Simultaneous Motion, Depth
And Slope Estimation With A Camera-Grid

Hanno Scharr and Tobias Schuchert
Institute for Chemistry and Dynamics of the Geosphere, ICG III
Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
H.Scharr@FZ-Juelich.de

Abstract

Standard optical flow methods are very successful for displacement estimation in image sequences using a brightness change constraint equation (BCCE). Recently a multi-camera generalization of the brightness constancy model has been proposed in [22] and tested for several special cases. The complete, unrestricted model has not been investigated yet. This paper bridges this gap and gives additional detail for the derivation of the model and an appropriately adapted total least squares (TLS) parameter estimation. We use 1d or 2d camera grids acquiring a 4 or 5 dimensional data set, i.e. spatio-temporal data coming from each camera with one or two additional dimensions for the camera position. From such a data set for each surface patch 3d motion, 3d position, and surface normals are estimated simultaneously using a weighted TLS estimation scheme. The implications coming from the 2d or 3d nullspace of the occurring data-matrix of the system of equations are presented in detail. In experiments we test the performance of this approach.

1 Introduction

Motion estimation as well as disparity estimation are standard tasks for optical flow algorithms as well known from early publications (e.g. [13, 18] and many more). The assumption used to estimate optical flow is usually formulated as brightness change constraint equation (BCCE). In [22] a multi-camera extension of the brightness constancy constraint model for simultaneous local estimation of 3d motion, 3d position and surface normals has been introduced. There the model has been tested for 4 special cases: estimation of depth only using 1d and 2d camera grids, estimation of depth and 3d motion, and estimation of depth and normals. In the current paper we simultaneously estimate all the parameters in the model.

An image sequence can be interpreted as data in a 3d $x$-$y$-$t$-space. For optical flow estimation in this space a BCCE defines a linear model for the changes of gray values due to local object motion. The result of the calculation is a displacement vector field. Assuming the data comes from a single fixed camera displacements come from object motion. Assuming a moving camera looking at a fixed scene displacements (then usually called disparities) are anti-proportional to local depth. This is known as structure from camera motion (e.g. [16]).

The basic idea for the new estimation technique

Figure 1: Top left: a camera grid. Top right: a input image acquired at one of the 25 camera positions. Bottom: reconstructed cube with 3d velocity (left) and normals (right).
presented here is to interpret the camera position $s = (s_x, s_y)$ as additional dimension(s) of the data (see Fig. 1, top left). Hence all image sequences acquired by a 1d camera grid can be combined to sample a 4d-Volume in $x$-$y$-$s_x$-$s_y$-$t$-space. If a 2d camera grid is used (e.g. in [19]) we get a 5d-Volume in $x$-$y$-$s_x$-$s_y$-$t$-space. We define brightness constancy constraint in this space as vanishing total differential of the intensity results in the sought estimation [8, 9], scene flow [27] and brightness changes [6, 11]. Robust [3] and variational estimators [4, 13, 21], regularization schemes [15], and special filters [7, 10, 14] have been studied. Recently depth estimation via optical flow and epipolar geometry has been presented relaxing the constraint of parallel lines of sight [24]. Most of these extension are either already in our model or can be applied seamlessly, which is an advantage over other reconstruction methods.

There already exist frameworks for simultaneous motion and stereo analysis (e.g. [5, 26, 28]). Also stereo for curved surfaces still finds attention [17]. To the best of our knowledge no available approach can deal with slanted surfaces and simultaneously estimate 3d motion. The current paper is an extension of [22] as stated above.

**Paper organization.** We start by deriving the constraint equation (Sec. 2) followed by a compact revision of the structure tensor TLS estimator (Sec. 3). Sec. 4 then treats the disentangling of the model parameters. An experimental validation of the novel model for 1d and 2d camera grids within this estimation framework is presented in Sec. 5.

## 2 Derivation of the BCCE

In this section the constraint equation describing local changes in data coming from a camera grid shall be developed, given a 3d object/motion model, a camera model and a brightness change model.

