Rigidity Checking of 3D Point Correspondences Under Perspective Projection

Daniel P. McReynolds
danm@cs.ubc.ca

David G. Lowe
lowe@cs.ubc.ca

Laboratory for Computational Intelligence
Department of Computer Science
2366 Main Mall
The University of British Columbia
Vancouver, B. C. V6T 1Z4
Canada

Abstract

An algorithm is described which rapidly verifies the potential rigidity of three dimensional point correspondences from a pair of two dimensional views under perspective projection. The output of the algorithm is a simple yes or no answer to the question "Could these corresponding points from two views be the projection of a rigid configuration?" Potential applications include 3D object recognition from a single previous view and correspondence matching for stereo or motion over widely separated views. Our analysis begins with the observation that it is often the case that two views cannot provide an accurate structure-from-motion estimate because of ambiguity and ill-conditioning. However, it is argued that an accurate yes/no answer to the rigidity question is possible and experimental results support this assertion with as few as six pairs of corresponding points over a wide range of scene structures and viewing geometries. Rigidity checking verifies point correspondences by using 3D recovery equations as a matching condition. The proposed algorithm improves upon other methods that fall under this approach because it works with as few as six corresponding points under full perspective projection, handles correspondences from widely separated views, makes full use of the disparity of the correspondences, and is integrated with a linear algorithm for 3D recovery due to Kontsevich. The rigidity decision is based on the residual error of an integrated pair of linear and nonlinear structure-from-motion estimators. Results are given for experiments with synthetic and real image data. A complete implementation of this algorithm is being made publicly available.

Keywords

rigidity verification, point correspondences, receiver operating characteristic, structure-from-motion

Contents

1 Introduction

An algorithm is given for accurately and rapidly verifying the potential rigidity of three dimensional point correspondences from a pair of two dimensional views under perspective projection. Our motivation comes from the problem of finding corresponding point features between two or more disparate views of an object. The output of the method is a simple yes or no answer based on the residual error of a minimum variance estimator for a parameter space of rigid transformations and structure. The rigidity verification approach proposed here is shared by other methods for verifying point correspondences using 3D recovery equations as a matching condition. The algorithm, however, substantially improves upon other methods because it works with as few as six corresponding points under full perspective projection, handles correspondences from widely separated views, makes full use of the disparity of the correspondences which necessarily involves the scene structure, and is integrated with a linear estimator based on a weak perspective model.

The matching condition for verifying rigidity is based on a set of 3D scene recovery constraints whose satisfaction minimizes the residual error of an iterative nonlinear optimization algorithm. Although iterative nonlinear methods can be computationally intensive, an accurate answer to the rigidity question is computed quickly for two main reasons. First, a reasonably good initial parameter estimate is computed from a linear algorithm, and secondly, the nonlinear model for 3D recovery makes full use of the image disparity. The 3D recovery equations proposed here are derived from the collinearity condition of the scene and image points under perspective projection which provides a natural and integrated approach to the simultaneous estimation of relative motion and scene structure.

Matching point-feature patterns between images is fundamental to computational vision as evidenced by the large body of literature addressing this problem. It is intrinsic to such tasks as stereo, structure-from-motion and model based recognition. Investigation into the correspondence problem for biological vision has a long history as well [5, 1]. This work is inspired by recently published ideas for recognition by Bennett *et al.* [4] and Kontsevich [17] and has similar goals to the work of Wei, He and Ma [38]. Three applications of the proposed method that are briefly considered include the recognition of a 3D object from a single previous view [4, 17], the detection of multiple moving objects from a sequence of views [33] and stereo correspondence [8].

The algorithm proposed here embodies the rigidity assumption first postulated by Ullman [35]. The rigidity assumption forms an organizing principle for the analysis of changing images due to the relative motion between the observer and the scene. The rigidity assumption can also be an organizing principle for the analysis of stereoscopic images as suggested by Kontsevich. Bennett cites evidence that the rigidity of motion is a key principle in organizing the world into discrete objects [4]. Marr provides a good overview of the utility of the rigidity assumption and places its history in context up to circa 1982 [25]. There is a body of work in the psychology literature which also deals with the perception of rigid and nonrigid 3D point configurations (c.f. [6] and references therein).

The matching condition for verifying the correctness of point correspondences under the rigidity assumption is based on a set of 3D recovery equations that are generally referred to as structure-from-motion equations. Accurately estimating the structure-from-motion parameters from a small set of point correspondences under perspective projection is a difficult problem. In general, if less than eight correspondences are available from two views then nonlinear equations must be solved to yield the parameter values. The resulting system of equations is inherently unstable given noisy observations, requires an initial estimate near the global minimum, and yields multiple solutions. Also, if the point features are projected approximately orthographically then a family of solutions exist for motion and structure. The problem of verifying the potential rigidity of a configuration of 3D points from a pair of views is reduced to an algorithm which minimizes the residual of a nonlinear objective function in the image space coordinates. By only considering the residual error of the estimator in the verification process, potential difficulties that could issue from the use of the estimated parameter values are avoided. This verification process we call rigidity checking and it is worth noting that rigidity checking provides a mechanism for finding geometric invariance from 2D views.

It is assumed that images are formed by a perspective projection and the camera model assumes knowledge of a scale factor that is related to the camera's image size and angular field of view. It is also assumed that the observation variances are known or can be measured. The rigidity decision is based on the residual error of an iterative least-squares estimator based on the Levenberg-Marquardt method. The convergence rate of the nonlinear estimator is improved substantially by an initial estimate of the rotation about the optical axis and a single parameter family of relative depths computed from a linear structure-from-motion algorithm due to Kontsevich [17]. Kontsevich's algorithm assumes images are formed by a scaled orthographic projection. Recent studies provide analysis and experimental evidence of the reliability of estimating the component of rotation about the optical axis [28, 30, 13]. By only relying on an image based residual error criterion to verify potential rigidity, with the ability to handle ill-conditioned solutions and widely separated viewpoints, and by working with a minimal number of correspondences, rigidity checking is seen to be different from traditional structure-from-motion.

2 Related Work In Structure-From-Motion

The rigidity checking method is based on a matching condition derived from a set of structure-from-motion equations. This section reviews solutions to the problem of structure-from-motion that are similar to the rigidity checking solution or are common to other methods for correspondence verification.

2.1 Structure-from-motion and Orthographic Projection

The nonlinear optimization algorithm is initialized with a motion and structure estimate provided by a linear algorithm based on Kontsevich's algorithm for a pair of scaled orthographically projected images [17]. Kontsevich's algorithm is similar to the one described by Koenderink and Van Doorn [16] and Basri [2] in that they make the same geometric arguments concerning recovery of scale and the rotation component in the image plane of the motion between a pair of views. The main contribution of Kontsevich's algorithm is its simple linear formulation for recovering the epipolar geometry as a change of basis transformation. The algorithm avoids computing rotation angles and is easily implemented. Koenderink and van Doorn do not specify an algorithm but outline the geometric constraints that determines a construction that will yield the scale and rotation parameters for the rotation component in the image plane. Kontsevich and Koenderink and Van Doorn both derive the same expression for depth recovery. Basri derives the same set of linear constraints for orthographically projected images as Kontsevich but determines the epipolar geometry by solving explicitly for the planar rotation angles. Also, Basri is not concerned with depth recovery. Basri also cites Ullman as having earlier derived essentially the same transformation as his but with a minimum of five corresponding points rather than four. Other derivations for recovering the epipolar geometry and scene structure from weak perspective views were reported by Huang and Lee [14], Lee and Huang [19], Bennett *et al.* [3], Harris [10], Hu and Ahuja [13] and more recently by Nishimura *et al.* [28]. Sudhir *et al.* [31] present related work for finding point correspondences under affine (not necessarily rigid) motion for weak perspective projection and a minimum of three views.

Huttenlocher and Kleinberg have published an interesting theoretical result for the problem of deciding whether two views of a 3D point set are of the same rigid object under orthographic and central projection. Their verification step for an hypothesized labeling of the points is based on exactly the same constraint as Kontsevich's for orthographic projection [15].

2.2 Structure-from-motion and Perspective Projection

Methods Based On Nonlinear Estimation

The nonlinear estimation process described here is based on the iterative least-squares Levenberg-Marquardt algorithm with additional parameter stabilization that effectively improves the reliability of the estimation process when the solution is minimally constrained and the observations are noisy. This stabilization is most frequently necessary when there is a large amount of perspective distortion and the views are nearly, or are, the same.

The basic nonlinear algorithm was developed earlier [26, 27]. Szeliski and Kang have recently and independently developed an algorithm for structure-from-motion which is similar to our method [32]. Their application is not the same as the one described here, however, in that they are seeking to recover object structure from many views and correspondences. Their formulation differs from ours in several ways. They incorporate and solve for a parameter that describes the global offset of the scene points from the camera coordinate frame. They also solve for a global scale factor related to the focal length and global depth offset parameter. We assume a fixed value for the global depth offset (as this cannot be determined from the images) and a known image scale factor which is related to the focal length and camera's field of view. They do not address the problems associated with having only a small number of views and correspondences and the use of stabilization methods to improve the algorithm's performance.

