USC-SIPI REPORT # 127

A Spatio-Temporal Approach to Motion Understanding

By

Min Shao and Rama Chellappa

8 July 1988

Signal and Image Processing Institute
Dept. of Electrical Engineering-Systems
PHE 324
University Park, MC-0272
Los Angeles, CA 90089, U.S.A.

Tel. #: (213) 743-8559

Address for correspondence: Professor Rama Chellappa at the above address.

Partially supported by the NSF Grant # IRI-87-13585.

A Spatio-temporal Approach To Motion Understanding

Min Shao and Rama Chellappa

Abstract

Algorithms for the interpretation of optical flow are difficult to design mainly due to the nonlinearity of constraint equations and the high dimensionality of the parameter space. Here we show that when two velocity fields from the same moving object are given, the rotational component of the motion parameters can be eliminated from the difference velocity field. Thus the translational component, or the focus of expansion (FOE) can be robustly found by solving a set of linear equations. This in turn facilitates closed-form solutions for the rotational component and environment depth. This approach can be applied to multi-object motion segmentation using the Hough transform. If a dense sequence of images is available, then the structure of the environment and the 3-D motion parameters can be recovered directly at every image point from the given velocity filed. In this approach both spatial and temporal information are used in a uniform way. The structure-from-motion (SFM) problem is then reduced to solving a quadratic equation. If the optical flow field is not available, the SFM problem based directly on the first-order derivatives of the image brightness is underdetermined. However, by exploiting the image brightness constancy constraint in both spatial and temporal domains we show that, given the first and second order spatio-temporal derivatives of image brightness, the SFM problem becomes overdetermined.

1 Introduction

The problem of recovering the structure of the environment and the 3-D relative motion between a sensor and the environment from the 2-D image data has been explored by many researchers in the area of computer vision. It is known as the *structure-from-motion* (SFM) problem. There are basically three different approaches to the SFM problem in the literature, namely, the discrete, the continuous, and the direct approach.

In the discrete approach [1, 2, 3], a finite number of well-separated features such as points, lines or contours are extracted and matched in a sequence of images. The displacements of these features between successive image frames are used to estimate motion parameters and depth of the environment at these features. It has been observed [1] that this approach tends to be unstable in the presence of even small amount of noise. As only a very sparse set of features are used, the depth of the environment obtained using the feature-based approach is also very sparse. The most difficult problem associated with this approach is the correspondence problem.

The continuous approach depends on the computation of apparent velocities, or optical flow, of brightness patterns in an image before the motion analysis begins [4, 5]. In principle, given the optical flow at five points, the motion parameters can be recovered. However, algorithms for the SFM problem are difficult to design due to mainly the nonlinearity of the constraint equations and the high dimensionality of the parameter space. Direct solutions are iterative in nature and good initial guesses are required to make the solution numerically stable. Most studies therefore seek to develop strategies where the dimensionality and /or nonlinearity are reduced.

As several authors have noticed, temporal information is as important as spatial information in a sequence of images. Bolles and Baker [6] used a solid of data called *spatio-temporal data*, with time as the third dimension, to compute 3-D locations of world features. It is constructed from a dense sequence of images-images taken close enough that none of the image features move more than a pixel or so. In this case the correspondence problem becomes trivial if the camera motion is known. Subbarao [7] and Wu [8] presented a formulation and solution procedure for reconstructing the structure of the environment and the motion parameters from the first-order spatio-temporal derivatives of the optical flow. It was pointed out that first-order temporal derivatives of the motion field is relatively more robust than its second-order spatial derivatives. But their analyses were restricted only to a small field of view around the line-of-sight.

In this paper we propose two methods to make the SFM problem linear and numerically efficient and robust. We first discuss how to use two velocity fields from the same environment taken at different time instants to decouple the rotational component from the optical flow field. We show that the difference of the two velocity fields does not contain the rotational component. Because the difference is taken at the same image location no correspondence is necessary. The ratio of the two components of the difference velocity field at each image location provides one linear constraint equation on the focus of expansion (FOE). Thus two or more image points on the same rigid body uniquely determine the FOE. By using the Hough transform one is able to segment the image into individual objects and compute the FOE for each rigid body.

Another way to linearize the SFM problem is using local derivatives of the optical flow fields. If we are able to estimate the first-order spatio-temporal derivatives of the velocity

filed, we can recover not only the motion parameters and the depth of the environment, but also the orientation of the underlying surfaces at each point. We thus unify the analysis of spatial and temporal information and obtain closed form solutions at each point. This is most desirable if one wants to consider non-rigid motion problems. As we will see, the results obtained by Subbarao [7] and Wu [8] are only first-order approximations to our solutions around the line of sight. Our methods are more efficient because we do not have to transform the optical flow field at each pixel.

The difficult problem associated with our approach described above is the assumption that optical flow fields are known . The direct approach [9, 10] bypasses the computation of optical flow and directly utilizes the derivatives of image brightness to estimate the motion parameters. Negahdaripour and Horn [9] exploited the spatial gradient and the time rate of change of brightness over the whole image and explicitly imposed the positive depth constraint to recover the FOE. As the problem of determining the motion parameters and the surface structure from a single brightness constancy equation is an underdetermined problem, this approach is successful only in some cases.

