discrepancies in the images can be handled.

It was seen that the initial estimation of this transformation, which is based on user-selected tie points, is rather rough and can provide an accuracy of about 4-5 pixels when transforming a point from one image to another. However, this accuracy improves considerably once more tie points become available through unconstrained matching. In fact, recomputation of M_I, as discussed in Section 2.3, generally yields a transformation that gives sub-pixel accuracy after discounting for the parallactic displacement from one image to the other. Because of this accuracy, in both unconstrained hierarchical matching and the epipolar matching, the search is typically initiated very close to the actual match point. This results in faster convergence.

After epipolar searching, exhaustive searching is done to compute match points on a closely spaced grid. These match points, about 3000 in all, are used recompute Q which in turn yields camera transformations P_1 and P_2. 3-D points, in a relative coordinate system, are then computed and are transformed using H — which is computed with the help of ground control points — to get the absolute 3-D location of the points in the object space. Fig. **??** shows the final 3-D model for the image pair in Fig. **??** after Rayleigh interpolation for the missing points. In all the processing steps mentioned here, it was observed that

reliance on linear, non-iterative computations results in considerable saving in run time.

An earlier version of our program used a non-linear, iterative technique, based on a modified Marquardt procedure, to compute the camera parameters. It was observed that unless good initial estimates for the camera parameters are provided, the process sometimes converged to a valid, but physically meaningless, set of parameters. The present technique, which does not require explicit camera modeling, is more robust as no information about camera parameters, exact or approximate, is required. The imagery, and a few control points in order to register the 3-D model to some world coordinate system, is all that is needed.

4 Extensions to Trinocular Stereo

A point in 3D is located by intersecting two rays originating from the camera points and passing through the corresponding points in the image plane. In practice, these rays rarely intersect and one is forced to take the point P in space which is closest to both these rays (i.e., sum of P's perpendicular distances from these two rays is minimum), as the intersection point. Since in trinocular vision the point P is required to be closest to three rays, it achieves better 3D localization of the point.

It is also possible to make the search for match points more efficient if more than 2 views of the same scene are available. In binocular stereo, the epipolar search for match point in the second image has one degree of freedom (i.e., it is confined to a line). For the third image, the degree of freedom can be reduced to zero as shown in Figure **??**.

Consider three images with M_{ij} and Q_{ij} denoting the M_I and Q matrices that take a point $\tilde{u}$ in image i and produce the corresponding match point and epipolar line, respectively, in the image j. A point $\tilde{u}$ in the first image can be transformed to $\tilde{v} = M_{12}\tilde{u}$ on its epipolar line $Q_{12}\tilde{u}$ in the second image as shown in Figure **??**. All the epipolar lines pass through the point C_{12}, the image of the first camera in the second image, as shown. Similarly, $\tilde{u}$ can be transformed to $\tilde{w} = M_{13}\tilde{u}$ on its epipolar line $Q_{13}\tilde{u}$ in the third image.

If one assumes that $\tilde{v}$ is the match point in the second image, then the match point in the third image can be easily found by intersecting $Q_{13}\tilde{u}$ and $Q_{23}\tilde{v}$. In general, since $\tilde{v}$ is not known, the search for the match point the third image can proceed simultaneously with that in the second image.

In order to confirm the match-triple $(\tilde{u}, \tilde{v}, \tilde{w})$, which has been obtained by starting with $\tilde{u}$ in the first image, we can rotate images 1, 2, and 3 and in turn regard image 2 and image 3

as the reference image. The same triple $(\tilde{u}, \tilde{v}, \tilde{w})$ should be obtained, no matter which image is the reference image, for three 3-way matching to succeed.

Acknowledgement

Experimental validation of much of the research presented in this paper would not have been possible without the STEREOSYS program developed by Marsha Jo Hanna at SRI. The authors thank her for use of the program and sharing the source code with them. They also wish to thank Pat Taylor for converting STEREOSYS to C++ and interfacing it to a general-purpose image class hierarchy.

References

[1] M.J. Hanna, *"Bootstrap stereo,"* Proc. Image Understanding Workshop, College Park, MD, April 1980, pp. 201–208.

[2] M.J. Hanna, *"A description of SRI's baseline stereo system,"* ARI International Artificial Intelligence Center Tech. Note 365, Oct. 1985. Workshop, College Park, MD, April 1980, pp. 201–208.

[3] H.C. Longuet-Higgins, *"A computer algorithm for reconstructing a scene from two projections,"* Nature, Vol. 293, 10 Sept. 1981.

[4] R. Hartley, *"Estimation of Relative Camera Positions for Uncalibrated Cameras,"*, Technical Report, GE Corporate R&D, 1 River Road, Schenectady, NY 12301, Oct., 1991.

[5] K.E. Atkinson, "An Introduction to Numerical Analysis," John Wiley and Sons, 2nd Edition, 1989.

[6] I.E. Sutherland, *"Three dimensional data input by tablet,"* Proceedings of IEEE, Vol. 62, No. 4, pp. 453–461, April 1974.

[7] L. H. Quam, *"Hierarchical Warp Stereo,"* Proc. DARPA Image Understanding Workshop, New Orleans, LA, pp. 149–155, 1984 (also in "Readings in Computer Vision, " M.A. Fischler and O. Firschein, Morgan Kaufmann Publishers, Inc., 1987).

[8] T.M. Strat, *" Recovering the camera parameters from a transformation matrix,"* Proc. of DARPA Image Understanding Workshop, New Orleans, LA, pp. 264–271, 1984.

Figure 1: The image hierarchy and the result of epipolar matching. **Note (2001) :** The original images from this paper have been lost.

Figure 2: Terrain elevation model for the image pair in Fig. 1.

Figure 3: 3-Way matching in trinocular stereo.