Other recent related work in nonlinear estimation of structure-from-motion is from Weng *et al.* [39, 40]. Their minimum variance optimal estimator minimizes an image based error function for point correspondences similar to the one described here. The major difference with the rigidity checking method concerns the structure parameters which are solved for simultaneously with the motion parameters but are factored out into a separate optimization step in their formulation. They argue for this decomposition because it reduces the computational load and more importantly because their minimum variance estimator only involves the minimum number of parameters for the structure-from-motion problem. The dimensionality of the rigidity checking parameter space is not large because, rather than solving for 3D scene coordinates, only the scene depths in a reference coordinate frame are determined in addition to the motion parameters. Their results for the Levenberg-Marquardt batch solution method are generally better than their iterated extended Kalman filter (IEKF) results. The better performance of batch techniques over IEKF methods for structure-from-motion is also supported by other experiments [18]. The proposed rigidity checking method is based on a Levenberg-Marquardt batch solution. They also analyze the limitation of small interframe motion and derive the Cramer-Rao lower error bound for the motion estimates. This lower bound predicts the instability and large expected error in the estimation of the motion and structure parameters from views that are close together. Briefly discussed in a later section is a stabilization method incorporated into the rigidity checking algorithm which addresses exactly this issue.

Methods Based On Linear Estimation

Another class of structure-from-motion algorithms for images under perspective projection solve a system of linear equations for a set of motion parameters. These algorithms are based on the coplanarity constraint of corresponding points from two views. The classic linear algorithm was first developed by Longuet-Higgins and requires a minimum of eight point correspondences to fully constrain the unknown parameters [21]. Wei *et al.* employ the linear formulation developed by Tsai and Huang [34] to solve the correspondence problem and estimate motion simultaneously. Their method, similar to the rigidity checking method, uses an hypothesize and test approach to determine the correct correspondences with the 3D recovery equations as a matching condition. Their method, however, requires a minimum of nine correspondences for a least-squares estimate compared to six for the rigidity checking method. The reliability of their matching constraint is reduced for images that are projected nearly orthographically or are noisy for reasons described in the next paragraph.

Weng *et al.* [39] discuss the limitations of the epipolar constraint for accurately estimating the motion and structure parameters. The main point of their discussion is that the epipolar constraint recovers the motion parameters independently of the scene structure. This independence has the advantage of yielding a linear system for estimating the motion parameters but has the disadvantage that the solution space accommodates a scene structure space which includes many physically impossible scene solutions, i.e., the motion solution leads to a violation of the "depths positive" criterion for the scene structure. The violation of the "depths positive" criterion is a result of the ambiguity of the motion solution in the presence of even a small amount of noise or when the images are projected nearly orthographically. This large scene space that includes physically impossible scenes accounts for the experimentally observed large false positive rate exhibited by epipolar constraint based methods for verifying correspondences. The rigidity checking method is based on the collinearity constraint for perspective projection and therefore makes full use of the image disparity which necessarily involves the scene structure. This constraint reduces the ambiguity of the motion solution by improving the discriminatory power of the estimator for rotation and translation by embedding local structure constraints on the image disparities within the global motion constraints. Also, because the "depth's positive" criterion is implicitly enforced in the simultaneous solution of motion and structure, the search in the motion parameter space is also appropriately restrained to a valid region.

3 Derivation of the Rigidity Checking Equations

3.1 Imaging Model

The standard perspective projection model is assumed. There are two coordinate systems relevant to this formulation, one camera-centered and the other object-centered. The object-centered frame can also be thought of as a intermediate camera frame which is identified with a reference camera frame by a translation along the optical axis. The object structure is recovered in this stable reference frame. Experiments show that estimating the motion and structure in an object-centered coordinate system noticeably improves the stability of the recovery method. This is especially true when the object motion is dominated by rotation. Intuitively, working in an object-centered frame helps to decouple the effects of rotational motion from translational motion since object rotations require smaller compensating translations. Recovering motion and structure in an object-centered coordinate frame is similar to the model described by Kumar *et al.* [18] with the exception that we are solving for discrete rigid body transformations rather than velocities. The transformation from the intermediate camera frame to the camera centered frame is given by a translation t_{cz} along the optical axis.

The rigid body motion of point $\mathbf{p}_j$ from the reference object frame to a subsequent camera frame is given by

$$\tilde{\mathbf{P}}_j = \mathbf{R}\mathbf{p}_j + \mathbf{t} + \mathbf{t}_c \tag{1}$$

where $\mathbf{t}_c$ maps the intermediate camera frame to the camera centered frame and is given by $\mathbf{t}_c = [0, 0, t_{cz}]^T$, $\mathbf{R}$ is a 3 by 3 orthonormal rotation matrix from the intermediate camera frame to the frame of the second image, $\mathbf{t}$ is the translation vector and $\tilde{\mathbf{P}}_j$ is the j^{th} object point in the second camera frame.

Image coordinates of the j^{th} point are given by

$$\mathbf{w}_j = \begin{bmatrix} u_j \\ v_j \end{bmatrix} = \frac{-f}{P_{jz}} \begin{bmatrix} P_{jx} \\ P_{jy} \end{bmatrix} \tag{2}$$

where f is a scale factor related to the image size and angular field of view of the camera.

3.2 Least-Squares Solution Of Nonlinear Equations

An image based error function $\mathbf{e}(\mathbf{w}, \mathbf{w}', \mathbf{m})$ can be written that relates the position of the projected feature points in different views to the rigid body transformation parameters and the depths of the points as follows

$$\mathbf{e}(\mathbf{w}, \mathbf{w}', \mathbf{m}) = \mathbf{w}' - \mathbf{y}(\mathbf{w}, \mathbf{m}) \tag{3}$$

where $\mathbf{m}$ is the set of parameters for motion and the depths of the scene points in the reference object frame, $\mathbf{w}$ is the coordinate vector of the point in the reference image frame and $\mathbf{w}'$ is the corresponding point in the second image frame.

The function $\mathbf{y}(\mathbf{w}, \mathbf{m})$ maps the image coordinates in the reference image frame into the nonreference image frame by first back-projecting the image coordinates into the reference camera centered frame, i.e., find P_x and P_y from (2) according to the latest depth estimates. This is followed by a transformation to the reference object frame by a translation along the optical axis, then to the nonreference camera frame according to the latest motion estimate by equation (1) and finally projecting into the nonreference image frame by (2). Because of the smoothness and well behaved properties of the nonlinear projection equation (2) applying the Gauss-Newton method to the estimation of the transformation and depth parameters is a good choice. Although the method requires an initial guess and there is a risk of converging to a local minimum, we show below that by incorporating stabilization methods and Levenberg-Marquardt extensions the use of the Gauss-Newton method can work well in practice.

Rather than solving directly for the parameters $\mathbf{m}$ which minimizes (3), the Gauss-Newton method computes a vector of corrections $\mathbf{h}$ which are added to the current parameter estimates. The $i^{th} + 1$ estimate is given by

$$\mathbf{m}^{i+1} = \mathbf{m}^i + \mathbf{h}. \tag{4}$$

Based on the assumption that (3) is locally linear, the effect of each parameter correction h_k on the error measurement is determined by multiplying the correction h_k by the partial derivative of the error with respect to that parameter. Therefore, we can solve for $\mathbf{h}$ in the following matrix system

$$\mathbf{J}\mathbf{h} = -\mathbf{E}$$

where $\mathbf{J}$ is the Jacobian matrix $J_{ik} = \frac{\partial E_i}{\partial h_k}$. $\mathbf{E}$ is the vector of $\mathbf{e}_j$ for correspondence j.

Computing The Partial Derivatives

The error measure is just the difference between the observed image point and the image point from the reference view mapped into the nonreference (or current) view, $\mathbf{w}'_j - \tilde{\mathbf{w}}_j = \mathbf{e}_j$, where $\tilde{\mathbf{w}}_j$ is a function of the depth of the point in the reference object-centered frame, and the translation and rotation between the two views. The partial derivatives of $\tilde{\mathbf{w}}_j$ are given by

$$\frac{\partial \tilde{u}_j}{\partial h_k} = \frac{-f}{\tilde{P}_{jz}}\left(\frac{\partial \tilde{P}_{jx}}{\partial h_k} - \frac{\tilde{P}_{jx}}{\tilde{P}_{jz}}\frac{\partial \tilde{P}_{jz}}{\partial h_k}\right) \qquad \text{and} \qquad \frac{\partial \tilde{v}_j}{\partial h_k} = \frac{-f}{\tilde{P}_{jz}}\left(\frac{\partial \tilde{P}_{jy}}{\partial h_k} - \frac{\tilde{P}_{jy}}{\tilde{P}_{jz}}\frac{\partial \tilde{P}_{jz}}{\partial h_k}\right).$$

The partial derivatives of $\tilde{P}_{jx}$, $\tilde{P}_{jy}$ and $\tilde{P}_{jz}$ with respect to h_k are the components of a set of directional derivatives in a camera centered frame. These parameters are the depth values of the points in the reference frame, P_{jz}, the three translation parameters t_x, t_y and t_z and the parameterization of the rotation component. Determining the partial derivatives with respect to rotation poses some difficulty, since rotation has only three degrees of freedom and any formulation with a rotation operator necessarily involves more than three terms which often then become the rotation parameters, e.g., the nine elements of an orthonormal rotation matrix. Furthermore, the formulation makes no suggestion as to the appropriate representation in terms of its underlying parameters. If we compose three rotations about individual axes to compute an arbitrary 3D rotation, singularities can occur if the sequential composition fails to specify independent directions of rotation. Therefore, we represent full three-degree-of-freedom rotations with a 3 by 3 orthonormal rotation matrix and compute corrections about each of the coordinate axes to be composed with this rotation. These corrections are approximately independent for small angles. They are also extremely efficient to compute. For example, the directional derivative of a point with respect to an incremental rotation about the x axis is the vector $(0, z, -y)$ for a change of basis convention for rotation, where z and y refer to the coordinates of the vector from the origin of the rotation to the point.

These derivatives are analytical closed form expressions. An important advantage of this fact is that the Jacobian matrix can be computed very efficiently and accurately. Hence, we are not required to use computationally expensive finite difference methods.