In our work we exploit the brightness constancy constraint in both spatial and temporal domain. The first and second order spatial and temporal derivatives of the image brightness provide four constraint equations for the motion parameters and the surface structure. At every image pixel there are three local unknowns, namely, the depth, and the orientation of the surface, and there are six global motion parameters. Thus by simply counting the number of constraints and the number of unknowns, it is easy to see that only the derivatives of the brightness at five different points are needed to recover the motion parameters (the magnitude of the translational component can not be recovered). If more points are used

the results will be more accurate. Unfortunately, these constraint equations are highly nonlinear and we have not been able to find closed form solutions.

The organization of this paper is as follows. General rigid body motion models are set up in Section 2. We then propose two linear methods in Section 3 to solve for the motion parameters and the structure of the environment from optical flow fields. The problem of recovering the motion and structure of the environment without explicitly computing the optical flow is considered in Section 4. Preliminary experimental results are presented in Section 5, which is followed by a summary in Section 6.

2 General Rigid Body Motion Models

In this paper we consider the relative motion between a sensor (usually a camera) and the environment. The environment may contain several moving objects. Each object is rigid and does not have to be separated from others a priori. The problem we are interested in is recovering the 3-D structure of the environment (depth and orientation relative to the camera) and the motion parameters of each individual rigid body from a sequence of time-varying images. For each independent object in the environment we consider the equivalent problem of a stationary object and a moving, monocular pin-hole camera represented by the spatial coordinate system (X, Y, Z), see Figure 1. The origin of this system is located at the vertex of the perspective projection, and the Z-axis is directed along the line-of-sight. The 2-D image sequence is created by the perspective projection of the objects onto an image plane. The focal length, from the nodal point to the image plane, is assumed to be known and, without loss of generality, normalized to 1. The origin of the corresponding coordinate

system on the image plane is located at $(0,0,1)$, and its x and y axis are parallel to X and Y axis respectively. Thus the perspective projection (x,y) on the image of a point (X,Y,Z) in the environment is:

$$x = \frac{X}{Z}, y = \frac{Y}{Z} \tag{1}$$

The instantaneous rigid body motion of the camera system can be decomposed into two components: translation $\underline{V} = (V_X, V_Y, V_Z)$ and rotation $\underline{\Omega} = (\Omega_X, \Omega_Y, \Omega_Z)$. In this paper we assume that the motion is uniform, that is, $\underline{V}, \underline{\Omega}$ remain constant for a short period of time.

Let $\underline{R}$ be the position vector of some point on the object, with camera coordinates given by (X,Y,Z). Due to the camera's motion, the point in space moves with relative velocity $\frac{d\underline{R}}{dt} = -(\underline{V} + \underline{\Omega} \times \underline{R})$. In component form one can write:

$$\begin{cases} \frac{dX}{dt} &= -V_X - \Omega_Y Z + \Omega_Z Y \\ \frac{dY}{dt} &= -V_Y - \Omega_Z X + \Omega_X Z \\ \frac{dZ}{dt} &= -V_Z - \Omega_X Y + \Omega_Y X \end{cases} \tag{2}$$

The corresponding projection (x,y) on the image plane of the point moves with a velocity (u,v), where [4]

$$\begin{cases} u(x,y) = (xV_Z - V_X)/Z + [xy\Omega_X - (1+x^2)\Omega_Y + y\Omega_Z] \\ v(x,y) = (yV_Z - V_Y)/Z + [(1+y^2)\Omega_X - xy\Omega_Y - x\Omega_z] \end{cases} \tag{3}$$

We will refer to (u,v) as the optical flow field or the velocity field interchangeably. The point $(V_X/V_Z, V_Y/V_Z)$ is called the focus of expansion (FOE). The underlying surface is described by the function $Z(X,Y)$, or equivalently by $Z(x,y)$, where Z is the depth of the

surface from the camera. The orientation of the surface relative to the camera is given by

$$\begin{cases} Z_X = \frac{\partial Z}{\partial X} \\ Z_Y = \frac{\partial Z}{\partial Y} \end{cases} \tag{4}$$

3 Interpretation Of Optical Flow

The problem of estimating the six motion parameters, depth and orientation of the underlying surface from optical flow fields is often referred as the *interpretation of optical flow* [4]. Solving the optical flow equation (2) robustly and efficiently for the motion parameters and surface structure is not as easy as it seems to be. A direct method would call for the observation of optical flow at five points on the same rigid body from which, without considering the scaling factor, ten equations in ten unknowns can be written. Five of these unknowns will be the relative depths of these five points. The other four are the global motion parameters (the magnitude of the translational component can not be recovered). Each additional observation introduces one more unknown and two more constraint equations. Thus the interpretation problem is basically overdetermined. Unfortunately, all these constraint equations are nonlinear. It is very difficult to find a numerically stable solution procedure for nonlinear systems with such a large number of unknowns without good initial guesses. Methods to find such good initial guesses have not yet been developed.

Our goal in this section is to linearize the interpretation problem, and to reduce the dimension of parameter space. We consider two different approaches to make the interpretation problem linear and well-behaved. The first approach comes from the idea of decoupling the rotational component from the optical flow. The second one uses first order spatio-temporal derivatives of the optical flow field at each point. In both these formulations

we are able to obtain closed-form solutions.

3.1 Decoupling rotation

The idea of decoupling rotational component from the optical flow has received some attention. Longuet-Higgins and Prazdny [4] observed that motion parallax which is observable only at depth discontinuities can be used to decouple the rotational component from the optical field. Unfortunately, this seems to be very hard to accomplish. Rieger and Lawton [11] took the difference of optical flow of two neighboring image points to eliminate the rotational component. However their analysis is valid only under some approximations

Our method of decoupling the rotational component from the optical flow relies on the availability of two velocity fields of the same rigid body under observation at two time instants. Once the component of the image displacement due to translation are separated from that due to rotation we have efficient algorithms for the computation of the 3-D motion parameters and structure.

