Egomotion Estimation Using Binocular Spatiotemporal Oriented Energy

Hao Zhong
zhhao31@gmail.com
Richard P. Wildes
http://www.cse.yorku.ca/~wildes

Department of Electrical Engineering
and Computer Science
York University
Toronto, Ontario, Canada

Abstract

Camera egomotion estimation is concerned with the recovery of a camera's motion (e.g., instantaneous translation and rotation) as it moves through its environment. It has been demonstrated to be of both theoretical and practical interest. This paper documents a novel algorithm for egomotion estimation based on binocularly matched spatiotemporal oriented energy distributions. Basing the estimation on oriented energy measurements makes it possible to recover egomotion without the need to establish temporal correspondences or convert disparity into 3D world coordinates. The resulting algorithm has been realized in software and evaluated quantitatively on a novel laboratory dataset with groundtruth as well as qualitatively on both indoor and outdoor real-world datasets. Performance is evaluated relative to comparable alternative algorithms and shown to exhibit best overall performance.

1 Introduction

Egomotion estimation recovers the time varying motion of a platform, typically in terms of instantaneous rotation and translation. Image-based egomotion estimation effects this recovery on the basis of visual information as well as camera calibration and thereby addresses a fundamental matter in visual information processing – how acquired imagery is related to an optical system's motion through the world. Successful egomotion estimation can provide vital input to a number of related processes, including 3D object modeling [30], Simultaneous Localization and Mapping (SLAM) [2] and sensor platform odometry [37]. In turn, these processes can contribute to larger systems, including mobile robots [6], vehicle guidance [38] and augmented reality [3]. In short, there is no lack of motivation for the development of approaches to camera-based egomotion estimation.

To estimate camera egomotion, monocular, binocular (stereo-based), or multiocular (more than two cameras) algorithms have been widely studied. Generally, monocular and binocular approaches are more popular, with binocular enjoying the advantage, shared with multiocular, of being able to resolve the scale ambiguity between 3D scene structure and camera translation (assuming appropriate calibration) that is inherent to monocular approaches [20]. Here, a brief survey of binocular approaches is provided, as they are most closely related to the proposed approach. (See, e.g., [33] for a review of monocular approaches.)

Many binocular approaches to egomotion estimation share a similar basic structure: Recover disparity between binocular views and then recover rigid body motion parameters by operating on disparity-based 3D point correspondences across time, as mediated by optical flow or 2D feature tracking, e.g., [4, 19, 26, 28, 31, 39, 40]. Other research has affected the recovery of egomotion in disparity space (e.g., [11, 12]), which avoids the need to convert the stereo correspondences into 3D world space. Another class of approach estimates egomotion more directly from a binocular sequence (i.e., without explicit correspondences) by considering image brightness derivatives [17], correlation surfaces [27] or trilinear brightness constraints [36]. Finally, it is notable that binocular approaches have been developed for the related area of visual odometry (e.g., [21, 22, 23, 25, 29]); although, this body of research is more distant from present concerns, as it goes beyond instantaneous egomotion estimation to consider temporal integration of such estimates and thereby obtain position and attitude estimates at any given time along a trajectory.

The proposed approach to egomotion estimation uses spatiotemporal oriented energy measurements that allow it to avoid converting stereo correspondences into explicit 3D, (X,Y,Z), world coordinates as well as avoid the need for explicit temporal correspondences. Spatiotemporal oriented energy measurements have been used previously for a variety of computer vision tasks; most closely related to current work are applications to optical flow [1, 16, 18], tracking [9] as well as stereo disparity and 3D scene flow [34, 35].

In the light of previous research, the contributions of the current work are threefold. 1) An analysis is presented that relates binocularly matched spatiotemporal oriented energies (SOEs) to camera egomotion, as the camera traverses an otherwise rigid 3D environment. It appears that this relationship has not been presented previously. 2) The formal analysis is embodied in a novel algorithm for stereo-based egomotion estimation. 3) The algorithm has been evaluated empirically in comparison to alternative state-of-the-art approaches. As part of the evaluation, a new binocular video dataset is introduced that includes groundtruth egomotion and is available to the community.