Partial Derivatives of ũ and ṽ

The partial derivative of the image coordinate $\tilde{u}_j$ with respect to the depth parameter is

$$\frac{\partial \tilde{u}_j}{\partial P_z} = \frac{\partial \tilde{u}_j}{\partial \tilde{P}_z}\frac{\partial \tilde{P}_z}{\partial P_z} = \frac{-f}{\tilde{P}_{jz}^2}\left(\tilde{P}_{jz}\frac{\partial \tilde{P}_{jx}}{\partial \tilde{P}_z} - \tilde{P}_{jx}\frac{\partial \tilde{P}_{jz}}{\partial \tilde{P}_z}\right)\frac{\partial \tilde{P}_z}{\partial P_z}.$$

Now write $\tilde{P}_{jx}$ and $\tilde{P}_z$ as

$$\tilde{P}_{jx} = \left(P_{jz}\mathbf{R}\bar{\mathbf{P}}_j + \mathbf{t}\right)_x$$
$$\tilde{P}_z = \left(P_z\mathbf{R}\bar{\mathbf{P}} + \mathbf{t}\right)_z$$

where $\bar{\mathbf{P}}_j$ is given by

$$\bar{\mathbf{P}}_j = \begin{bmatrix} \frac{-u_j}{f} \\ \frac{-v_j}{f} \\ 1 \end{bmatrix}.$$

Then,

$$\frac{\partial \tilde{P}_{jx}}{\partial \tilde{P}_z} = \frac{\frac{\partial \tilde{P}_{jx}}{\partial P_z}}{\frac{\partial \tilde{P}_z}{\partial P_z}} = \frac{(\mathbf{R}\bar{\mathbf{P}}_j)_x}{(\mathbf{R}\bar{\mathbf{P}}_j)_z}$$

when $\tilde{\mathbf{P}} = \tilde{\mathbf{P}}_j$ else $\frac{\partial \tilde{\mathbf{P}}_j}{\partial \tilde{P}_z} = \mathbf{0}$. Therefore

$$\frac{\partial \tilde{u}_j}{\partial P_z} = \frac{-f}{\tilde{P}_{jz}^2}\left(\tilde{P}_{jz}\left(\mathbf{R}\bar{\mathbf{P}}_j\right)_x - \tilde{P}_{jx}\left(\mathbf{R}\bar{\mathbf{P}}_j\right)_z\right)$$

when $\tilde{\mathbf{P}} = \tilde{\mathbf{P}}_j$ else it equals 0.

Rotations are parameterized by incremental rotations about the coordinate axes. Let these parameters be designated by ϕ_x, ϕ_y and ϕ_z. Then the partial derivative with respect to ϕ_x is

$$\frac{\partial \tilde{u}_j}{\partial \phi_x} = \frac{-f}{\tilde{P}_{jz}^2}\left(\tilde{P}_{jz}\frac{\partial \tilde{P}_{jx}}{\partial \phi_x} - \tilde{P}_{jx}\frac{\partial \tilde{P}_{jz}}{\partial \phi_x}\right) = \frac{-f\tilde{P}_{jx}p'_{jy}}{\tilde{P}_{jz}^2}$$

where

$$p'_{jy} = (\mathbf{R}\mathbf{p}_j)^T \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

The partial derivative with respect to ϕ_y is

$$\frac{\partial \tilde{u}_j}{\partial \phi_y} = \frac{-f}{\tilde{P}_{jz}^2}\left(\tilde{P}_{jz}\frac{\partial \tilde{P}_{jx}}{\partial \phi_y} - \tilde{P}_{jx}\frac{\partial \tilde{P}_{jz}}{\partial \phi_y}\right) = f\left(\frac{p'_{jz}}{\tilde{P}_{jz}} + \frac{\tilde{P}_{jx}p'_{jx}}{\tilde{P}_{jz}^2}\right)$$

where

$$p'_{jx} = (\mathbf{R}\mathbf{p}_j)^T \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } p'_{jz} = (\mathbf{R}\mathbf{p}_j)^T \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The partial derivative with respect to ϕ_z is

$$\frac{\partial \tilde{u}_j}{\partial \phi_z} = \frac{-f}{\tilde{P}_{jz}^2}\left(\tilde{P}_{jz}\frac{\partial \tilde{P}_{jx}}{\partial \phi_z} - \tilde{P}_{jx}\frac{\partial \tilde{P}_{jz}}{\partial \phi_z}\right) = \frac{-fp'_{jy}}{\tilde{P}_{jz}}.$$

The partial derivatives of $\tilde{u}_j$ with respect to the three translation components t_x, t_y and t_z are now given. The partial with respect to t_x is

$$\frac{\partial \tilde{u}_j}{\partial t_x} = \frac{-f}{\tilde{P}_{jz}^2}\left(\tilde{P}_{jz}\frac{\partial \tilde{P}_{jx}}{\partial t_x} - \tilde{P}_{jx}\frac{\partial \tilde{P}_{jz}}{\partial t_x}\right) = \frac{-f}{\tilde{P}_{jz}}.$$

Similarly, the partials with respect to t_y and t_z are

$$\frac{\partial \tilde{u}_j}{\partial t_y} = 0 \text{ and } \frac{\partial \tilde{u}_j}{\partial t_z} = \frac{f\tilde{P}_{jx}}{\tilde{P}_{jz}^2}.$$

The partial derivatives of $\tilde{v}_j$ are derived similarly.

The solution for motion and depth can only be recovered up to a global scale factor. To reduce the number of degrees of freedom it is convenient to fix the global scale by either fixing one of the depth parameters or by constraining the translation to have a unit norm. Here, the value of the first depth parameter is set to zero in the object frame. Hence, from a simple counting argument for corresponding points, a minimum of two views and five points are required to yield a system of equations of full rank where each match contributes two independent constraints.

4 Implementation Issues

4.1 Integrating The Linear and Nonlinear Algorithms

Kontsevich describes a fully linear algorithm for recovering the epipolar geometry of two sets of corresponding points under scaled orthographic projection by formulating the problem in terms of a change of basis in Euclidean coordinates [17]. The rigid 3D structure of the configuration is also recovered. The epipolar geometry is recovered under the rigidity assumption. The change of basis transformation maps the two point sets into a common object centered frame which is referenced to a rotation axis that corresponds to the rotation component in depth. The translation component of the motion is excluded immediately by considering the transformation of the edge vectors formed by joining the object points. The change of basis transformation excludes the scaling component and the rotation component about the viewing direction. The recovery problem is reduced to the standard problem of binocular stereo where the unknown rotation component is about the vertical axis. A one parameter family of object structure is recovered, parameterized by the angle of rotation about the vertical axis. A minimum of four non-coplanar point correspondences are required to satisfy the constraints.

The novel geometric interpretation of the constraint that determines the rotation component about the viewing direction (a vector orthogonal to the image plane) is the observation that the orthogonal projection of an edge onto an axis V (which is identified with the moving object) is invariant to rotation about the viewing direction. The axis V is the 3D rotation axis orthogonally projected onto the image plane of the first view and is derived from the decomposition of an arbitrary rotation into a rotation about some axis V in the image plane and a rotation about an axis orthogonal to the image plane. The rotation component about the viewing direction maps axis V in the first view to axis V' in the second view. The projected edge on V is identical (up to a scale factor) to the corresponding edge projected onto V'. Once the axes V and V' are determined the scale factor is also determined and a basis transformation can be formed for each image which, when applied to the two edge sets, results in a binocular stereo problem in an object centered frame for rotation about the vertical axis. The object structure is uniquely determined by the rotation angle about the vertical axis which is a free variable, hence, there is a one parameter family of solutions for object structure.

The axes V and V' can only be determined unambiguously (up to a sign) if the 3D configuration is non-coplanar and the two sets of image measurements are not related by an affine transformation (which would be the case for a rotation only about the viewing direction).

The constraint that the orthogonal projection of any edge onto axis V equals the projection of the corresponding edge onto axis V' can be used as a consistency criterion for verifying point correspondences under scaled orthographic projection as described by Kontsevich. This consistency criterion has recently been described by Huttenlocher and Kleinberg [15] (1994) in a paper that addresses the combinatoric problem of verifying point correspondences under orthographic and central projection for noise free observations. Huttenlocher and Kleinberg describe a new algorithm with a low order polynomial time complexity for the problem of labeling and verifying the correspondences between two sets of projected points. Their verification criterion for orthographically projected point correspondences is identical to Kontsevich's (note that Kontsevich's criterion is more general since it handles scaled orthographic images). Unfortunately, it is an open question if the complexity of Huttenlocher's algorithm for noisy observations approaches the bounds for noise free observations.

Kontsevich's algorithm is applicable to multiple frames. In the case of three frames it is sufficient to compare the structure estimates from two pairs of comparisons. If the structure is unique then real scalars, λ and μ, can be found which satisfy the structure equalities. There is a problem, however, with the solution strategy suggested by Kontsevich for a special motion case. The special motion case occurs, for example, when there is no rotation about the viewing direction between any pair of views. The details of the problem will be omitted here since we are primarily concerned with the two view case, however, it should be noted that the solution for the two scalars is not a simple linear problem. In fact, solving for the scalars involves finding the real roots of a fourth order polynomial. The solution is further complicated when the image measurements are noisy.

The linear estimation algorithm provides a single parameter family of solutions for depth parameterized by the rotation angle about the vertical axis. For the purpose of initializing the rotation estimate this value is arbitrarily set to 45 degrees. For the purpose of initializing the depth values a rotation angle of approximately 10 degrees is chosen. This value overestimates the scale of the depth solution if the actual rotation angle is greater than 10 degrees since depth is proportional to the product of disparity and the inverse sine of the rotation angle. The mean value of inverse sine over the range of 1 to 45 degrees corresponds to an angle of about 10 degrees. The depth scaling corresponding to a 10 degree rotation angle was found experimentally to give the best convergence results for images with small or large perspective distortions.