Suppose that we observe two velocity fields of the same environment due to uniform motion of the environment relative to the camera at two time instants. At time $t1$ the optical flow at the image coordinate (x, y) is given by $u(x, y, t1), v(x, y, t1)$. At time $t2$ the optical flow at the same coordinate (x, y) is given by $u(x, y, t2), v(x, y, t2)$. We assume that the 3-D velocity $\underline{V}, \underline{\Omega}$ remain constant between the two time instants. Of course the images at the same location at different times correspond to the projection of two different points on the underlying surface. From (3) we have

$$\begin{cases} u(x,y,t1) = (xV_Z - V_X)/Z(t1) + [xy\Omega_X - (1+x^2)\Omega_Y + y\Omega_Z] \\ v(x,y,t1) = (yV_Z - V_Y)/Z(t1) + [(1+y^2)\Omega_X - xy\Omega_Y - x\Omega_z] \end{cases} \tag{5}$$

and

$$\begin{cases} u(x,y,t2) = (xV_Z - V_X)/Z(t2) + [xy\Omega_X - (1+x^2)\Omega_Y + y\Omega_Z] \\ v(x,y,t2) = (yV_Z - V_Y)/Z(t2) + [(1+y^2)\Omega_X - xy\Omega_Y - x\Omega_z] \end{cases} \tag{6}$$

We now take the difference of the optical flows at the same image location (x, y)

$$\begin{cases} \Delta u(x,y) = u(x,y,t1) - u(x,y,t2) = (xV_Z - V_X)(1/Z(t1) - 1/Z(t2)) \\ \Delta v(x,y) = v(x,y,t1) - v(x,y,t2) = (yV_Z - V_Y)(1/Z(t1) - 1/Z(t2)) \end{cases} \tag{7}$$

Thus the rotation parameters have been eliminated from the difference field. Unlike stereoscopic motion problem [12, 13], we do not have to establish the correspondence between two velocity fields in order to linearize the SFM equations.

To ensure that the difference velocity field is accurate, the difference of $Z(t1)$ and $Z(t2)$ has to be relatively large. This means that the surface must have enough variation in depth. A discussion on the importance of variation in depth in human vision can be found in [11]

Once the difference velocity field has been found, it is relatively easy to compute the FOE. By taking the ratio of the two components of the difference velocity at (x, y), one obtains:

$$\frac{\Delta v(x,y)}{\Delta u(x,y)} = \frac{y - y_0}{x - x_0} \tag{8}$$

where

$$x_0 = \frac{V_X}{V_Z}, y_0 = \frac{V_Y}{V_Z} \tag{9}$$

Thus the difference velocity field is everywhere directed away from (or towards) some image location (x_0, y_0) called FOE [4].

Equation (8) is valid at (x, y) as long as the image at (x, y) corresponds to the two different points on the same rigid body. If the equation is violated, then (x, y) must be

on the boundary of an object. This is the basis of motion segmentation [14]. The above formulation illustrates how we can utilize more than two frames of images (or more than one velocity field) to improve the accuracy in the SFM problem.

Theoretically, we need only the difference velocities at two different image locations in order to solve (8) for the FOE. But if we take into account the noise in the computation of optical flow, we may get very undesirable results using so little information. We now propose two ways to combine the global information to reduce the effect of noise.

3.1.1 Least-squares formulation

If the scene has already been segmented into individual objects, we can use linear least-squares methods to improve the accuracy. Assume that a set of image points $(x_i, y_i), i = 1, 2, ..., M$ where the difference velocities are available belong to the same object in the scene. For each point we have a linear equation in x_0, y_0:

$$\Delta v(x_i, y_i)x_0 - \Delta u(x_i, y_i)y_0 = \Delta v(x_i, y_i)x_i - \Delta u(x_i, y_i)y_i \quad \text{for } i = 1, 2, ..., M \tag{10}$$

The least-squares solution of the above set of linear equations is given by

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = (H^t H)^{-1} H^t \underline{C} \tag{11}$$

where $H_{M\times 2}$ is given by

$$H = \begin{bmatrix} \Delta v(x_1, y_1) & -\Delta u(x_1, y_1) \\ \vdots & \vdots \\ \Delta v(x_M, y_M) & -\Delta u(x_M, y_M) \end{bmatrix} \tag{12}$$

and $\underline{C}_{M\times 1}$ given by

$$\underline{C} = \begin{bmatrix} \Delta v(x_1, y_1)x_1 - \Delta u(x_1, y_1)y_1 \\ \vdots \\ \Delta v(x_M, y_M)x_M - \Delta u(x_M, y_M)y_M \end{bmatrix} \tag{13}$$

The solution procedure only involves inverting a 2 by 2 matrix, thus is very efficient and robust.

3.1.2 Hough Transform Formulation

If the scene contains several rigid objects undergoing relative motion, the above method can be applied only after the scene has been segmented into individual objects. If all the FOEs are known, (8) can be used to identify if the point is on the boundary of an object or not. Thus the remaining task is to find the FOEs of all the rigid body motions in the scene.

From (10) we see that each measurement of difference velocity provides a linear constraint equation on the FOE, thus 'voting' for a set of values of FOEs. This is an ideal situation in which we can apply the Hough Transform, as the dimension of the parameter is low and the constraint equation is linear.