### 2.1 Surface Patch Model

For each pixel at a given point in time we want to estimate depth, motion and surface normals from a local neighborhood. Thus as object/motion model we use a surface patch valid for the whole neighborhood. This surface element at world coordinate position $(X_0, Y_0, Z_0)$ is modelled as a function of time $t$ and local world coordinates $(\Delta X, \Delta Y)$ with $X$- and $Y$-slopes $Z_x$ and $Z_y$, moving with velocity $(U_x, U_y, U_z)$:

$$
\begin{bmatrix}
X \\
Y \\
Z
\end{bmatrix} =
\begin{bmatrix}
X_0 + U_xt + \Delta X \\
Y_0 + U_yt + \Delta Y \\
Z_0 + U_zt + Z_x\Delta X + Z_y\Delta Y
\end{bmatrix}
$$

(1)

The surface normal is then $(-Z_x, -Z_y, 1)$. 

**Related work.** There is rich literature on optical flow estimation techniques (see the overviews [1, 12]) and many extensions have been developed. There are extensions towards affine motion estimation [8, 9], scene flow [27] and brightness
2.2 Camera Model

As camera model we use a pinhole camera at world coordinate position \((s_x, s_y, 0)\), looking into \(Z\)-direction (cmp. Fig. 2)

\[
\begin{pmatrix} x \\ y \\ \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} X - s_x \\ Y - s_y \\ \end{pmatrix}
\]

where camera position space is sampled equidistantly using a camera grid (cmp. Fig. 1 top left).

2.3 Brightness Change Model

We sample \((s_x, s_y)\)-space using several cameras each of which acquires an image sequence. We combine all of these sequences into one 5d data set sampling the continuous intensity function \(I(x, y, s_x, s_y, t)\). We will now formulate the brightness model in this 5d space. We assume that the acquired brightness of a surface element is constant under camera translation, meaning we are looking at a Lambertian surface. In addition this brightness shall be constant in time, i.e. we need temporally constant illumination, which can be relaxed easily using the approach shown in [11]. We see that brightness does not change with one temporal and two spatial coordinates. In other words there is a 3d manifold in our 5d space in which \(I\) does not change. Thus the total differential \(dI\) vanishes in this manifold. The brightness model therefore is

\[
dI = I_x dx + I_y dy + I_{s_x} ds_x + I_{s_y} ds_y + I_t dt = 0
\]

We use the notation \(I_* = \frac{\partial I}{\partial *}\) for derivatives of the image intensities \(I\). Please note that all derivatives and differentials in this equation have physical units. Image coordinates \(x\) and \(y\) are given in pixel, i.e. the physical length of a sensor element on the camera chip. Camera coordinates \(s_x\) and \(s_y\) are given in camera to camera location. And time \(t\) is given in "frames", i.e. the inverse of the temporal acquisition rate. This is important when interpreting unit free parameters estimated from the model equations.

2.4 Combination of the 3 Models

In order to derive a single linear equation from the above models, we first apply the pinhole camera (Eq. 2) to the moving surface element (Eq. 1)

\[
\begin{pmatrix} X \\ Y \\ Z \\ \end{pmatrix} = \begin{pmatrix} x \\ y \\ f \end{pmatrix}
\]

Consequently we can calculate the differentials \(dx\) and \(dy\) for a fixed surface location (i.e. for constant \(\Delta X\) and \(\Delta Y\))

\[
\begin{pmatrix} dX \\ dY \\ \end{pmatrix} = \frac{f}{Z} \begin{pmatrix} dx \\ dy \end{pmatrix} - \frac{f}{Z} \begin{pmatrix} ds_x \\ ds_y \end{pmatrix}
\]

This equation is nonlinear in \(U_z\) as \(Z = Z_0 + U_z t + Z_x \Delta X + Z_y \Delta Y\). We linearize it via the assumption that \(U_z t\) is small compared to the overall depth \(|Z_0| \gg |U_z t\) and omit \(U_z t\) in the denominator.