2 Technical approach

2.1 Spatiotemporal oriented energy

Binocularly matched, local measurements of spatiotemporal oriented energy (SOE) serve as the data on which the developed approach to egomotion estimation operates. SOEs provide an integrated way to capture spatial appearance and temporal characteristics of an image sequence [34]; therefore, they have the potential to support recovery of egomotion via consideration of temporal dynamics of spatial information as a function of egomotion.

For present purposes, local SOE measurements are recovered separately in the left and right streams of the binocular video via convolution with a bank of Gaussian second derivative filters, $G_2(\hat{\mathbf{w}})$, and their Hilbert transforms, $H_2(\hat{\mathbf{w}})$, which are combined in quadrature to yield energy measurements

$$E(I(\mathbf{x});\hat{\mathbf{w}}) = [G_2(\hat{\mathbf{w}}) * I(\mathbf{x})]^2 + [H_2(\hat{\mathbf{w}}) * I(\mathbf{x})]^2 \qquad (1)$$

where I is an image, $\mathbf{x} = (x,y,t)^\top$ are spacetime image coordinates, the unit vector $\hat{\mathbf{w}}$ specifies the 3D direction of the filter and $*$ is the convolution operator [14].

Most practical uses of energy filtering, (1), involve normalization to make responses invariant to multiplicative bias and bring values to the uniform scale 0 to 1. The necessary operation is realized via pointwise division by the local sum of consort energies at a point

$$\hat{E}(I(\mathbf{x});\hat{\mathbf{w}}_j) = \frac{E(I(\mathbf{x});\hat{\mathbf{w}}_j)}{\sum_i^N E(I(\mathbf{x});\hat{\mathbf{w}}_i)}, \tag{2}$$

with N the number of orientations that span orientation space for the order of filter employed. Indeed, the filter results can serve as a basis set from which energy at any other orientation can be calculated via a weighted combination. Here, since 2^{nd}-order Gaussian filters and Hilbert transforms are used, $N = 10$ is required [14], with their orientations chosen to uniformly sample 3D orientation as the normals to the faces of an icosahedron with antipodal directions identified. The result of this computation is that a set of N (normalized) SOEs are available at each spacetime point, $\mathbf{x}$, in both the left and right image sequences.

Finally, correspondences must be established between points in the left and right image sequences. In general, any reliable algorithm for establishing binocular correspondence could be applied on a framewise basis to the original image sequences [7]. Here, since SOEs are available and previously have been shown useful for stereo video matching [34], that matching approach is applied to establish the needed left-right correspondences.

2.2 Egomotion in visual spacetime

In this subsection, a novel parameterization of 3D directions, $\hat{\mathbf{w}}$, in visual spacetime, (x,y,t), is given in terms of camera egomotion parameters. The derivation begins by reviewing standard material on the visual motion field [20]. Let a Euclidean coordinate system, $(\hat{\mathbf{X}},\hat{\mathbf{Y}},\hat{\mathbf{Z}})^\top$, be defined at the projection centre of the left camera in a rectified binocular pair, with the optical axis and stereo baselines along the $\hat{\mathbf{Z}}$ and $\hat{\mathbf{X}}$, axes, resp., and the $\hat{\mathbf{Y}}$ axis chosen to complete a right-handed system. Under perspective projection, the image coordinates in the left camera are given as $\mathbf{x}^l = (x,y,t)^\top = (X/Z,Y/Z,t)^\top$, with focal length set to unity for conciseness. The coordinates of a corresponding point in the right camera are then given as $\mathbf{x}^r = (x+d,y,t)^\top$, where $d = B/Z$ is stereo disparity and B the stereo baseline.