Following the Hough transform formalism [15], one obtains the following algorithm:

ALGORITHM: FOE detection using the Hough transform.

1. Quantize parameter space between appropriate maximum and minimum values for x_0 and y_0.

2. Form an accumulator array $A(x_0, y_0)$ whose elements are initially zero

3. For each point (x, y) where the difference velocity is available, and exceeds some threshold, increment all points in the accumulator array along the appropriate line, i.e,

$$A(x_0, y_0) = A(x_0, y_0) + 1$$

for x_0 and y_0 satisfying $\Delta v(x, y)x_0 - u(x, y)y_0 = \Delta v(x, y)x - \Delta u(x, y)y$

4. Local maxima in the accumulator array now correspond to the FOEs of the objects in the scene

Using the FOEs as the features of objects, we can use standard clustering techniques in pattern recognition to classify all the image points into rigid objects. Of course the translational component is not a unique criterion to segment motion in images. But it is the first stage of motion segmentation. The second stage uses the Hough transform to cluster the rotational component based on the result of the first stage. The constraint equations are again linear. The details of motion image segmentation will be the topic of a forthcoming paper.

Once the scene has been segmented, and the motion parameters are found using the Hough transform followed by least-squares, it is quite straightforward to find the depth.

3.2 A Spatio-Temporal Solution

An alternative way of linearizing the SFM problem is using the first and second order spatial derivatives of the optical flow [4, 16]. In this section, we show that one can use the more robust first order temporal derivatives of the optical flow field in place of the noise-sensitive second order spatial derivatives. We show that the depth and the motion parameters as

well as the orientation of the underlying surface at each point can be recovered from the optical flow and its derivatives at that point. The pixel-wise solution of the SFM problem is very important because it enables us to deal with more general problems such as non-rigid body motion.

Let $u(x,y,t), v(x,y,t)$ denote the optical flow at time t and image location (x,y). The surface can be described either by $Z(x,y,t)$ or by $Z(X,Y,t)$, where (x,y) and (X,Y) are related by the perspective projection in (1). Note that here the third dimension –time t and the 2-D spatial dimension are treated in a uniform way.

By including the temporal dimension in (3), one obtains:

$$\begin{cases} u(x,y,t) = (xV_Z - V_X)/Z(x,y,t) + [xy\Omega_X - (1+x^2)\Omega_Y + y\Omega_Z] \\ v(x,y,t) = (yV_Z - V_Y)/Z(x,y,t) + [(1+y^2)\Omega_X - xy\Omega_Y - x\Omega_Z] \end{cases} \tag{14}$$

In this section, we assume that the relative motion remains constant for a short period of time. By differentiating the above equation with respect to the spatial coordinates x and y, one obtains.

$$\begin{cases} u_x(x,y,t) = [\frac{V_Z}{Z(x,y,t)} - \frac{xV_Z - V_X}{Z^2(x,y,t)} Z_x] + [y\Omega_X - 2x\Omega_Y] \\ v_x(x,y,t) = [-\frac{yV_Z - V_Y}{Z^2(x,y,t)} Z_x] + [-y\Omega_Y - \Omega_Z] \end{cases} \tag{15}$$

and

$$\begin{cases} u_y(x,y,t) = [-\frac{xV_Z - V_X}{Z^2(x,y,t)} Z_y] + [x\Omega_X + \Omega_Z] \\ v_y(x,y,t) = [\frac{V_Z}{Z(x,y,t)} - \frac{yV_Z - V_Y}{Z^2(x,y,t)} Z_y] + [2y\Omega_X - x\Omega_Y] \end{cases} \tag{16}$$

One also takes the temporal derivatives of the velocity field to obtain:

$$\begin{cases} u_t(x,y,t) = [-\frac{xV_Z - V_X}{Z^2(x,y,t)}] \frac{\partial Z(x,y,t)}{\partial t} \\ v_t(x,y,t) = [-\frac{yV_Z - V_Y}{Z^2(x,y,t)}] \frac{\partial Z(x,y,t)}{\partial t} \end{cases} \tag{17}$$

From (2) we know that

$$\frac{dZ(x,y,t)}{dt} = -V_Z - \Omega_X Y + \Omega_Y X \tag{18}$$

And from the chain rule of differentiation, we also have

$$\frac{dZ}{dt} = \frac{\partial Z}{\partial x}\frac{dx}{dt} + \frac{\partial Z}{\partial y}\frac{dy}{dt} + \frac{\partial Z}{\partial t} \tag{19}$$

Noting that $\frac{dx}{dt} = u$, $\frac{dy}{dt} = v$ is exactly the optical field, one can solve for $\frac{\partial Z}{\partial t}$:

$$\frac{\partial Z(x,y,t)}{\partial t} = -V_Z - \Omega_X Y + \Omega_Y X - Z_x u(x,y,t) - Z_y v(x,y,t) \tag{20}$$

Thus the temporal derivative of the velocity field is given by:

$$\begin{cases} u_t(x,y,t) = [\frac{V_X - xV_Z}{Z^2(x,y,t)}][-V_Z - \Omega_X Y + \Omega_Y X - Z_x u(x,y,t) - Z_y v(x,y,t)] \\ v_t(x,y,t) = [\frac{V_Y - yV_Z}{Z^2(x,y,t)}][-V_Z - \Omega_X Y + \Omega_Y X - Z_x u(x,y,t) - Z_y v(x,y,t)] \end{cases} \tag{21}$$

The variables $\frac{\partial Z}{\partial x}$, $\frac{\partial Z}{\partial y}$ are related to the orientation of the surface Z_X, Z_Y by the following lemma.