For image-based expressions we use the following notation (cmp. Fig. 2)

local pixel coordinates

\[
\begin{align*}
x &= x_0 + \Delta x \\
y &= y_0 + \Delta y
\end{align*}
\]

with pixel to patch center pixel distances

\[
\begin{align*}
\Delta x &= \frac{f(1 - Z_x \frac{dy}{f})}{Z_0} \Delta X \\
\Delta y &= \frac{f(1 - Z_y \frac{dx}{f})}{Z_0} \Delta Y
\end{align*}
\]

3d optical flow and disparity

\[
\begin{align*}
u_x &= \frac{f}{Z_0} U_z \\
u_y &= \frac{f}{Z_0} U_y \\
u_z &= -\frac{1}{Z_0} U_z \\
v &= -\frac{f}{Z_0} \\
z_x &= \frac{Z_x}{Z_0(1 - Z_x \frac{dy}{f})} \\
z_y &= \frac{Z_y}{Z_0(1 - Z_y \frac{dx}{f})}
\end{align*}
\]

The expression for the slopes become obvious using the following linearization

\[
\begin{align*}
\frac{-f}{Z_0} (1 - Z_x \frac{\Delta x}{Z_0} - Z_y \frac{\Delta y}{Z_0}) &\approx v + z_x \Delta x + z_y \Delta y
\end{align*}
\]

assuming \(|v| \gg |z_x \Delta x|\) and \(|v| \gg |z_y \Delta y|\). Then \(|v| \gg |z_x \Delta x| \iff 1 \gg |Z_x \Delta x|/(f - Z_x x)|. This is usually well fulfilled for not too large slopes \(Z_x\) and...
normal angle lenses as then $Zx$ is substantially smaller than the focal length $f$ and $\Delta x \ll f$. In our Sony XC-55 camera pixel size is 7.4$\mu$m and $f = 12.5$mm, i.e. $n\Delta x / f \approx n0.6e - 3$ for the $n$th neighbor pixel. An analogue calculation is true for the $y$-direction.

Plugging Eqs. 5–7 into the brightness model Eq. 3 yields the sought for BCCE

$$0 = I_x (v ds_x + (ux + xo u_z) dt) + I_z \Delta x (z_x ds_x + u_z dt) + I_z \Delta y z_y ds_x + I_{sz} ds_x + I_y (v ds_y + (uy + y_0 u_z) dt) + I_y \Delta y (z_y ds_y + u_z dt) + I_y \Delta y (z_x z_y ds_y + I_{sy} ds_y + I_t dt)$$

(8)

where all nonlinear terms coming from the multiplications with $z_x \Delta x$ and $z_y \Delta y$ are neglected. This is allowed as long as approximation Eq. 7 holds. We decompose (and rearrange) this equation into data vector $d$ and parameter vector $p$:

$$d = (I_x, I_y, I_x \Delta x, I_x \Delta y, I_y \Delta y, I_y \Delta x, I_{sz}, I_{sy}, I_t)^T$$

$$p = (v ds_x + (ux + xo u_z) dt, v ds_y + (uy + y_0 u_z) dt, z_x ds_x + u_z dt, z_y ds_y, z_x ds_y + u_z dt, z_y ds_y + u_z dt, ds_x, ds_y, dt)^T$$

(9)

and get the model equation $d^T p = 0$. This linear BCCE can be used for combined estimation of 3d optical flow disparity and normal of the imaged surface element. Looking at the data vector we observe that this model is very similar to an affine optical flow model (see e.g. [9]), but here it has 3 dimensions that take the same role time has in an affine model. The different terms and properties of this model have been investigated individually in [22] by looking at several special cases in experiments. Here we examine the full model for a 1d and a 2d camera grid. We get the constraint equation for a 1d grid by setting $ds_y = 0$ in Eq. 8. Written as data-and parameter vector we get

$$d = (I_x, I_y, I_x \Delta x, I_x \Delta y, I_y \Delta y, I_{sz}, I_t)^T$$

$$p = (v ds_x + (ux + xo u_z) dt, u_y + y_0 u_z) dt, z_x ds_x + u_z dt, z_y ds_y, z_x ds_y + u_z dt, ds_x, ds_y, ds_x, ds_y, dt, dt)^T$$

(10)

All experiments are done using a simple total least squares parameter estimation scheme, which we briefly review in the next section.