Let egomotion of the camera be given in terms of instantaneous translational, $\mathbf{T} = (t_x,t_y,t_z)^\top$, and rotational, $\Omega = (\omega_x,\omega_y,\omega_z)^\top$, velocities with respect the centre of projection. Correspondingly, the 3D velocity of a point, $\mathbf{P} = (X,Y,Z)^\top$, relative to the camera is then

$$\dot{\mathbf{P}} = \begin{pmatrix} \dot{X} \\ \dot{Y} \\ \dot{Z} \end{pmatrix} = -\mathbf{T} - \Omega \times \mathbf{P} = \begin{pmatrix} -t_x - \omega_y Z + \omega_z Y, \\ -t_y - \omega_z X + \omega_x Z, \\ -t_z - \omega_x Y + \omega_y X, \end{pmatrix} \tag{3}$$

with "dot notation" used to denote temporal derivatives. In the usual way, the visual motion field, $(u,v)^\top$, which captures the perspective image projection of the relative 3D motion between a camera and 3D world, now can be parameterized in terms of egomotion parameters

$$\begin{pmatrix} u(\mathbf{x};\mathbf{T},\Omega) \\ v(\mathbf{x};\mathbf{T},\Omega) \end{pmatrix} = \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \frac{\dot{X}}{Z} - X\frac{\dot{Z}}{Z^2} \\ \frac{\dot{Y}}{Z} - Y\frac{\dot{Z}}{Z^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{Z}(xt_z - t_x) + \omega_x xy - \omega_y(x^2+1) + \omega_z y \\ \frac{1}{Z}(yt_z - t_y) + \omega_x(y^2+1) - \omega_y xy - \omega_z x \end{pmatrix}. \tag{4}$$

Further, since binocular disparity, d, is assumed available, substituting $\frac{1}{Z} = \frac{d}{B}$ allows for an expression that avoids explicit reference to the 3D world coordinate Z. Similarly, the visual motion field at the corresponding point in the right view is given in terms of the temporal derivative of $(x+d,y)^\top$, i.e., $(\dot{x}+\dot{d},\dot{y})^\top = (u+\dot{d},v)^\top$, where

$$\delta d(\mathbf{x};\mathbf{T},\Omega) = \dot{d} = -B\frac{\dot{Z}}{Z^2} = d\left(\frac{1}{Z}t_z + \omega_x y - \omega_y x\right) = d\left(\frac{d}{B}t_z + \omega_x y - \omega_y x\right), \tag{5}$$

with δd simply an alternate symbol for $\dot{d}$, analogous to the roles of u, v for $\dot{x}, \dot{y}$, resp., in (4).

Finally, image spacetime, $(x,y,t)^\top$, directions associated with visual motion field $(u,v)^\top$ at corresponding points across a binocular sequence is given in terms of unit vectors, $\hat{\mathbf{v}}^l$ and $\hat{\mathbf{v}}^r$, in the left and right imagery (resp.) and parameterized by egomotion parameters, $\mathbf{T}, \Omega$ as

$$\begin{aligned} \hat{\mathbf{v}}^l(\mathbf{x};\mathbf{T},\Omega) &= norm\left(u(\mathbf{x};\mathbf{T},\Omega), v(\mathbf{x};\mathbf{T},\Omega), 1\right)^\top, \\ \hat{\mathbf{v}}^r(\mathbf{x};\mathbf{T},\Omega) &= norm\left(u(\mathbf{x};\mathbf{T},\Omega) + \delta d(\mathbf{x};\mathbf{T},\Omega), v(\mathbf{x};\mathbf{T},\Omega), 1\right)^\top, \end{aligned} \tag{6}$$

where $norm()$ denotes normalization while $u(\mathbf{x};\mathbf{T},\Omega)$, $v(\mathbf{x};\mathbf{T},\Omega)$ and $\delta d(\mathbf{x};\mathbf{T},\Omega)$ are given by their defining equations, (4) and (5).