Lemma 1

$$\begin{cases} \frac{\partial Z}{\partial x} &= Z(x,y,t)\frac{Z_X(X,Y,t)}{1 - xZ_X - yZ_Y} \\ \frac{\partial Z}{\partial y} &= Z(x,y,t)\frac{Z_Y(X,Y,t)}{1 - xZ_X - yZ_Y} \end{cases} \tag{22}$$

Proof: From the perspective projection we have

$$Z(X,Y) = Z(x,y) \Big| \begin{cases} x = \frac{X}{Z(X,Y)} \\ y = \frac{Y}{Z(X,Y)} \end{cases} \tag{23}$$

Taking the partial derivatives of $Z(X,Y)$ with respect to X, Y on both sides of the above equation, and applying the chain rule, one obtains

$$\begin{cases} Z_X &= Z_x\frac{\partial x}{\partial X} + Z_y\frac{\partial y}{\partial X} \\ Z_Y &= Z_x\frac{\partial x}{\partial Y} + Z_y\frac{\partial y}{\partial Y} \end{cases} \tag{24}$$

where

$$\begin{cases} \frac{\partial x}{\partial X} &= \frac{1-xZ_X}{Z} \\ \frac{\partial x}{\partial Y} &= -\frac{x}{Z}Z_Y \\ \frac{\partial y}{\partial X} &= -\frac{y}{Z}Z_X \\ \frac{\partial y}{\partial Y} &= \frac{1-yZ_Y}{Z} \end{cases} \tag{25}$$

Substituting the above equation into (24), it is easy to obtain (22) by solving (24) for Z_x, Z_y using Cramer rule. Q.E.D

An alternative proof can be found in [17]

Using the above lemma, and the perspective projection property, (15)-(16), and (21) can be written as:

$$\begin{cases} u_x(x,y,t) = [\frac{V_Z}{Z(x,y,t)} - \frac{xV_Z - V_X}{Z(x,y,t)}\frac{Z_X}{1-xZ_X-yZ_Y}] + [y\Omega_X - 2x\Omega_Y] \\ v_x(x,y,t) = [-\frac{yV_Z - V_Y}{Z(x,y,t)}\frac{Z_X}{1-xZ_X-yZ_Y}] + [-y\Omega_Y - \Omega_Z] \end{cases} \tag{26}$$

and

$$\begin{cases} u_y(x,y,t) = [-\frac{xV_Z - V_X}{Z(x,y,t)}\frac{Z_Y}{1-xZ_X-yZ_Y}] + [x\Omega_X + \Omega_Z] \\ v_y(x,y,t) = [\frac{V_Z}{Z(x,y,t)} - \frac{yV_Z - V_Y}{Z(x,y,t)}\frac{Z_Y}{1-xZ_X-yZ_Y}] + [2y\Omega_X - x\Omega_Y] \end{cases} \tag{27}$$

and

$$\begin{cases} u_t(x,y,t) = [\frac{V_X - xV_Z}{Z(x,y,t)}][-\frac{V_Z}{Z(x,y,t)} - \Omega_X y + \Omega_Y x - \frac{Z_X}{1-xZ_X-yZ_Y}u(x,y,t) - \frac{Z_Y}{1-xZ_X-yZ_Y}v(x,y,t)] \\ v_t(x,y,t) = [\frac{V_Y - yV_Z}{Z(x,y,t)}][-\frac{V_Z}{Z(x,y,t)} - \Omega_X y + \Omega_Y x - \frac{Z_X}{1-xZ_X-yZ_Y}u(x,y,t) - \frac{Z_Y}{1-xZ_X-yZ_Y}v(x,y,t)] \end{cases} \tag{28}$$

At each point, (14),(27) and (28), together with (29) provide eight nonlinear equations in nine unknowns. However, the depth $Z(x,y,t)$ always appears in ratio with the translational velocity $\underline{V}$ and thus is not recoverable. We can arbitrarily assume $Z(x,y,t) = 1$. Equations (14),(27),(28) and (29) can then be simplified as (we now omit the time argument in all the variables):

$$\begin{cases} u(x,y) = [xV_Z - V_X] + [xy\Omega_X - (1+x^2)\Omega_Y + y\Omega_Z] \\ v(x,y) = [yV_Z - V_Y] + [(1+y^2)\Omega_X - xy\Omega_Y - x\Omega_Z] \end{cases} \tag{29}$$

$$\begin{cases} u_x(x,y) = [V_Z - (xV_Z - V_X)\frac{Z_X}{1-xZ_X-yZ_Y}] + [y\Omega_X - 2x\Omega_Y] \\ v_x(x,y) = -[yV_Z - V_Y]\frac{Z_X}{1-xZ_X-yZ_Y} + [-y\Omega_Y - \Omega_Z] \end{cases} \tag{30}$$

and

$$\begin{cases} u_y(x,y) = [-(xV_Z - V_X)\frac{Z_Y}{1-xZ_X-yZ_Y}] + [x\Omega_X + \Omega_Z] \\ v_y(x,y) = [V_Z - (yV_Z - V_Y)\frac{Z_Y}{1-xZ_X-yZ_Y}] + [2y\Omega_X - x\Omega_Y] \end{cases} \tag{31}$$

$$\begin{cases} u_t(x,y) = [(V_X - xV_Z)][-V_Z - \Omega_X y + \Omega_Y x - \frac{Z_X}{1-xZ_X-yZ_Y}u(x,y) - \frac{Z_Y}{1-xZ_X-yZ_Y}v(x,y)] \\ v_t(x,y) = [(V_Y - yV_Z)][-V_Z - \Omega_X y + \Omega_Y x - \frac{Z_X}{1-xZ_X-yZ_Y}u(x,y) - \frac{Z_Y}{1-xZ_X-yZ_Y}v(x,y)] \end{cases} \tag{32}$$

The above relations form eight nonlinear equations in eight unknowns. The solutions of these nonlinear equations are generally iterative. Yet we are able to reduce the problem to solving a quadratic equation in one unknown.