3 Revision of the Structure Tensor

In this total least squares parameter estimation method a model with linear parameters $p_x$ has to be given in the form $d^T p = 0$ with data depending vector $d$ and parameter vector $p$. In our case we have one equation for each 4d or 5d pixel. In order to get an over-determined system of equations, one uses the assumption, that within a small neighborhood $\Omega$ of pixels $i$ all equations are approximately solved by the same set of parameters $p$. We get

$$d_i^T p = e_i \text{ for all pixels } i \in \Omega$$

(11)

with errors $e_i$ which have to be minimized by the sought for solution $\hat{p}$. Using a matrix $D$ composed of the vectors $d_i$ via $D_{ij} = (d_i)_j$. Eq. 11 becomes $Dp = e$. We minimize $e$ in a weighted 2-norm

$$||e|| = ||Dp|| = p^T D^T W D p =: p^T Jp \overset{!}{=} \text{min} \quad ||e||$$

(12)

where $W$ is a diagonal matrix containing the weights. In our case Gaussian weights are used (see Sec. 5). The matrix $J = D^T WD$ is the so called structure tensor. The space of solutions $\hat{p}$ is spanned by the eigenvector(s) to the smallest eigenvalue(s) of $J$. We call this space the null-space of $J$ as the smallest eigenvalue(s) are 0 if the model is fulfilled perfectly, i.e. $e = 0$. For standard optical flow the null-space is 1d, here it is 2d or 3d, depending on the model used. If there is not enough variation in the data, the so called aperture problem occurs, meaning the null-space has a higher dimension as indicated by the model. Here, we assume that enough data variation is present. Handling of other cases can be deduced from literature (e.g. [25]).

4 Disentangling the parameters

Using a 1d camera grid and model Eq. 10, the 2 eigenvectors with respect to the smallest eigenvalues span the nullspace of $J$. For a 2d grid and Eq. 9 3 eigenvectors span the nullspace. For all models, we linearly combine these eigenvectors such that all but exactly one component of $\{ds_x, ds_y, dt\}$ vanish. From the linear combination with $dt \neq 0$ and $ds_x = ds_y = 0$, we calculate the motion components. From the other 1 or 2 eigenvectors with $dt = 0$ we derive depth and normals.
4.1 Depth and normals

For the 1d camera grid we build linear combinations \( \tilde{p}_1 \) and \( \tilde{p}_2 \) of the eigenvectors such that component 6 of \( \tilde{p}_1 \) (i.e. \( ds_x \)) and component 7 of \( \tilde{p}_2 \) (i.e. \( dt \)) vanish, respectively, i.e. \( \tilde{p}_{1,6} = 0 \) and \( \tilde{p}_{2,7} = 0 \). Then we get the \( v, z_x, \) and \( z_y \) via

\[
v = \frac{\tilde{p}_{2,1}}{\tilde{p}_{2,6}} \quad z_x = \frac{\tilde{p}_{2,3}}{\tilde{p}_{2,6}} \quad z_y = \frac{\tilde{p}_{2,4}}{\tilde{p}_{2,6}}
\]

(13)
as \( dt = 0 \) for this parameter vector (cmp. Eq. 10).

For the 2d camera grid, 2 linear independent eigenvectors with \( dt = 0 \) can be derived. Thus we could estimate \( v, z_x, \) and \( z_y \) from either of the 2 eigenvectors, but could not decide which one is better. Fortunately one can estimate the error covariance matrix (see [20]) for the estimation pairs and rotate \( s_x-s_y \)-space such that this matrix becomes diagonal. Then the estimates are independent and can be recombined respecting their then known individual errors. A more detailed description of this approach can be found in [22]. For the estimation of \( u_z \) a similar subspace rotation has to be applied. But there the rotation is in \( x-y \)-space not in \( s_x-s_y \)-space. This is described next.