2.3 Egomotion estimation

If a 3D, $(x,y,t)^\top$, spacetime direction, $\hat{\mathbf{v}}$, is associated with a 2D, $(x,y)^\top$, image flow, $(u,v)^\top$, then it yields minimal energy across orientations, as brightness constancy assumes uniform intensity along the flow direction. Thus, to solve for the appropriate direction, the basis set of oriented energy measurements, (2), can be steered to the direction that yields minimal energy response, as parameterized by the global egomotion parameters, $\mathbf{T}$, Ω. Let oriented energy measurements for corresponding points in left and right imagery be $\hat{E}^l\left(I^l(\mathbf{x}^l); \hat{\mathbf{v}}^l(\mathbf{x}^l;\mathbf{T},\Omega)\right)$ and $\hat{E}^r\left(I^r(\mathbf{x}^r); \hat{\mathbf{v}}^r(\mathbf{x}^r;\mathbf{T},\Omega)\right) = \hat{E}^r\left(I^r(\mathbf{x}^l + \mathbf{d}); \hat{\mathbf{v}}^r(\mathbf{x}^l + \mathbf{d};\mathbf{T},\Omega)\right)$, with $\mathbf{d} = (d,0,0)^\top$, because $\mathbf{x}^l$ and $\mathbf{x}^r$ binocularly correspond. Then matched energies at a point sum to

$$E^{stereo}(I^l(\mathbf{x}^l), I^r(\mathbf{x}^r);\mathbf{T},\Omega) = \hat{E}^l\left(I^l(\mathbf{x}^l); \hat{\mathbf{v}}^l(\mathbf{x}^l;\mathbf{T},\Omega)\right) + \hat{E}^r\left(I^r(\mathbf{x}^l + \mathbf{d}); \hat{\mathbf{v}}^r(\mathbf{x}^l + \mathbf{d};\mathbf{T},\Omega)\right), \tag{7}$$

with $\hat{E}^l$ and $\hat{E}^r$ given by (2) applied to the left, I^l, and right, I^r, image streams, resp.

Within the developed framework, the solution to egomotion estimation can be stated as

$$\arg\min_{\mathbf{T},\Omega} \sum_{\mathbf{x}^l \in \mathcal{S}} E^{stereo}(I^l(\mathbf{x}^l), I^r(\mathbf{x}^l + \mathbf{d});\mathbf{T},\Omega) \tag{8}$$

with $\mathcal{S}$ the set of image points considered in the estimation, as indexed to the left image. Due to the nonlinear dependence of the objective function (8), on $\mathbf{T}$ and Ω, Gauss-Newton refinement is employed to obtain the solution. While alternative optimization methods could be employed [10], Gauss-Newton will be shown useful in the current text when empirical results are presented in Sec. 3. For the sake of conciseness, let $\mathcal{G}^l = G_2\left(\hat{\mathbf{v}}^l(\mathbf{x}^l;\mathbf{T},\Omega)\right) * I^l(\mathbf{x}^l)$ and $\mathcal{H}^l = H_2\left(\hat{\mathbf{v}}^l(\mathbf{x}^l;\mathbf{T},\Omega)\right) * I^l(\mathbf{x}^l)$ and similarly for the right image stream, I^r. Then, egomotion parameters are estimated in terms of the objective function, (8), residual

$$\mathbf{r}(\mathbf{x};\mathbf{T},\Omega) = (\mathcal{G}^l, \mathcal{H}^l, \mathcal{G}^r, \mathcal{H}^r)^\top \tag{9}$$

and Jacobian (using subscripts to denote differentiation)

$$\mathsf{J}(\mathbf{x};\mathbf{T},\Omega) = \begin{pmatrix} \mathcal{G}^l_{t_x} & \mathcal{G}^l_{t_y} & \mathcal{G}^l_{t_z} & \mathcal{G}^l_{\omega_x} & \mathcal{G}^l_{\omega_y} & \mathcal{G}^l_{\omega_z} \\ \mathcal{H}^l_{t_x} & \mathcal{H}^l_{t_y} & \mathcal{H}^l_{t_z} & \mathcal{H}^l_{\omega_x} & \mathcal{H}^l_{\omega_y} & \mathcal{H}^l_{\omega_z} \\ \mathcal{G}^r_{t_x} & \mathcal{G}^r_{t_y} & \mathcal{G}^r_{t_z} & \mathcal{G}^r_{\omega_x} & \mathcal{G}^r_{\omega_y} & \mathcal{G}^r_{\omega_z} \\ \mathcal{H}^r_{t_x} & \mathcal{H}^r_{t_y} & \mathcal{H}^r_{t_z} & \mathcal{H}^r_{\omega_x} & \mathcal{H}^r_{\omega_y} & \mathcal{H}^r_{\omega_z} \end{pmatrix} \tag{10}$$

