CDS

TECHNICAL MEMORANDUM NO. CIT-CDS 95-008
March,1995

Depth from Brightness of Moving Images*

Stefano Soatto and *Pietro Perona*

Control and Dynamical Systems
California Institute of Technology
Pasadena, CA 91125

Depth from Brightness of Moving Images[*]

Stefano Soatto and *Pietro Perona*

Control and Dynamical Systems
California Institute of Technology 116-81
Pasadena – CA 91125, USA
soatto@benissimo.caltech.edu

March 12, 1995

Abstract

In this note we describe a method for recursively estimating the depth of a scene from a sequence of images. The input to the estimator are brightness values at a number of locations of a grid in a video image, and the output is the relative (scaled) depth corresponding to each image-point. The estimator is invariant with respect to the motion of the viewer, in the sense that the motion parameters are not part of the state of the estimator and therefore the estimates do not depend on motion as long as there is enough parallax (the translational velocity is nonzero). This scheme is a "direct" version of an other algorithm previously presented by the authors for estimating depth from point-feature correspondence independent of motion.

Consider a sequence of images, consisting of a map from some location on a pixel grid $\mathbf{x}$ and a particular time instant t onto a brightness value in $\mathbb{R}_+$.

$$\begin{array}{rcl} I : \mathbb{R}^{2\times 2} \times \mathbb{R}_+ & \longrightarrow & \mathbb{R}_+ \\ (\mathbf{x}, t) & \mapsto & I(\mathbf{x}, t). \end{array} \tag{1}$$

In practice the brightness values are quantized, and we will lump the effects of the quantization errors and other sensor noises into an additive Gaussian noise component, so that we measure

$$I(\mathbf{x}, t) + n_I(\mathbf{x}, t) \qquad\qquad n_I \in \mathcal{N}(0, \sigma). \tag{2}$$

As the camera moves relative to the scene, the brightness patches on the image plane move accordingly. Under somewhat restrictive circumstances, we can assume that the brightness of each point in the scene remains unchanged. This assumption can be violated in a number

[*] Research sponsored by NSF ERC in Neuromorphic Systems Engineering at Caltech, ONR grant N00014-93-1-0990. This work is registered as CDS technical report n. CIT-CDS 95-008, March 1995.

of cases (specularities, reflections, non-uniform lightening etc.), but is by and large satisfied in many practical circumstances [4].

The image brightness constancy assumption corresponds to enforcing that the total time-derivative of the image at each pixel location remains constant:

$$\frac{d}{dt} I_r(\mathbf{x}(t), t) = 0 \qquad \forall \mathbf{x} \in \mathbb{R}^2; t \in \mathbb{R};\ r \in N \tag{3}$$

where r indicate a particular level of resolution. By expanding the above derivative into its spatial gradient and its temporal derivative we get

$$\nabla_{\mathbf{x}} I_r(\mathbf{x}, t)\dot{\mathbf{x}}(t) + \frac{\partial I_r(\mathbf{x})}{\partial t} = 0 \tag{4}$$

where $\dot{\mathbf{x}}$ is the velocity of the brightness pattern at the image location $\mathbf{x} = [x\ y]^T$ (optical flow). The optical flow can be loosely related to the velocity of the projection of any particular point $\mathbf{X}$ in the scene (motion field). In particular, under the assumption that the relative motion between the viewer and the scene is rigid with translational velocity V and rotational velocity Ω, the motion field can be written as

$$\hat{\dot{\mathbf{x}}}(t) = \left[\frac{1}{\mathbf{X}_3}\mathcal{A} \mid \mathcal{B}\right] \begin{bmatrix} V(t) \\ \Omega(t) \end{bmatrix} \tag{5}$$

where

$$\mathcal{A} \doteq \begin{bmatrix} 1 & 0 & -x \\ 0 & 1 & -y \end{bmatrix} \qquad \mathcal{B} \doteq \begin{bmatrix} -xy & 1+x^2 & -y \\ -1-y^2 & xy & x \end{bmatrix}. \tag{6}$$

The above equation can be written more concisely as

$$\hat{\dot{\mathbf{x}}} = \mathcal{C}(\mathbf{x}, d) \begin{bmatrix} V \\ \Omega \end{bmatrix}. \tag{7}$$

where

$$d \doteq \frac{1}{\mathbf{X}_3} \tag{8}$$

is the inverse depth of the point with coordinates $\mathbf{X}$ and

$$\mathcal{C}(\mathbf{x}, d) \doteq [d\mathcal{A} \mid \mathcal{B}]. \tag{9}$$

Under the assumption that the scene has Lambertian properties and constant illumination, we can assume that the optical flow and the motion field coincide, $\hat{\dot{\mathbf{x}}} = \dot{\mathbf{x}}$, so that we can substitute (5) into (4) in order to get

$$\nabla_{\mathbf{x}} I_r(\mathbf{x}, t)\mathcal{C}(\mathbf{x}, d) \begin{bmatrix} V \\ \Omega \end{bmatrix} + \frac{\partial I_r(\mathbf{x})}{\partial t} = 0. \tag{10}$$