In the following derivation we use a set of auxiliary variables defined as follows :

$$\begin{cases} T &= xV_Z - V_X \\ T_1 &= yV_Z - V_Y \end{cases} \tag{33}$$

$$\begin{cases} p &= T\frac{Z_X}{1-xZ_X-yZ_Y} \\ q &= T\frac{Z_Y}{1-xZ_X-yZ_Y} \end{cases} \tag{34}$$

and

$$w = V_Z + y\Omega_X - x\Omega_Y \tag{35}$$

It is clear that the two sets of independent variables $V_X, V_Y, V_Z, \Omega_X, \Omega_Y, \Omega_Z, Z_X, Z_Y$ and $T, T_1, w, p, q, \Omega_X, \Omega_X, \Omega_Z$ are equivalent, that is, there is a one-to-one correspondence

between them. Besides there is almost no computation involved from one set of variables to another. Therefore we can try to solve for the second set of variables from the velocity field and its spatio-temporal derivatives.

Equations (30)-(33) can be reformulated in terms of the new set of independent variables:

$$\begin{cases} u &= T + xy\Omega_X - (1+x^2)\Omega_Y + y\Omega_Z \\ v &= \alpha T + (1+y^2)\Omega_X - xy\Omega_Y - x\Omega_Z \\ u_x &= w - p - x\Omega_Y \\ v_x &= -\alpha p - y\Omega_Y - \Omega_Z \\ u_y &= -q + x\Omega_X + \Omega_Z \\ v_y &= w - \alpha q + y\Omega_X \\ u_t &= Tw + pu + qv \\ \frac{T_1}{T} &= \alpha \end{cases} \tag{36}$$

where

$$\alpha = \frac{v_t}{u_t} \tag{37}$$

is known.

Since the first six equations are linear, we can solve them for $\Omega_X, \Omega_Y, \Omega_Z, w, p, q$ in terms of T:

$$\begin{cases} T_1 &= \alpha T \\ \Omega_X &= -\frac{\alpha x}{1+y^2}p - \alpha\frac{1}{1+y^2}T + [\frac{v}{1+y^2} - \frac{xv_x}{1+y^2}] \\ \Omega_Y &= \frac{y}{1+x^2}q + \frac{1}{1+x^2}T + [\frac{yu_y}{1+x^2} - \frac{u}{1+x^2}] \\ \Omega_Z &= -(v_x + \alpha p + y\Omega_Y) \end{cases} \tag{38}$$

and

$$\begin{cases} p &= C_1T + C_2 \\ q &= C_3T + C_4 \\ w &= C_5T + C_6 \end{cases} \tag{39}$$

where $C_1, C_2, C_3, C_4, C_5, C_6$ are the constants defined as follows. Let

$$\begin{cases} a_{11} &= -\frac{\alpha x^2}{1+y^2} - \alpha \\ a_{12} &= -\frac{y^2}{1+x^2} - 1 \\ a_{21} &= 1 - \frac{\alpha yx}{1+y^2} \\ a_{22} &= \frac{xy}{1+x^2} - \alpha \\ c_{11} &= \frac{\alpha x}{1+y^2} + \frac{y}{1+x^2} \\ c_{12} &= v_x + u_y + \frac{x^2v_x - xv}{1+y^2} + \frac{y^2u_y - yu}{1+x^2} \\ c_{21} &= \frac{\alpha y}{1+y^2} - \frac{x}{1+x^2} \\ c_{22} &= v_y - u_x + \frac{xyv_x - vy}{1+y^2} + \frac{ux - xyu_y}{1+x^2} \end{cases} \tag{40}$$

Then,

$$\begin{cases} C_1 &= \frac{c_{11}a_{22} - c_{21}a_{12}}{a_{11}a_{22} - a_{12}a_{21}} \\ C_2 &= \frac{c_{12}a_{22} - c_{22}a_{12}}{a_{11}a_{22} - a_{12}a_{21}} \\ C_3 &= \frac{c_{21}a_{11} - c_{11}a_{21}}{a_{11}a_{22} - a_{12}a_{21}} \\ C_4 &= \frac{c_{22}a_{11} - c_{12}a_{21}}{a_{11}a_{22} - a_{12}a_{21}} \\ C_5 &= C_1 + \frac{yx}{1+x^2}C_3 + \frac{x}{1+x^2} \\ C_6 &= C_2 + \frac{yx}{1+x^2}C_4 + u_x + \frac{xyu_y - ux}{1+x^2} \end{cases} \tag{41}$$

Substituting the above equations into (43), one obtains a quadratic equation in T:

$$C_5T^2 + (C_6 + C_1u + C_3v)T + (C_2u + C_4v - u_t) = 0 \tag{42}$$

This equation generally has two solutions. It means that there is a two-fold ambiguity in interpreting the optical flow at each point. But this ambiguity can be easily resolved using the rigidity assumption. Hough Transform can again be employed to segment the image, and combine the solution at each point into a consistent and robust global solution.