which are stacked across all n points, $\mathbf{x}_i$, considered in the computation according to

$$\begin{aligned} \rho(\mathbf{T},\Omega) &= \left(\mathbf{r}(\mathbf{x}_1;\mathbf{T},\Omega)^\top, \quad \mathbf{r}(\mathbf{x}_2;\mathbf{T},\Omega)^\top, \quad \ldots, \quad \mathbf{r}(\mathbf{x}_n;\mathbf{T},\Omega)^\top\right)^\top, \\ \mathcal{J}(\mathbf{T},\Omega) &= \left(\mathsf{J}(\mathbf{x}_1;\mathbf{T},\Omega)^\top, \quad \mathsf{J}(\mathbf{x}_2;\mathbf{T},\Omega)^\top, \quad \ldots, \quad \mathsf{J}(\mathbf{x}_\mathsf{n};\mathbf{T},\Omega)^\top\right)^\top \end{aligned} \tag{11}$$

so that iterative refinement proceeds according to

$$\begin{pmatrix} \mathbf{T} \\ \Omega \end{pmatrix}^{k+1} = \begin{pmatrix} \mathbf{T} \\ \Omega \end{pmatrix}^{k} - \left(\mathcal{J}(\mathbf{T},\Omega)^\top \mathcal{J}(\mathbf{T},\Omega)\right)^{-1} \mathcal{J}(\mathbf{T},\Omega)^\top \rho(\mathbf{T},\Omega), \tag{12}$$

with k and $k+1$ successive iterations.

2.4 Salient feature selection

When dealing with real-world images, feature selection can play an important role. Restricting subsequent analysis to reliable features can greatly improve an algorithm's robustness to noise. Feature selection for the developed egomotion algorithm is based on the match score map produced by the stereo matching algorithm used to provide input to the egomotion estimator, e.g., [34], which is combined with a sampling strategy to ensure selected features are reasonably distributed across the images and thereby ameliorate difficulties that arise when global egomotion parameters are estimated based on spatially biased feature distributions.

To select points with reliable disparity estimates, local extrema of curvature of the match score map (*e.g.*, correlation surface) are employed. While a variety of approaches to feature selection might be considered, match score curvature is known to provide reliable (if conservative) indication of loci where stereo correspondence is good [13]. Curvature is calculated as the 2^{nd} spatial derivative of the map along the horizontal axis (assuming horizontally aligned epipolar lines). Nonmax. suppression is used to select local extrema. To ensure the selected points are well distributed across the image, the image is gridded spatially (currently 9×12) and within each grid cell a threshold on the local extrema is set adaptively such that the number of points selected lie between specified min. and max. values. Example selected features are shown in Fig. 1. Notice that selecting points where the stereo match is well defined locally also finds well textured points that will yield correspondingly well defined SOEs (1). Also, gridded adaptive thresholding yields features well distributed spatially.

2.5 Recapitulation

Given a temporal stream of calibrated and rectified binocular imagery, processing proceeds as follows. 1) The left and right image sequences are independently filtered to extract pointwise SOE measurements, (2). 2) Binocular disparity is estimated pointwise [34]. 3) Salient feature points are extracted, Sec. 2.4. 4) The egomotion estimator is executed, (12). Egomotion parameters are initialized identically to zero; estimation ends when the residual change between iterations is below a threshold (10^{-6}) or a maximum number of iterations (50) is reached. To facilitate efficient processing with large capture range, the entire approach is embedded in a coarse-to-fine refinement scheme [5] within Gaussian pyramids [8].