Given a number of locations on the image plane, for example on a pixel grid,

$$\mathbf{x}^i \qquad \forall i = 1 \ldots n \tag{11}$$

we can collect the constraints in (10) written at $\mathbf{x}^i \ \forall\ i = 1 \ldots n > 6$:

$$\begin{bmatrix} \nabla_{\mathbf{x}^1} I_r(\mathbf{x}^1, t)\mathcal{C}(\mathbf{x}^1, d^1) \\ \vdots \\ \nabla_{\mathbf{x}^n} I_r(\mathbf{x}^n, t)\mathcal{C}(\mathbf{x}^n, d^n) \end{bmatrix} \begin{bmatrix} V \\ \Omega \end{bmatrix} + \begin{bmatrix} \frac{\partial I_r(\mathbf{x}^1)}{\partial t} \\ \vdots \\ \frac{\partial I_r(\mathbf{x}^n)}{\partial t} \end{bmatrix} = 0 \tag{12}$$

and solve for the motion parameters V, Ω as a function of the image derivatives and the inverse depth d in a least-squares sense:

$$\begin{bmatrix} \hat{V} \\ \hat{\Omega} \end{bmatrix} = \mathcal{G}^{\dagger}(\mathbf{x}, \nabla_{\mathbf{x}}\mathbf{I}_r, d)\frac{\partial \mathbf{I}_r(\mathbf{x})}{\partial t} \tag{13}$$

where

$$\mathcal{G}(\mathbf{x}, \nabla_{\mathbf{x}}\mathbf{I}_r, d) \doteq \begin{bmatrix} \nabla_{\mathbf{x}^1} I_r(\mathbf{x}^1, t)\mathcal{C}(\mathbf{x}^1, d^1) \\ \vdots \\ \nabla_{\mathbf{x}^n} I_r(\mathbf{x}^n, t)\mathcal{C}(\mathbf{x}^n, d^n) \end{bmatrix} \tag{14}$$

$$\frac{\partial \mathbf{I}_r(\mathbf{x})}{\partial t} \doteq \begin{bmatrix} \frac{\partial I_r(\mathbf{x}^1)}{\partial t} \\ \vdots \\ \frac{\partial I_r(\mathbf{x}^n)}{\partial t} \end{bmatrix} \tag{15}$$

and $\dagger$ denotes the pseudo-inverse. If we substitute the estimate of the motion parameters $\hat{V}$ and $\hat{\Omega}$ back into equation (12), written for a number of points larger than 6, we end up with a subspace constraint involving only the inverse depths p^i and the derivative of the image brightness:

$$\mathcal{G}\mathcal{G}^{\dagger}(\mathbf{x}, \nabla_{\mathbf{x}}\mathbf{I}_r, d) - \frac{\partial \mathbf{I}_r(\mathbf{x})}{\partial t} = 0 \tag{16}$$

which can be written as

$$\mathcal{G}^{\perp}(\mathbf{x}, \nabla_{\mathbf{x}}\mathbf{I}_r, d)\frac{\partial \mathbf{I}_r(\mathbf{x})}{\partial t} = 0 \tag{17}$$

where

$$\mathcal{G}^{\perp} \doteq Idn - \mathcal{G}\mathcal{G}^{\dagger} \tag{18}$$

where Idn is the identity matrix. Now, since we measure the image brightness at each level of resolution, modulo some noise that we model as a white, zero-mean and Gaussian, we can view the above equation as a nonlinear, implicit dynamical model with parameters p on an $n-$dimensional sphere:

$$\begin{cases} \mathcal{G}^{\perp}(\mathbf{x}, \nabla_{\mathbf{x}}\mathbf{I}_r, d)\frac{\partial \mathbf{I}_r(\mathbf{x})}{\partial t} = 0 \\ Y_r(\mathbf{x}, t) = I_r(\mathbf{x}, t) + n_I(\mathbf{x}, t) \end{cases} \qquad p \in \mathbf{S}^{n-1} \qquad n_I(\mathbf{x}, t) \in \mathcal{N}(0, \Sigma). \tag{19}$$

The normalization of the depth parameters p is due to the inherent scale-factor ambiguity [2].

The above is a dynamical model in nonlinear implicit form, and estimating depth amounts to identifying its parameters $p \in \mathbf{S}^{n-1}$. The Essential Filter [2] is a local recursive observer that accomplishes the task. Therefore we could implement one essential filter at each level of resolution r, and interconnect them by propagating the estimates across scales starting from the coarser level.

Note that this filter estimates the depth at each grid point $\mathbf{x}_i$ (at all pixels, in the limit in which gradients are computed on the whole image), relative to the moving camera but *independent of the motion of the camera.* This holds for any motion such that $V \neq 0$, which is a non-observable configuration for the depth parameters [2]. The motion components have been decoupled from the estimation process in deriving the subspace constraint. This method is inspired by [1], who derive a similar subspace constraint for the direction of translation. However, they do not view the constraint as a dynamic model, and formulate an optimization task between each two views which they solve by exhaustive search over the all possible directions of translation.

This method is a direct extension of the work presented in [3], where depth is estimated recursively and independent of motion from a sequence of feature-point correspondences.

References

[1] D. Heeger and A. Jepson. Subspace methods for recovering rigid motion i: algorithm and implementation. *Int. J. Comp. Vision vol. 7 (2)*, 1992.

[2] S. Soatto, R. Frezza, and P. Perona. Motion estimation via dynamic vision. *Submitted to the IEEE Trans. on Automatic Control. Registered as Technical Report CIT-CDS-94-004, California Institute of Technology. Reduced version to appear in the proc. of the 33 IEEE Conference on Decision and Control. Available through the Worldwide Web Mosaic* (http://avalon.caltech.edu/cds/techreports/) , 1994.

[3] S. Soatto, R. Frezza, and P. Perona. Structure from visual motion as a nonlinear observation problem. In *Proceedings of the IFAC Symposium on Nonlinear Control Systems NOLCOS*, Tahoe City, June 1995.

[4] A. Verri and T. Poggio. Against quantitative optical flow. *Proc. of the 1st Int. Conf. on Computer Vision pp.171-180*, 1987.