4 Direct Recovery Of Motion And Structure From Image Brightness

All of the previous discussions emphasize the importance of optical flow fields. They are based on the fact that accurate velocity fields are available. Yet we still have not been able to come up with a satisfying algorithm for computing the velocity fields, although progresses in this regard has been reported.

Recently several SFM algorithms that directly use brightness derivative information have been proposed [9, 10]. All these algorithms use the first-order spatio-temporal derivatives of the brightness, and hence are relatively robust. However, the problem of recovering the motion parameters and the environment structure from first-order derivatives of image brightness is underdetermined. Thus these algorithms can only deal with restricted cases, and usually require heuristic search. However as we will show, the SFM problem from the first and second order derivatives are overdetermined.

Assume that the image brightness $E(x, y, t)$ at each point remains the same as the camera moves. By taking the derivatives of the image brightness with respect to t,We have the following well-known equation [18] relating the optical flow the the first-order derivatives

of the brightness.

$$E_t(x,y,t) + E_x(x,y,t)u(x,y,t) + E_y(x,y,t)v(x,y,t) = 0 \tag{43}$$

At each point, the velocity field $u(x,y,t), v(x,y,t)$ is a function of a local variables Z(x,y,t) and the six global motion parameters. Yet the above equation provides only one constraint at each point. Thus the problem of recovering the motion parameters and the depth of the environment is underdetermined. We need to find more constraints.

One way to transform the underdetermined SFM problem into an overdetermined one is to use second-order derivatives of the image brightness[19]. By taking the derivatives of the both sides of the brightness constancy equation with respect to x, y, t, we obtain three more constraint equations:

$$E_{tx} + E_{xx}u(x,y) + E_{xy}v(x,y) + E_x u_x(x,y) + E_y v_x(x,y) = 0 \tag{44}$$

$$E_{ty} + E_{xy}u(x,y) + E_{yy}v(x,y) + E_x u_y(x,y) + E_y v_y(x,y) = 0 \tag{45}$$

$$E_{tt} + E_{xt}u(x,y) + E_{yt}v(x,y) + E_x u_t(x,y) + E_y v_t(x,y) = 0 \tag{46}$$

From (14),(27)-(29) we know that $u_x, u_y, v_x, v_y, u_t, v_t$ introduce two more local variables, namely, the orientation of the underlying surface which is also very useful information.

Thus by plugging equations (14),(27)-(29) into (54)-(57), we have four constraint equations in three local variables, and six global motion parameters at each point. The magnitude of the translational component can not be recovered, as discussed in the previous sections. Thus we only have five global motion parameters. Theoretically, only the first and second-order derivatives of the image brightness at five points are needed. The least-squares approach can then be used to exploit the abundance of available data to improve the accuracy.

The major difficulties in this formulation of the SFM problem are still the nonlinearity of the constraint equations and the high dimensionality of the parameter space. Although the SFM problem is no longer underdetermined if we use the second-order derivatives of the image brightness, the solution is still very unstable. We are still working on robust solution procedures to solve these constraint equations. Again the rotation decoupling might be useful. Or we can use a relatively long sequence of images.

5 Simulation Results

In this section, we present some simulation results to illustrate the robustness of the rotation decoupling algorithms. We assume that two velocity fields are available and we use the rotation decoupling algorithms (least-squares and Hough transform) to find the FOEs. Various degrees of noise are added to the optical flow fields. As we have pointed out, the accuracy of the algorithms depends on how much depth variations the underlying scene contains.The depth variation of the scene is measured by the relative difference of the depth of the scene observed at the same image location at different times.

The scene we use is a sphere of unit radius moving with constant translational velocity $V_X = 2.0, V_Y = 4.0, V_Z = 2.0$, and constant rotational velocity $\Omega_X = 0.0, \Omega_Y = 2.0, \Omega_Z = 1.0$. Thus the FOE is at $x_0 = 1, y_0 = 2$. The input data are the simulated optical flow fields corresponding to the sphere at two different locations. The depth variation is controlled by the positions of the sphere at two different times. Various degrees of noise are added to the optical flow fields. The computed FOEs under a number of depth variation and noise are shown in Table 1.

Depth Variation	Noise	FOE Least-squares	FOE Hough Transform
33%	1%	(1.0,2.0)	(0.99,1.98)
33%	10%	(0.93,1.84)	(0.86,1.69)
33%	20%	(0.75,1.44)	(0.44,0.75)
33%	50%	(0.40,0.69)	(0.26,0.34)
62%	1%	(1.0,2.0)	(0.99,1.98)
62%	10%	(0.96,1.91)	(0.98,1.94)
62%	20%	(0.88,1.74)	(0.95,0.87)
62%	50 %	(0.59,1.08)	(0.77,0.64)

Table 1: Simulation results of Rotation decoupling algorithms with ideal FOE at: $(1, 2)$.

Obviously, the least-squares method is more robust than the Hough transform. This observation justifies the use of least-squares following the Hough transform, which has been explained in the paper.

6 Summary

In this paper, we proposed two methods to linearize the SFM problem. The rotation decoupling algorithm works well if the scene contains enough depth variations. The surface orientation can not be recovered directly using this algorithm. The spatio-temporal approach can be viewed as an extension of the rotation decoupling algorithm when the time interval between two optical flow fields is small. Closed form solutions are obtained for the motion parameters and the surface structure at each pixel. Currently, we are applying the algorithm to real images and extending it to non-rigid body motion analysis.