4.2 Motion components

Here, we argue for the 1d camera grid model, but for the 2d camera grid an analogous derivation is valid. The motion components are calculated from \( \tilde{p}_1 \) where \( ds_x = 0 \). We can calculate \( u_z \) from components 3 and 5 individually, but want to treat them as 2 independent measurements, \( u_{z,1} \) and \( u_{z,2} \). They are independent only in the eigensystem of their error covariance matrix \( C \), which we approximate as derived in [20]. It can be diagonalized by rotation \( R \) as \( C = R^T \Sigma R \), where \( \Sigma_{ii} = \sigma_i^2 \) is a diagonal matrix and \( \sigma_i^2 \) are the error variances of the independent measurements. As \( R \) rotates the \( x-y \)-coordinate system we show how this affects \( u_{z,1} \) and \( u_{z,2} \). For \( ds_x = 0 \) we identify the remaining nonzero components of the model equation (10) as an affine flow model \( I_t + \nabla I(A x + u) = 0 \)

\[
I_t + (I_x, I_y) \begin{pmatrix} u_{z,1} & 0 \\ 0 & u_{z,2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} u_x \\ u_y \end{pmatrix} = 0
\]

(14)

Rotating the coordinate system by \( R \) yields

\[
I_t + R \nabla I(A R x + u) = 0 \quad \Leftrightarrow \quad I_t + \nabla I(R^T A R x + \nabla I(R^T u)) = 0
\]

(15)

The diagonal entries of \( R^T A R \) are the independent measurements \( u_{z,1} = (c^2 \tilde{p}_{1,3} + s^2 \tilde{p}_{1,5})/\tilde{p}_{1,7} \) and \( u_{z,2} = (s^2 \tilde{p}_{1,3} + c^2 \tilde{p}_{1,5})/\tilde{p}_{1,7} \), where \( c \) and \( s \) are the components of the normalized first eigenvector of \( C \). The two measurements are combined via

\[
u_z = \frac{u_{z,1}/\sigma_1^2 + u_{z,2}/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}
\]

(16)

respecting their errors \( \sigma_i \). We then get

\[
u_x = \frac{\tilde{p}_{1,1}/\tilde{p}_{1,7} - x_0 u_z}{y_0 u_z} \quad \nu_y = \frac{\tilde{p}_{1,2}/\tilde{p}_{1,7} - y_0 u_z}{y_0 u_z}
\]

(17)

with central pixel coordinates \( x_0 \) and \( y_0 \).

5 Experiments

We want to test the performance of the 1d and 2d camera grid model for simultaneous estimation of motion, depth and normals. In [22] a 1d camera grid model for motion and depth, but without normals has been investigated. We get it from Eq. 9 by setting \( z_x = 0, z_y = 0, \) and \( ds y = 0 \)

\[
d = (I_x, I_y, I_x \Delta x + I_y \Delta y, I_s, I_t)^T \quad p = ((v ds_x + (u_x + x_0 u_z) dt),
\]

\[ 
(y_0 u_z) dt, u_x dt, ds_x dt)^T \]

(18)

In a first experiment we compare the influence of the additional estimation of normals to the other parameters in the 1d case (via Eq. 10) using synthetic data. In a second experiment we test the benefit of a 2d grid (Eq. 9) compared to the 1d grid (Eq. 10). Results for real data acquired under controlled lab conditions have already been depicted in Fig. 1.

5.1 Setup and Prerequisites

In [22] properties of this estimation approach have been investigated with respect to filter choices, linearization effects, and noise influence. We use the results here by choosing 5 cameras (i.e. sampling points in \( s_x \) and \( s_y \)) for the 1d grid and 5 \times 5 cameras for the 2d grid, because the \( s_x \) - and \( s_y \) - derivatives can be calculated with low systematic error. Linearization effects due to the approximation \( |Z_0| \gg |U_z t| \) spoil estimation results for large \( |U_z t| \), as expected. Thus we select our frame rate \( \gg |U_z t|/|Z_0| \). Noise increases the absolute error of the estimated components of the parameter vectors in the structure tensor. Thus image sequences should be acquired such that the estimable signal is
maximal. Especially absolute values of displacements, i.e. uncorrected optical flow components \((u_x + x_0 u_z)\) and \((u_y + y_0 u_z)\) as well as disparity \(v\), should be as large as possible. Unfortunately the estimation process using the structure tensor breaks down around and above 0.25 of the dominating spatial wavelength in the image data. Therefore we smooth the acquired data in \(x\)- and \(y\)-directions in order to suppress smallest wavelengths. For reasonable camera distances in the grid disparity is much larger than 1 pixel. Therefore we define a working depth with an integer disparity \(v_0\) and shift our data in \(s_x\)- and \(s_y\)-directions such that this disparity becomes zero. Then we can estimate disparity reliably in a depth interval around the working depth, where extremal disparities should not exceed \(\pm 1\) pixel. We select our acquisition configuration accordingly. Please note that this restriction is not inherent to the model but is due to the use of the structure tensor and can be relaxed using warping or multiscale estimators (see e.g. [4, 21]).