4.2 Stabilizing The Nonlinear Solution

As long as there are significantly more constraints on the solution than unknowns, the Gauss-Newton method, as described earlier, will usually converge in a stable manner from a wide range of starting positions. However, in motion analysis and structure recovery, it is often the case that the observations can lead to an ill-conditioned solution even when the parameters are over-constrained. Viewing situations that can lead to ill-conditioning include views that are closely spaced together or that are projected approximately orthographically. This treatment of regularizing the structure-from-motion problem is based on material presented by Lowe for a model based vision problem [23].

Specifying Prior Constraints. Convergence problems can be ameliorated by introducing prior constraints on the desired solution that specify the corrections to make in the absence of further data. For our problem the prior expectation will simply be to solve for zero corrections to the current parameter estimates. This constraint can be seen at the core of the Levenberg-Marquardt method. In its simplest form the Levenberg-Marquardt method stabilizes the solution by including a global regularization parameter which does not adequately address the issue of the appropriate relative weighting of the parameters and the contribution of a parameter's error to the global error measure. Prior knowledge of the expected value of the parameters, even if only in terms of order of magnitude, allows each parameter to be stabilized according to its contribution to the global error measure. One appropriate measure of a parameter's contribution to the residual error is its variance. Each parameter should be weighted so they all contribute equally according to their distance from their expected value in standard deviations.

Prior constraints on the solution can be incorporated by adding rows to the linear system constraining the value that is to be assigned to each parameter correction

$$\begin{bmatrix} \mathbf{J} \\ \mathbf{I} \end{bmatrix} \mathbf{h} = \begin{bmatrix} -\mathbf{E} \\ \mathbf{0} \end{bmatrix}. \tag{5}$$

The identity matrix $\mathbf{I}$ adds one row for specifying the value of each parameter correction, and we specify a zero *a priori* value for the correction. Each iteration of the Gauss-Newton method determines a new correction to be added to the current parameter estimate. The prior constraint is set to zero and stipulates that the correction should be to leave the parameter estimate at its current value in the absence of a strong constraint from the data. The relative weighting of these constraints will be described in the next section.

Knowledge about the scene geometry can aid in setting bounds on the range of expected parameter values which in turn can be used to specify standard deviations. Rotations for example will have a standard deviation of at most $\pi/2$ and translations must be limited to keeping the object in the field of view. We make an assumption about the approximate location of the object-centered frame relative to the camera frame. The standard deviation of the depth parameters is proportional to this distance as a function of the degree of stabilization desired. A relatively small standard deviation corresponds to a greater degree of stabilization which will cause the depth estimates to remain closer to their initial values.

These deviations may be large relative to the deviations arising from the image measurements, but they still play an important role in stabilizing the solution for ill-conditioned problems.

Efficient Computation of Stabilization. The prior estimates of the correction values will be weighted by a diagonal matrix $\mathbf{W}$ in which each weight is inversely proportional to the standard deviation σ_i for parameter i

$$W_{ii} = \frac{1}{\sigma_i}.$$

This matrix is used to scale each row of the prior values in the lower part of (5). We assume that the constraints based on image measurements in the upper part of the equation have already been scaled to a unit standard deviation.

$$\begin{bmatrix} \mathbf{J} \\ \mathbf{W} \end{bmatrix} \mathrm{h} = \begin{bmatrix} -\mathbf{E} \\ \mathbf{0} \end{bmatrix}.$$

We will minimize this system by solving the corresponding normal equations

$$\left[\mathbf{J}^T\mathbf{W}^T\right] \begin{bmatrix} \mathbf{J} \\ \mathbf{W} \end{bmatrix} \mathrm{h} = \left[\mathbf{J}^T\mathbf{W}^T\right] \begin{bmatrix} -\mathbf{E} \\ \mathbf{0} \end{bmatrix}$$

which multiplies out to

$$\left(\mathbf{J}^T\mathbf{J} + \mathbf{W}^T\mathbf{W}\right) \mathrm{h} = -\mathbf{J}^T\mathbf{E}.$$

Since $\mathbf{W}$ is a diagonal matrix, $\mathbf{W}^T\mathbf{W}$ is also diagonal but with each element on the diagonal squared. This means that the computational cost of the stabilization is negligible, as we can first form $\mathbf{J}^T\mathbf{J}$ and then simply add small constants to the diagonal that are the inverse of the variance of each parameter.

Forcing Convergence. Even after incorporating this stabilization based on reasonable assumptions of the expected values of the parameters, it is possible that the system will fail to converge to a minimum due to the fact that this is a linear approximation of a nonlinear system. The standard method to deal with this situation is to use the Levenberg-Marquardt extension to iterative nonlinear least squares [20][24]. The Levenberg-Marquardt scaling parameter λ is used to increase the weight of stabilization whenever divergence occurs. Increasing the value of λ will essentially freeze the parameters having the lowest standard deviations and therefore solve first for those with higher standard deviations. For our problem, this implies convergence for difficult problems will proceed by solving first for translations and rotations and then proceeding on subsequent iterations to solve for depths.

4.3 Scaling The Image Coordinates And The Global Depth Offset

The translation vector $\mathbf{t}_c$ in equation (1) that transforms scene points from the object centered frame to the camera centered frame can be viewed as a global depth offset. This offset is unknown and cannot be determined from the image data. Its value essentially determines the degree of "perspective". The closer the object is to the camera for a given object size the greater the amount of perspective distortion in the object's image. The global depth offset is defined to be proportional to the focal length of the camera as are the scene depths. It is important, therefore, that the initially estimated depths from the linear estimator be proportional to the focal length as well. This is achieved through the scaling of the image measurements by $1/f$ where f is the known image magnification factor (see equation (2)). Experiments with synthesized data for a typical range of object sizes and distances reveals that a value of 2 focal lengths for the global depth offset is a good compromise. A value of 2 is a deliberate under-estimate because convergence is noticeably improved.

An issue related to the global offset is the choice of the coordinate frame origin for the structure estimated by the nonlinear algorithm. The idea is to determine which scene point is closest to the camera in order not to solve for its depth. This implies that all other estimated depths should be further away from the camera. This helps to reduce the possibility that the structure estimate will violate the "depths positive" criterion due to a poor initial estimate of the global offset. Given the linearly estimated structure, the list of correspondences is reordered to place the correspondence with the most positive depth first (note that the optical axis is defined to be in the negative z direction). Since the nonlinear estimator assigns a fixed depth value to the first correspondence on the list, the expected depths relative to the first depth should all be farther away from the image plane.

4.4 Translation Initial Estimate

The default initial estimate for translation is zero except in the special case where the scale factor estimated by the linear algorithm suggests that the scene receded from the camera. In this case, the image scale factor is used to estimate the translation component in depth, t_z. The assumption is that the global scale factor between views estimated by Kontsevich's algorithm is a reasonable indication of the motion of the object in depth. From Kontsevich's algorithm, the scale factor s is given by $s|\mathbf{r}| = |\mathbf{r}'|$ where $\mathbf{r}$ is a 3D edge vector and $\mathbf{r}'$ is the corresponding edge after the view transformation. The value of s is determined from the average ratio of the projected edge vector magnitudes before the view transformation to the magnitudes after the view transformation.

To see the relationship between translation in depth and s under perspective projection, the scale factor can be written in terms of, say, the u coordinates of a pair of corresponding points as $s = \frac{u'}{u}$ where u under perspective projection is given by equation (2). The u component from equation (2) can be rewritten as

$$u = \frac{-kx}{\eta z_o + 1}$$

where $k = \frac{f}{t_{cz}}$, $\eta = t_{cz}^{-1}$, z_o is the point's depth in the object frame, and t_{cz} is the global depth offset. Now the scale factor can be rewritten as

$$s = \frac{(-kx')\ (\eta z_o + 1)}{(-kx)\ (\eta z_o' + 1)}. \tag{6}$$

Under the assumption that the global scale change is due to a motion dominated by translation in depth (i.e., the rotation component is negligible), and assume that there is no translation parallel to the image plane (the effect of translation parallel to the image plane can be minimized by considering the distances between points), then $x' \simeq x$ and $z_o' \simeq z_o + t_z$. With these assumptions equation (6) simplifies to

$$t_z = \frac{(\eta z_o + 1)\,(1 - s)}{s\,\eta}.$$

Under the assumption that $\eta\ z_o << 1$, which holds reasonably well for distant compact objects, the expression for t_z reduces to

$$t_z = \frac{1 - s}{s\,\eta} = \frac{(1 - s)}{s}\ t_{cz}. \tag{7}$$

Equation (7) is the expression that is used in the algorithm's implementation. The initial estimate for t_z is only bound to the given expression if s is less than one indicating the object translated away from the camera. There are two reasons for this: the first is that experimental results indicated that the nonlinear estimator had less trouble converging to the correct solution if the object loomed rather than receded, and secondly, the global offset is set to a small value and initializing the motion to bring the object closer to the camera increased the risk of violating the "depths positive" criterion upon convergence. Monte Carlo test results for the scenario with a large amount of perspective distortion indicated that the correct verification rate improved by approximately 7 percent with this initialization for t_z incorporated in the algorithm.

4.5 Extending The Set Of Matches

The description of the rigidity checking algorithm up to this point has stressed the method's performance with only 6 point correspondences. The algorithm, however, readily accommodates extending the set of matches by adding additional correspondences to the currently verified set and reverifying the new larger set. A larger set of correspondences improves the accuracy of the verification decision. With a sufficiently large number of correspondences verified, a modified version of the rigidity checking method or some other method can be used to determine the epipolar lines reliably. With the epipolar geometry reliably estimated, the search space for more correspondences is reduced to a one dimensional search in the image.