References

[1] R.Y. Tsai and T.S. Huang, "Uniqueness and Estimation of Three Dimensional Motion Parameters of Rigid Objects with Curved Surfaces", *IEEE Trans. on Patt. Anal. and Mach. Intel.*, vol. PAMI-6, pp. 13–27, January 1984.

[2] S. Ullman, *The Interpretation of Visual Motion*, M.I.T. Press, Cambridge, MA, 1979.

[3] T. J. Broida and R. Chellappa, "Estimation of Object Motion Parameters from Noisy Images", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-8, pp. 90–99, Jan. 1986.

[4] H. C. Longuet-Higgins and K. Prazdny, "The Interpretation of a Moving Retinal Image", *Proc. Royal Society of London*, vol. B-208, pp. 385–397, July 1980.

[5] A. Bruss and B. K. P. Horn, "Passive Navigation", *Computer Vision, Graphics and Image Processing*, vol. 21, pp. 3–20, 1983.

[6] R.C. Bolles and H.H Baker, "Epipolar-plane image analysis: a technique for analyzing motion sequences ", In *Proc. Third IEEE Workshop on Computer Vision: Representation and Control*, pp. 168–178, October 1985.

[7] M. Subbarao, "Interpretation of Image Motion fields: a Spatio-Temporal Approach", In *Proc. IEEE Workshop on Motion: Representation and Analysis*, pp. 157–166, May 1986.

[8] J. Wu, *"Motion Estimation from Image Sequences"*, PhD thesis, Harvard University, Cambridge, Massachusetts, September 1987.

[9] S. Negahdaripour and B.K.P. Horn, *"A Direct Method for Locating the Focus of Expansion"*, Technical Report AI Memo No. 939, MIT Artificial Intelligence Lab., January 1987.

[10] B.K.P Horn and Jr. E.J. Weldon, "Computationally efficient Methods For Recovering Translational Motion", In *Proc. International Conference on Computer Vision*, pp. 2–11, London, England, June 1987.

[11] J.H. Rieger and D.H. Lawton, "Determining the Instantaneous Axis of Translation from Optical Flow Genearted by Arbitrary Sensor Motion", In *Proc. Workshop on Motion: Representation and Perception*, pp. 371–378, July 1983.

[12] A. M. Waxman and J.H. Duncan, "Binocular Image Flows: Steps Toward Stereo-Motion Fusion", *IEEE Trans. on Patt. Anal. and Mach. Intel.*, vol. PAMI-8, pp. 715–729, November 1986.

[13] P. Balasubramanyam and M.A. Synder, "Computation of Motion In Depth Parameters: A First Step in Stereoscopic Motion Interpretation", In *Proc. DARPA Image Understanding Workshop*, Cambridge, Massachusetts, April 1988.

[14] G. Adiv, "Determining Three–Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects", *IEEE Trans. on Patt. Anal. and Mach. Intel.*, vol. PAMI-7, pp. 384–401, July 1985.

[15] D.H. Ballard and C.M. Brown, *Cpmputer Vision*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1982.

[16] A. M. Waxman and S. Ullman, "Surface Structure and 3-D Motion from Image Flow: A Kinematics Analysis", *International Journal of Robotics Research.*, vol. 4(3), pp.

72–94, 1985.

[17] J. Aloimonos and A. Basu, "Combining Information in Low-level Vision", In *Proc. DARPA Image Understanding Workshop*, Cambridge, Massachusetts, April 1988.

[18] B. K. P. Horn and B. G. Schunck, "Determining Optical Flow", *Artificial Intelligence*, vol. 17, pp. 185–203, August 1981.

[19] H.-H. Nagel, "On the Estimation of Optical Flow: Relations between Different Approaches and Some New Results", *Artificial Intelligence*, vol. 33, pp. 299–324, November 1987.

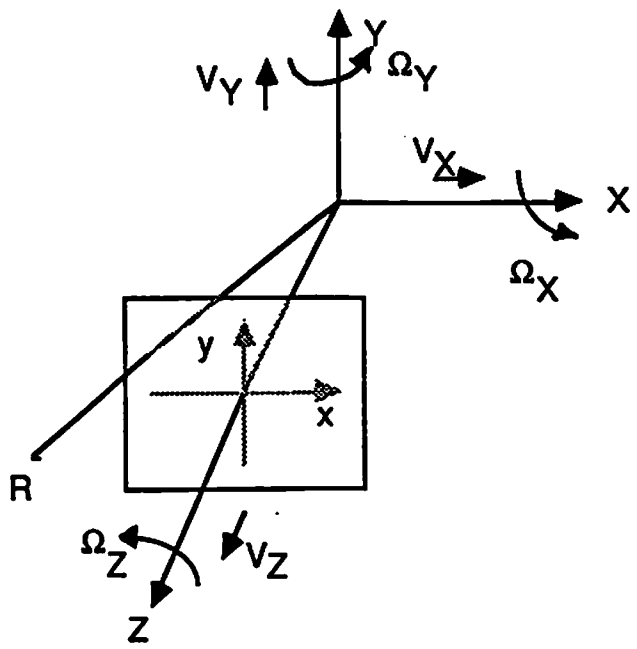

Figure1:The Pin-hole camera and its 3D Motion Model.