As synthetic data we calculate sinusoidal patterns with wavenumbers \(K_x = 2\pi f / (Z_0 \lambda_x)\) and \(K_y = 2\pi f / (Z_0 \lambda_y)\), focal length \(f\), depth \(Z_0\) at the origin, and wavelengths \(\lambda_x\) and \(\lambda_y\) in the image plane (length in pixels depends on the pixel size, we use 7.4\(\mu\)m). The sinusoidal pattern is then \(\cos(K_x (c_\alpha \Delta X + s_\alpha \Delta Y)) \cos(K_y (-s_\alpha \Delta X + c_\alpha \Delta Y))\) where \(\Delta X\) and \(\Delta Y\) are calculated by inverting Eq. 4, \(c_\alpha = \cos(\alpha)\) and \(s_\alpha = \sin(\alpha)\) for a given rotation angle \(\alpha\). Two example images can be found in Fig. 3.

**5.2 1d Grid with and without Normals**

In order to show the influence of the additional estimation of surface normals in the 1d camera grid model, we generate noise free, synthetic sinusoidal images (see above) with varying \(Z_x\). The other parameters were set to \(\lambda_x = \lambda_y = \lambda\), \(\alpha = 0^\circ\), \(Z_0 = 100\)mm, \(f = 10\)mm, \(Z_y = 0\). From these sequences we estimate local depth \(Z\) and depth velocity \(U_z\) for both 1d grid models, and \(Z_x\) for the model additionally handling surface normals. The respective plots can be found in Fig. 4 and Fig. 5. There \(Z_x\) is given in degree, i.e. \(Z_x [^\circ] = \arctan(Z_x)\). Looking at Fig. 5 we observe that left and right plots are almost identical. Thus the model with the normals performs as stable as the simpler model without normals. In the case of noise free synthetic data, a positive influence of the normals is visible for the estimation of \(U_z\) but only for short wavelengths (cmp. Fig. 5a and b). For all the other cases depicted in Fig. 5 the influence is negligible. Looking at Fig. 4 we see that the absolute error of the estimated \(Z_x\) is in the range of \(0.001^\circ\) to \(0.01^\circ\) in the noise free case and in the range of \(0.01^\circ\) to \(1^\circ\) when 2.5% Gaussian noise is added to the data.

**5.3 1d Grid versus 2d Grid**

The benefit of a second camera-shift direction is expected to lie in the reduction of a potential aperture problem. This means, looking at a strongly oriented pattern, a 1d camera grid can not resolve...
observe that both grid models perform comparable well below 0.05° and 0.03° in the 1d and the 2d case, respectively. For the standard deviations we observe that both grid models perform comparable in the noise free case. In the noisy data case the aperture problem becomes visible in the 1d model: estimation of $Z_0$ gets worse for increasing $\alpha$, $Z_y$ is not estimable. The 2d model still gives $Z_0$ reliably, $Z_y$ has a high but limited variance.

### 6 Summary and Outlook

We described and tested a novel model for simultaneous estimation of 3d motion, depth and surface normals in multi-camera sequences. We demonstrated its use within the structure tensor estimation framework but due to its BCCE-like form other optical flow estimation methods may be applied as well. This will provide additional benefit e.g. higher robustness and/or faster performance. Using the structure tensor we have investigated the full model and derived a method to disentangle the mixed motion components. In the experiments we compared the 1d model without normals from [22] to the full model. All parameters can be estimated as well or better with the full model. The additional parameters do not lead to instabilities. In addition the 2d model is more robust with respect to the aperture problem. In future work, we will investigate other estimation schemes instead of the structure tensor. Further we will investigate additional constraints e.g. enforcing that slopes of the estimated depth field and the estimated slopes agree.

**Acknowledgement** We thank Andres Chavarria for fruitful discussions. This work has partly been funded by DFG SPP1114 (SCHA 927/1-3).

### References