4.6 Nonlinear Estimation With Disparity Rather Than Depth

A study of the parameterization of the nonlinear estimator revealed a small improvement in the stability and rapidity of convergence when inverse depth (which we call disparity) was estimated rather than depth. Distant scene points have small disparity values which improves the conditioning of the Hessian matrix. Harris and Pike also estimate inverse depth to avoid a nearness bias for their estimation process due to their use of ellipsoids to model structure uncertainty [11].

5 Experimental results

Monte Carlo simulations were run to determine the parameter space bounds over which the algorithm is effective. The algorithm was also tested on real images with manually selected matches.

For the simulated data, unless stated otherwise, the camera focal length was set to unity and the field of view was specified by the size of the image frame, i.e., the image frame is s by s where s is 0.7. The number of corresponding points is 6 over 2 frames. If the resolution of the camera is m x m then the image coordinates vertically and horizontally are digitized to m equally spaced intervals. In all synthetic data cases m was set to 512 pixels. Normally distributed random values were added to the image coordinates to simulate the effects of noise.

For synthetic and real image data, the elements of the stabilization value $\mathbf{W}$ for stabilizing the disparity parameters was set to 50 focal lengths (corresponding to a disparity standard deviation of 1/50 focal lengths). Disparity stabilization values at or near zero (i.e. no stabilization) resulted in a small decrease in performance for the scene and camera geometry standard scenario discussed in the following section. For the standard scenario the correct verification rate typically decreases by about 3 percent with the absence of stabilization. For the scenario which yields image data with large amounts of perspective distortion the lack of disparity stabilization caused the correct verification rate to decrease by about 10 percent. The stabilization values for the motion parameters were set to zero as it was found to be unnecessary to stabilize these parameters.

The linear algorithm supplies two initial estimates because of the ambiguity in the rotation sense for rotation in depth. The second initial estimate is only computed if the nonlinear estimator fails to verify potential rigidity given the first linear estimate.

Stopping Criterion For Convergence

The algorithm iterates until one of the following stopping criterion is met.

1. The norm of the parameter correction vector $\mathbf{h}$ is less than 10^{-2}.
2. The residuals do not decrease by more than the relative amount 10^{-3} over 2 successive iterations.
3. The number of iterations exceeds an upper bound of 10.

4. The residual is less than a threshold value determined from *a priori* knowledge of the observation variances. The threshold value is determined by the expression

$$T_E = k\sqrt{(N2m - n)\sigma^2}$$

where N is the number of images (2 for the results described here), k is a factor setting the confidence level and is set to approximately 2 for the Monte Carlo trials described below, σ^2 is the observation variance and is assumed to be equal for all measurements, m is the number of correspondences and n is the number of estimated parameters. This expression is from the unbiased estimator of the variance of the data for a least squares estimator. The factor of $N2m - n$ is the number of degrees of freedom for the estimator. This follows from an interpretation of the estimator as determining the pose of the object in the N camera frames relative to the object defined in a world coordinate frame whose 3D structure must also be determined. Hence, each observation contributes 2 constraints (vertical and horizontal position) for each of the N images. Now, by identifying the origin and orientation of the world coordinate frame with one of the camera frames, the pose transformation for that (reference) camera frame is the identity transformation $\mathbf{R} = \mathbf{I}$ and $\mathbf{t} = \mathbf{0}$. Thus, the number of pose (or motion) parameters to be estimated is reduced by 6. Now, the actual number of independent constraints is reduced by 2m which is the number of constraints from the reference image since the structure is dependent on these observations by the back-projection calculation. This dependency reduces the number of structure parameters for each point from 3 to 1. The number of estimable depth parameters is $m - 1$ since one of the depths is fixed to set the scale of the solution. Thus, the total number of estimable parameters is

$$(N-1)6 + (m-1)$$

which is the value bound to the variable n in the expression above for T_E. The total number of independent constraints is $(N-1)2m$. However, the total residual error upon convergence is contributed to by the $N2m$ measurements for N views since the total residual error for the correct parameter values can be written as

$$\sqrt{\sum_{j=0}^{N-1}\sum_{i=0}^{m-1} \| \mathbf{w}_{ij} - \mathbf{y}(\mathbf{x}_{i0}, \mathbf{m}_j) \|^2}$$

where $\mathbf{m}_0$ is the identity pose transformation and correct depth parameter set for the reference frame, $\mathbf{x}_{i0}$ is the i$^{\mathrm{th}}$ noise free image point from the projection of the correct scene point in the reference frame, $\mathbf{w}_{ij}$ is the i$^{\mathrm{th}}$ noisy observed image point in the j$^{\mathrm{th}}$ frame, and $\mathbf{m}_j$ is the correct pose and scene depths for the j^{th} frame relative to the zero$^{\mathrm{th}}$ frame. This residual error has a total of $N2m$ terms, which is the value used to compute T_E.

5.1 Monte Carlo Simulation

The rigidity checking algorithm requires the specification of the global offset of the intermediate camera (object) frame from the camera centered frame. Its value was set to 2 focal lengths for reasons discussed earlier.

Monte Carlo Results

The standard scenario for camera and scene geometry. 10,000 trials for the following scenario were run. The scene consists of 6 feature points. Translation was uniformly distributed between -500 and 500 focal lengths. Rotation about the optical axis was uniformly distributed in the interval ± 180 degrees and rotation in depth was uniformly distributed in the interval ± 90 degrees. Object size was in the range 10 to 5000 focal lengths with the closest object point 2 to 5000 focal lengths away from the camera with both variables uniformly distributed in their respective ranges. The motion is applied to the points in the object centered frame. The global offset that defines the scene coordinates in the object centered frame is a uniformly distributed random variable in the range 2 to 5000 focal lengths. Image noise was simulated by adding normally distributed random values to the exact image coordinates with a zero mean and a variance of 1 pixel.

On average, 1000 trials required a total of 3.9 seconds to verify at a convergence rate of 97.9 percent with double precision floats on a Sparc 10 processor. The threshold T_E

was computed for an observation variance of 1 pixel for all correspondences. The nonlinear estimation algorithm is typically bypassed for 50 to 60 percent of the trials for the standard scenario described above because the residual for the linear algorithm is below the specified threshold. If bypassing the nonlinear estimator is prevented then the time would increase proportionately.

The receiver operating characteristic (ROC) curve is given in figure 1 for three image noise levels. The curves for increasingly noisy observations shift progressively to the right. 100,000 trials for rigid configurations and 100,000 trials for nonrigid configurations were run to generate each curve. The same camera and scene geometry random variables described above were used. The noise is normally distributed with a mean of 0 and a standard deviation of 1, 2 and 3 pixels. For the nonrigid trials image data was generated randomly for the two views. The convergence rate for a separate test running 10,000 nonrigid trials was 1.3 percent taking an average of 13.6 seconds to complete 1000 trials for the same threshold, T_E, used for the rigid trials.

It is important to note that the nonlinear estimator is never bypassed for the ROC trials, since it was desired to assess the end to end performance of the method. For the rigid trials, experiments show that if the linear estimator's residual error was below the rigidity threshold, T_E, than it was almost always the case that the nonlinear estimator's residual value was also below T_E upon convergence. For the nonrigid trials, the nonlinear estimator almost never converged with a residual value below T_E if the linear estimator's residual was below T_E. The performance improvement from not bypassing the nonlinear estimator for nonrigid trials amounts to a decline of about 0.2 percent in the false positive rate which coincides with the 0.2 percent proportion of nonrigid trials where the linear estimator's residual error was below T_E. The trade-off clearly favors bypassing the nonlinear estimator when the linear estimator's residual is below T_E.

Images with large perspective distortions. Figures 2 and 3 are ROC plots for images with large amounts of perspective information. The closest scene point is now fixed in the range 2 to 50 focal lengths while the remaining points lie in a depth range between 10 and 5000 focal lengths from the closest depth. Some of the samples from this scenario may

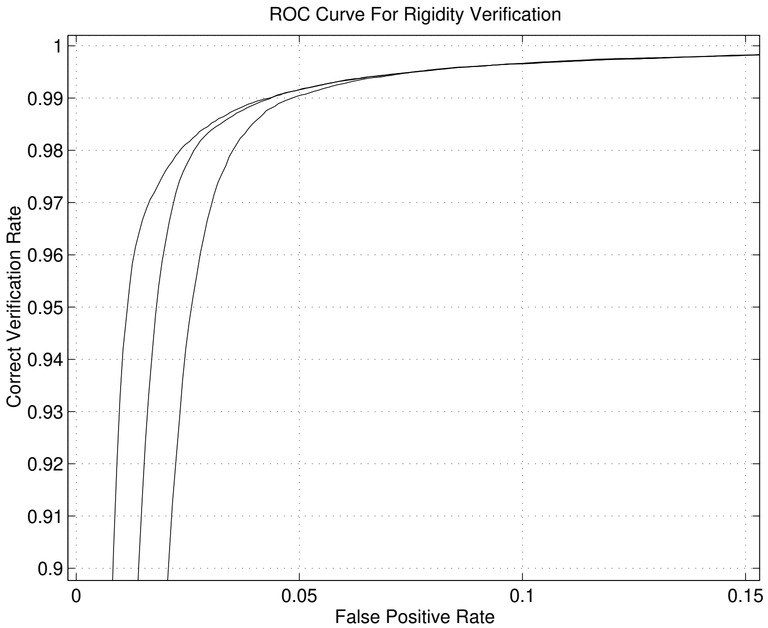

Figure 1: Receiver operating characteristic (ROC) for noisy observations with a variance of 1, 4 and 9 pixels. Motion and structure variables are for the standard scenario. Close up of the curve's knee. The curve for the noisiest data is furthest to the right.

be somewhat unrealistic since some of the points probably would not be in focus given a real camera with a 40 degree field of view (the value used here). Added noise has a variance of 1. 100,000 trials for rigid and nonrigid configurations were run for this curve. Approximately 15 percent of the rigid object trials resulted in the linear estimator's residual error falling below the rigidity threshold for a noise variance of 1 pixel. Recall, however, that the nonlinear estimator was not bypassed for these ROC plots.

5.2 Real image sequence

Two images of a Lego object on a motion table were taken by a monochrome camera with a 480 by 512 pixel image. The object's center was approximately 13 inches from the camera's projection center. The image magnification factor was determined to be 609 pixels. The total motion between the two views was a translation of 2.5 inches and a rotation of 15 degrees.

Figure 4 is the first frame marked with seven manually determined feature points. Figure

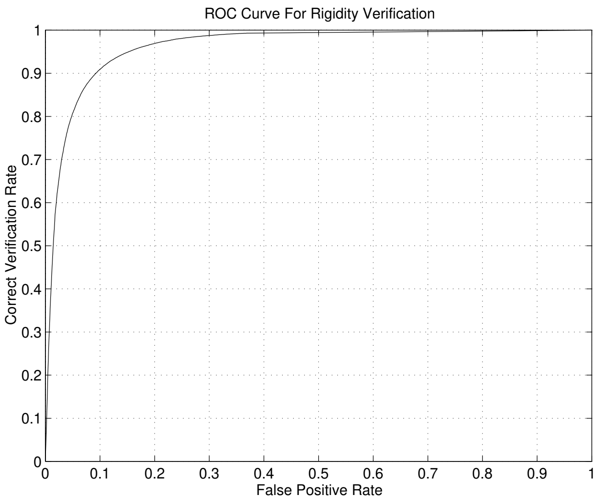

Figure 2: Receiver operating characteristic (ROC) for noisy observations with a variance of 1 pixel. Large perspective distortion.

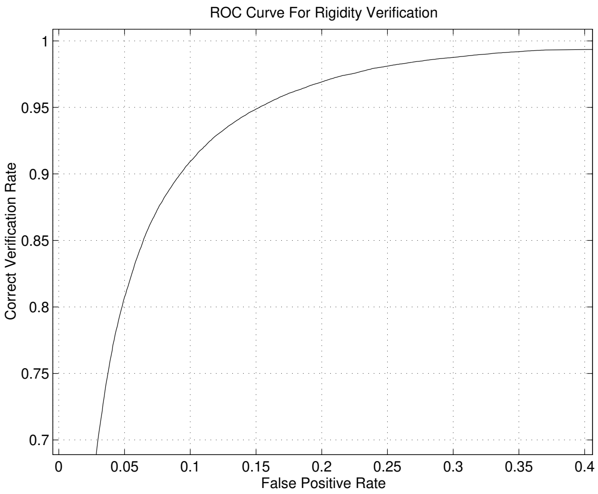

Figure 3: Receiver operating characteristic (ROC) for noisy observations with a variance of 1 pixel. Close up of the curve's knee. Large perspective distortion.

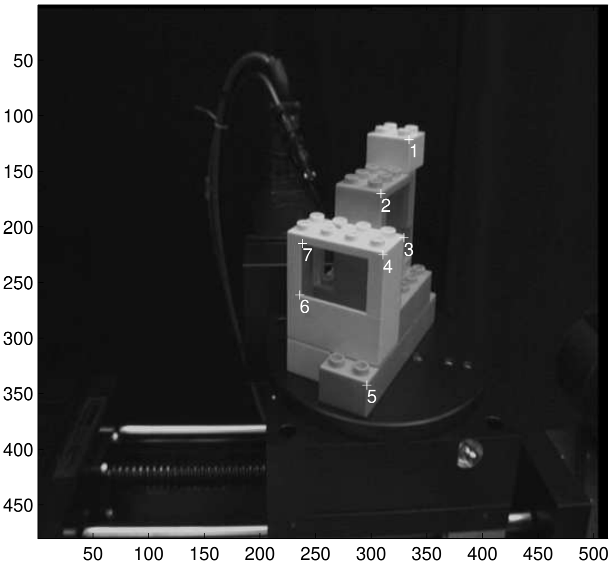

Figure 4: First frame of Lego sequence with labeled correspondences.

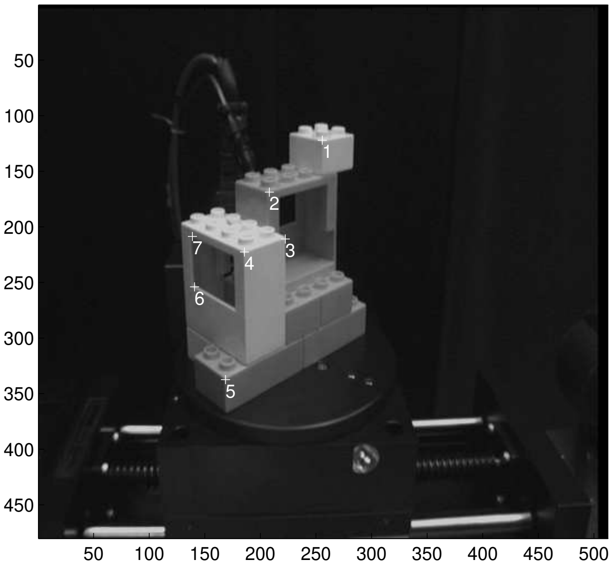

Figure 5: Second frame of Lego sequence with labeled correspondences.

5 is the second frame with the corresponding points marked.

As with the synthetic data the global offset is set to 2 focal lengths. For the correctly matched correspondences the nonlinear estimator was bypassed since the residual error from the linear estimator was below the threshold for an assumed noise variance of 1 pixel. This is a strong indication that the object projection is approximately orthographic.

For the first experiment only the correspondences labeled one to six in both views are considered. Given the six correspondences in the first frame, the same label set for the correspondences in the second image was permuted and verified. Figure 6 shows a portion of the bar graph of sorted residual errors for the 720 possible permutations of the label set in the second image. The figure shows the number of permutations that yielded residual errors below 10 pixels. For an expected noise variance of 1 pixel, 34 incorrect correspondence label sets fall below the rigidity threshold, a 4.7 percent false positive rate. The linear estimator's residual for the correct correspondence was 1.0 pixels. Only one of the incorrect correspondences yielded a residual below this value. The permutation with the lowest residual value of 0.77 pixels is (5,3,4,6,1,2).

For the second experiment all seven correspondences from both views are considered. Given the seven correspondences in the first image, the same label set for the correspondences in the second image was permuted and verified. Figure 7 shows a portion of the bar graph of sorted residual errors for the 5040 possible permutations of the label set in the second image. The figure shows the number of permutations that yielded residual errors at or below 10 pixels. For an expected noise variance of 1 pixel, 24 incorrect correspondence label sets fall below the rigidity threshold, a 0.48 percent false positive rate. The correct correspondence yielded the lowest residual of 1.0 pixels. The next highest residual is 1.36 pixels for the permutation (1,2,7,4,5,6,3), i.e., points 3 and 7 swapped. The epipolar lines for these two points appear to be in close proximity and therefore a mismatch between them could lead to a low residual solution.

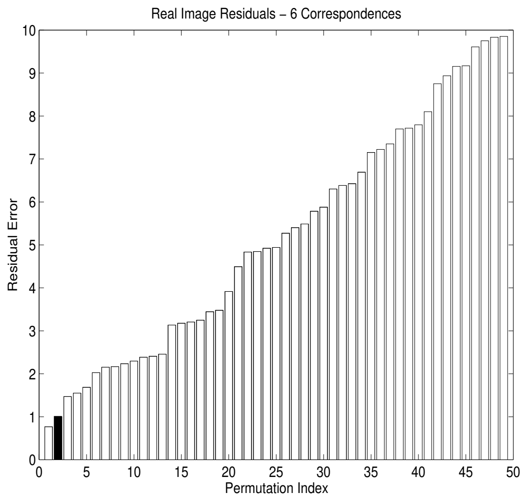

Figure 6: Bar graph of sorted residual error for permutations of six corresponding points. Only the lowest residual values are shown, approximately 10 pixels and less. For an expected noise variance of 1 pixel, 35 permutations out of 720 (4.9 percent) are classified as potentially rigid. Filled bar is correct correspondence.

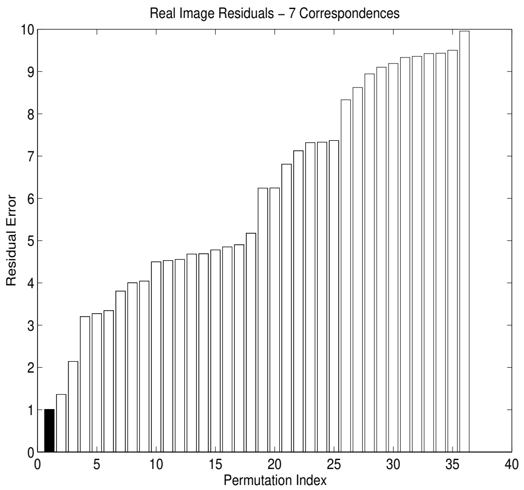

Figure 7: Similar to the previous figure, but with seven correspondences. For an expected noise variance of 1 pixel, 25 permutations out of 5040 (0.5 percent) are classified as potentially rigid. Filled bar is correct correspondence.

6 A Comparison To Other Methods

6.1 Horn's Algorithm For Relative Orientation

A version of Horn's algorithm for determining the rotation and translation between a pair of perspective views of a scene is parameterized in terms of quaternions [12]. The objective function is based on the well known coplanarity constraint for corresponding points and is formulated in terms of a scalar triple product appropriately weighted by a term based on the estimated observation variances. It is interesting to note that Weng *et. al.*'s epipolar improvement method [39] is based on the same objective function for a different parameterization. Horn's algorithm was implemented in Matlab and tested. Figure 8 is the ROC curve for Horn's method compared to the rigidity checking method for three different image noise levels and figure 9 is a close up of the knee of the curves. The same standard scenario was used as described earlier. Linear estimates are provided as initial guesses to Horn's iterative algorithm. This contrasts with Horn's approach which involves running a large number of initial guesses to find all of the solutions. Ten iterations maximum were allowed for each initial estimate. 10,000 trials for both rigid and nonrigid 3D configurations were run for each of Horn's ROC curves.

The high false positive rate exhibited by Horn's method results from the property of the epipolar constraint that allows for a larger space of motion solutions which are determined independently of the physical plausibility of the corresponding structure estimate. In addition, the potential ambiguity of the motion estimate for nearly orthographic images or noisy image measurements also contributes to the algorithm's weaker performance. A discussion of the drawbacks of the epipolar constraint for rigidity verification can be found in the earlier section on related work.

In contrast to epipolar constraint based methods, the rigidity checking method converges to fewer local minima with low residual values primarily because the collinearity constraint formulation makes use of all the image measurement information. Rigidity checking searches a rigid transformation space while implicitly enforcing a simultaneously consistent and physically plausible depth estimate. Deviation from rigidity is reliably signalled by a large residual error.

Of the three sets of 10,000 nonrigid trials run on Horn's algorithm, approximately 55 percent of the trials in each set were discarded because the "depths positive" criterion was violated. If this criterion was not applied the false positive rate would, in fact, be much higher. This consideration accounts for the blip in the ROC curve at around the 0.45 false positive rate.

6.2 Comparison to Essential Matrix Methods

Wei *et al.* employ the linear formulation developed by Tsai and Huang [34] to solve the correspondence problem and estimate motion simultaneously. Preliminary Monte Carlo testing with a Matlab implementation of their algorithm indicates a large false positive rate for the standard Monte Carlo scenario described earlier. Monte Carlo testing for nine corresponding points and a noise variance of one pixel yielded a true positive rate of approximately 36 percent versus 99 percent for the rigidity checking method and 85 percent for Horn's method at

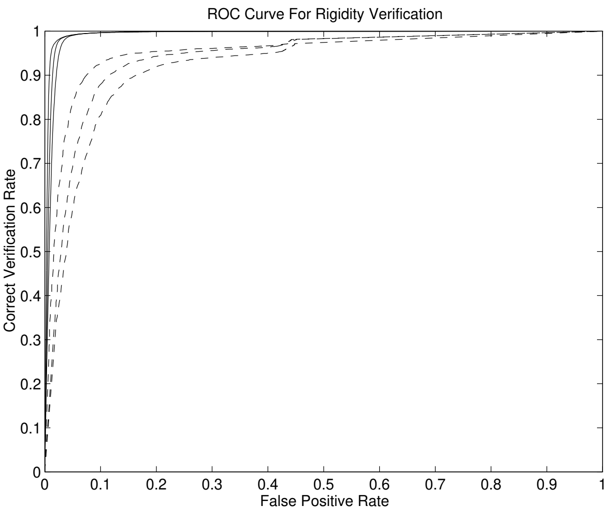

Figure 8: A comparison with Horn's method. ROC for noisy observations with a variance of 1, 4, and 9 pixels. Horn's method are the 3 dashed curves. The curve for the noisiest data is furthest to the right for each method. Motion and structure variables are for the standard scenario.

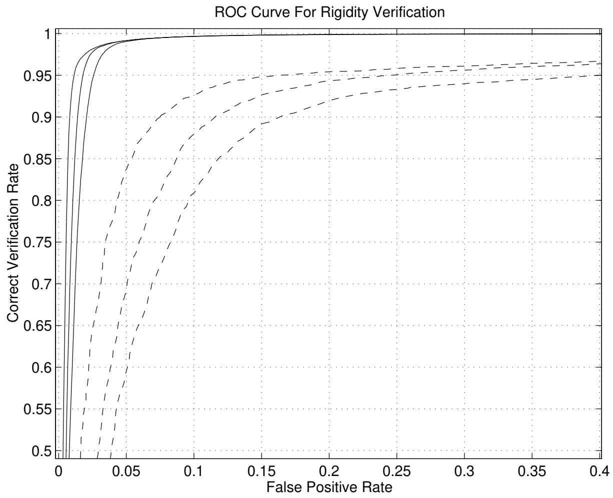

Figure 9: A comparison with Horn's method. Close up of the knee of the ROC curves from the previous figure.

a false positive rate of 5 percent. The performance contrast is greater still considering that the results for Horn's method and the rigidity checking method are for only six correspondences. Note that Wei *et al.* do not check if the motion solution corresponds to a "depths positive" reconstruction. Such a check would improve their performance.

7 Applications Of Rigidity Checking

7.1 Recognition From a Single View

Kontsevich discusses the application of his linear structure-from-motion algorithm for weak perspective images to the problem of correspondence checking. As a corollary to the "viewpoint consistency constraint" [22] or "generic view assumption" [7][37], he states the following assumption for pairwise comparisons:

> If, for some pair of projections, correct [rigid] correspondence exists, the projections are views of the same object.

A relation is established, then, between verification of potential rigid correspondence, scene structure and object recognition. The relation between correspondence and structure referred to by Kontsevich as the "structural theory" (citing earlier work due to Ullman and, Grzywacz and Yuille [9]) describes the mutually supporting processes of correspondence and 3D interpretation that operate simultaneously. The theory and performance of the rigidity checking method can be viewed as a significant improvement in the formulation of the rigidity component of the "structural theory" hypothesis of Grzywacz and Yuille and extends the implementation of the hypothesis to images formed under full perspective projection by building on the work of Kontsevich. Object recognition can be viewed as a registration problem between a model of the object and a representation of the object derived from a novel view or set of views of the object. The different approaches to recognition lie in the representation of the object's model, the representation of the object from the novel view(s) and the matching or registration process used to verify the presence of the object.

The significance of the relation between a recognition process for point configurations and structure-from-motion has been noted by Ullman and Basri [36] and Shashua [29]. Ullman and Basri describe a linear relation between two model views of an object and a novel view under scaled orthographic projection where the transformation between views can be modeled by a general linear transformation. Under a similarity transformation between views and orthographic projection, three model views are required for recognizing a rigid object. They note that it is possible to recover 3D structure and motion based on three orthographic views using the linear equations they derived under a similarity transformation between views. However, they presuppose the existence of correspondences when performing recognition. They also present a novel scheme for performing recognition by linear combinations of models using subsets of corresponding points that avoids the necessity of point to point correspondence [36](1989 technical report). The idea of using the alignment constraints to simultaneously verify correspondences was not presented. However, they echo the theme quoted above when they comment in their conclusion that "The linear combination scheme reduces the recognition problem in a sense to the problem of establishing correspondence between the viewed object and candidate models" [36](1989 technical report). Shashua similarly draws a comparison between recognition and structure-from-motion for objects undergoing an affine transformation in space. He is able to recover full correspondence between two orthographic views with at least four point correspondences by utilizing the brightness constancy constraint for infinitesimal motion. Shashua's results are applied to the recognition problem under the paradigm of generating novel views from one or more model views and verification by alignment.

Bennett *et al.* derive recognition polynomials that assume point correspondences have been established between a single 2D previous view of an object and a novel view under orthographic projection [4]. Their polynomials can be constructed for different transformations between views, e.g., similarity or affine. They have not, however, derived a polynomial for perspective projection although they claim that, in principle, such a polynomial can be constructed. Their method, which only requires one model view, contrasts with Ullman and Basri's method which requires at least two model views or Shashua's method which assumes brightness constancy. Their method could be used to establish correspondences and this is implied when they discuss the extraction of rigid configurations from image sequences.

The rigidity checking method has characteristics common to the methods described above. Like Ullman and Basri and Shashua the rigidity checking scheme is based on an alignment paradigm that makes a verification decision based on the distance between corresponding features in a common image frame. The rigidity checking method, however, can extend this recognition paradigm to images formed under full perspective projection as well as handling scaled orthographically projected images. Rigidity verification is a profound matching constraint useful for recognition purposes. In combination with other matching conditions, a reliable system for verifying point correspondences could be devised which should prove to be an effective component in a system for recognizing 3D objects from a single previous view.

7.2 Motion Segmentation And Stereo Correspondence

Motion Segmentation. The problem of detecting multiple moving objects from a sequence of images taken by a moving camera has been addressed by a variety of methods. The methods cited below depend on a matching condition that is defined by the epipolar constraint. The rigidity checking method is based on a structure-from-motion matching condition and could be used as an alternative module for testing the consistency of hypothesized point matches.

Thompson *et al.* look for outliers using least median squares to segment points inconsistent with an orthographic structure-from-motion constraint [33]. Nishimura *et al.* also use the epipolar constraint to segment out differently moving objects under weak perspective projection [28]. They use a Hough transform method to detect clusters in a scaling and frontoparallel rotation space which correspond to the differently moving objects.

The approach of Thompson *et al.* exploits the rigidity constraint under orthographic projection to find sets of points that are inconsistent with the motion of the camera relative to the static environment. They use Huang and Lee's algorithm [14] for structure-from-motion as the matching condition for a set of hypothesized correspondences provided by a separate frame-to-frame point matching and tracking process. Points that are inconsistent with the rigid interpretation are segmented out as determined by the residual error in Huang and Lee's structure-from-motion constraint. The actual structure and motion estimate are not important: only the reliability of the algorithm's response to outliers is required to ensure the detection of such outliers. They use least median squares to detect the outliers which they admit is a computationally intensive method.

The rigidity checking method would fit well into their approach. Like Thompson, attention is not paid to the actual motion and structure estimates but only relies on the residual error of the matching criterion. Rigidity checking would extend the domain of input images from those formed by weak perspective projection to full perspective projection. Similarly, rigidity checking could be substituted for Nishimura *et al.*'s scaled orthographic epipolar constraint that is used to verify the rigidity of point correspondences.

Stereo Correspondence. In stereo correspondence, the use of the epipolar constraint reduces the search space to a one dimensional search along the corresponding epipolar lines. In the absence of extrinsic camera calibration, the problem of stereo correspondence is equivalent to the 2D motion correspondence problem. Hence, if the epipolar geometry is unknown or poorly estimated, then the rigidity checking method would be a suitable module for disambiguating true from false matches and together with other binocular matching rules should prove to be a reliable approach to solving the stereo correspondence problem.

8 Conclusions

An algorithm has been described that reliably and rapidly verifies the potential rigidity of three dimensional point correspondences from a pair of two dimensional views under perspective projection. The method, which we call rigidity checking, is useful for finding corresponding point features between two or more views of an object. The rigidity decision is based on the residual error of an integrated pair of linear and nonlinear structure-from-motion estimators. The matching condition is based on a set of 3D recovery equations derived from the collinearity condition of points under perspective projection. This choice for the 3D recovery model contributes significantly to the performance improvement of the algorithm relative to other methods because, unlike recovery based on the epipolar constraint, the collinearity condition uses all of the information in the image measurements. This improvement in performance comes at a small additional cost in computational complexity due to the choice of parameterization. In Monte Carlo simulations over the entire set of rigid and nonrigid trials, a single trial took in the order of 10 milliseconds to execute on a Sparc 10 processor. In summary, rigidity checking works with as few as six corresponding points under weak or full perspective projection, handles correspondences from widely separated views, makes full use of the disparity of the correspondences, and is integrated with an initial parameter estimator based on a linear weak perspective algorithm. Results from extensive Monte Carlo simulations and from real images were presented. A comparison was made with methods based on the epipolar constraint such as Horn's nonlinear algorithm for structure-from-motion and Wei *et al.*'s linear method that illustrated the disadvantages of the epipolar constraint as a matching condition. Applications of this algorithm as a module for performing rigidity checking are numerous; 3D recognition from a single previous view, motion segmentation and stereo correspondence were briefly discussed.

Public Availability Of Implementation

This algorithm has been implemented in C and is freely available by anonymous FTP from ftp.cs.ubc.ca in directory pub/local/danm.

References

[1] Anstis, S.M., The perception of apparent motion, *Phil Trans. R. Soc. London B 290*, 153-168, 1980.

[2] Basri, R., On the uniqueness of correspondence under orthographic and perspective projections, *A.I. Memo No. 1333*, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, December 1991.

[3] Bennett, B.M., Hoffman, D.D., Nicola J.E., Prakash C., Structure from two orthographic views of rigid motion, *J. Opt. Soc. Am. A*, 6(7), 1052-1069, 1989.

[4] Bennett, B.M., Hoffman, D.D., Prakash, C., Recognition Polynomials, *J. Opt. Soc. Am. A*, 10(4), 759-764, April 1993.

[5] Braddick, O.J., Low-Level and high-level processes in apparent motion, *Phil Trans. R. Soc. London B 290*, 137-151, 1980.

[6] Braunstein, Myron L.,Hoffman, Donald D., Pollick, Frank E., Discriminating rigid from nonrigid motion: minimum points and views,*Perception and Psychophysics*, 47(3), 205-214, March 1990.

[7] Freeman, W.T., Exploiting the generic view assumption to estimate scene parameters, *Proc. 4th International Conference on Computer Vision (ICCV 93)*, 347-356, 1993.

[8] Frisby, J.P., Pollard, S.B., Computational issues in solving the stereo correspondence problem, in *Computational Models of Visual Processing*, Landy, M.S., Movshon, J.A., eds., The MIT Press, Cambridge, MA, 1991.

[9] Grzywacz, N.M., Yuille, A.L., Massively parallel implementations of theories for apparent motion, Spatial Vision, Vol. 3, No. 1, 15-44, 1988.

[10] Harris, Chris., Structure-from-motion under orthographic projection, *Proc. of 1st European Conference on Computer Vision (ECCV 90)*, 118-123, 1990.

[11] Harris, C.G., Pike, J.M., 3D positional integration from image sequences, *Image And Vision Computing*, Vol. 6, No. 2, 87-90, 1988.

[12] Horn, B.K.P., Relative orientation revisited, *J. Opt. Soc. Am. A*, 8(10), 1630-1638, Oct 1991.

[13] Hu, X., Ahuja, N., Motion estimation under orthographic projection, *IEEE Transactions on Robotics and Automation*, Vol. 7, No. 6, 848-853, 1991.

[14] Huang, T.S., Lee, C-H., Motion and structure from orthographic projections, *IEEE Trans. PAMI*, Vol. 11, No. 5, 536-540, 1989.

[15] Huttenlocher, D.P., Kleinberg, J.M., Comparing point sets under projection, *ACM-Symposium on discrete algorithms*, 1-7, 1994.

[16] Koenderink, J.J., van Doorn, A.J., Affine structure-from-motion, *J. Opt. Soc. Am. A*, 8(2), 377-385, Feb 1991.

[17] Kontsevich, L.L., Pairwise comparison technique: a simple solution for depth reconstruction, *J. Opt. Soc. Am. A*, 10(6), 1129-1135, Jun 1993.

[18] Kumar, R.V.R., Tirumalai, A., Jain, R.C., A non-linear optimization algorithm for the estimation of structure and motion parameters, *IEEE Computer Society Conference On Computer Vision and Pattern Recognition (CVPR 89)*, 136-143, 1989.

[19] Lee, C-H., Huang, T., Finding point correspondences and determining motion of a rigid object from two weak perspective views, *Comp. Vis. Graph. and Image Processing* 52, 309-327, 1990.

[20] Levenberg, K., A method for the solution of certain nonlinear problems in least squares, *Quart. Appl. Math.*, vol. 2, pp. 164-168, 1944.

[21] Longuet-Higgins, H.C., A computer program for reconstructing a scene from two projections, *Nature*, vol. 293, 133-135, Sept 1981.

[22] Lowe, D.G., *Perceptual organization and visual recognition*, Kluwer, Boston, MA, 1985.

[23] Lowe, D.G., Fitting parameterized three-dimensional models to images, *IEEE Trans. PAMI*, vol. 13, no. 5, 441-450, 1991.

[24] Marquardt, D.W., An algorithm for least squares estimation of nonlinear parameters, *SIAM J. of applied math*, vol. 11, no. 2, pp. 431-441, 1963.

[25] Marr, D., *Vision: a computational investigation into the human representation and processing of visual information*, Freeman, San Francisco, Calif., 1982.

[26] McReynolds, D.P., *Solving for relative orientation and depth*, MSc thesis, Department of Computer Science, University of British Columbia, October 1988.

[27] McReynolds, D.P., Determining the motion of a remotely piloted vehicle from a sequence of images, *IEEE Proceedings Of The National Aerospace And Electronics Conference (NAECON 89)*, vol. 3, 1097-1104, 1989.

[28] Nishimura, E., Xu, G., Tsuji, S., Motion segmentation and correspondence using epipolar constraint, *Asian Conference on Computer Vision (ACCV 93)*, 199-204, 1993.

[29] Shashua, A., Correspondence and affine shape from two orthographic views: motion and recognition,*A.I. Memo No. 1327*, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, December 1991.

[30] Sinclair, D., Motion segmentation and local structure, *Proc. 4th International Conference on Computer Vision (ICCV 93)*, 366-373, 1993.

[31] Sudhir, G., Banerjee, S., Zisserman, A., Finding point correspondences in motion sequences preserving affine structure, *Proceedings British Machine Vision Conference 1993 (BMVC93)*, 359-368, 1993.

[32] Szeliski, R., Kang, S.B., Recovering 3D shape and motion from image streams using non-linear least squares, *IEEE computer society conference on computer vision and pattern recognition (CVPR 93)*, 752-753, 1993. Also published as *CRL 93/3*, Cambridge Research Lab, Digital Equipment Corp., March 1993.

[33] Thompson, W.B., Lechleider, P., Stuck, E.R., Detecting moving objects using the rigidity constraint, *IEEE Trans. PAMI*, 15(2), 162-166, Feb 1993.

[34] Tsai, R.Y., Huang, T.S., Uniqueness and estimation of 3-D motion parameters of rigid bodies with curved surfaces, *IEEE Trans. PAMI*, 6(1). 13-27, 1984.

[35] Ullman, S., *The interpretation of visual motion*, MIT Press, Cambridge, Mass., 1979.

[36] Ullman, S., Basri, R., Recognition by linear combinations of models, *IEEE Trans. PAMI*, 13(10), 992-1006,1991. Originally published as technical report *A.I. Memo No. 1152*, MIT, Aug 1989.

[37] Wagemans, J., Perceptual use of nonaccidental properties, *Canadian Journal of Psychology*, 46:2, 236-279, 1992.

[38] Wei, G.-Q., He, Z., Ma, S.D., Fusing the matching and motion estimation of rigid point patterns, *Proceedings 1990 IEEE International Conference on Robotics and Automation*, Cincinnati, OH, USA, 2017-22 vol.3, 13-18 May 1990.

[39] Weng, J., Ahuja, N., Huang, T.S., Optimal motion and structure estimation,*IEEE Trans. PAMI*, vol.15, no.9,pp. 864-884, 1993.

[40] Weng, J., Huang, T.S., Ahuja, N., *Motion and structure from image sequences*, Springer series in information sciences 29, Springer-Verlag, 1993.