3 Empirical evaluation

The proposed approach to egomotion estimation (**SOE**) has been evaluated empirically on three datasets. The first was acquired in a calibrated laboratory setting and includes groundtruth egomotion [41]. This dataset consists of 7 videos capturing all different combinations of 3 degree-of-freedom (DOF) motion in a plane with systematic variation of velocities. Under the current notation, the parameters are given as t_x, t_z, and ω_y. These parameters are selected as they capture an important practical situation (ground plane motion) and due to mechanical constraints in the lab. The second and third datasets were captured in more naturalistic indoor (an office) and outdoor (building exterior with foreground ground cover), resp. Both consist

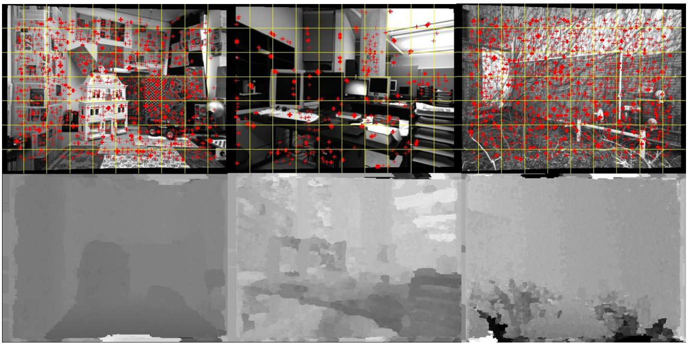


Figure 1: Examples of the images in various datasets. The images are from (left-to-right) the laboratory, indoor and outdoor datasets. Selected features are marked as red plus signs, with the spatial distribution grid also overlaid (see Sec. 2.4 for details). Their corresponding disparity maps are presented on the second row, darker means closer.

of single video sequences emcompassing full 6 DOF egmotion. All imagery was captured with the same binocular video camera with a 6 cm stereo baseline using 75 deg. horizontal field of view lenses for capture at 1024×768 spatial resolution and 30 frames/second. Example (left) images for each dataset are shown in Fig. 1. Below each image are representative disparity maps as estimated by the employed spatiotemporal stereo matcher [34]. An additional dataset [15] was not included in the evaluation as it focuses on visual odometry and thereby lacks framewise egomotion groundtruth.

Two alternative egomotion algorithms are considered for comparison. The first, **DC**, is selected as it is an alternative that, similar to the proposed **SOE**, works without explicit projection of disparity measurements into world, $(X,Y,Z)^\top$ space, and previously outperformed such approaches [12]. This algorithm requires disparities that are matched across time. For the sake of fair comparison, the same disparity measurements and feature point selection used for the proposed approach also are used as input to **DC**. Temporal correspondences are established using the Lucas-Kanade algorithm [24] as implemented in OpenCV [32], with pyramids to increase capture range. The second algorithm (**KGL**) is selected as a state-of-the-art algorithm for binocular-based egomotion estimation as applied to visual odometry [22]. This algorithm makes uses its own techniques for matching between images. Code for the second algorithm was downloaded from its authors' website; code for the first was developed by the present authors, as none appeared available elsewhere. Parameter values for both algorithms were as suggested by their authors or as tuned for best performance on present datasets.

For lab experiments, separate videos were captured for all combinations t_x, t_z and ω_y with speed increased in 15 steps for translation and rotation ranging $2.1-90$ cm/sec. and $0.9-13.5$ deg./sec., resp. Egomotion was realized by attaching the cameras to an automated high precision 3 DOF motion control platform mounted on an optical bench, which also provided groundtruth readings. The same scene was imaged throughout; see Fig. 1. The instantaneous egomotion estimates of each algorithm were compared to groundtruth at 10 equally spaced times across each video; mean and standard deviation of errors were calculated. Algorithms estimated 6 DOF egomotion, even though only at most 3 were actuated.

Lab results are shown in Figs. 2 and 4. For the pure t_x case, **SOE** exhibits smaller error than the alternatives on the actuated t_x, essentially 0 error on the Ω parameters and small error on the nonactuated t_y, t_z. **KGL** also shows small errors, but tends to oscillate about 0. **DC**

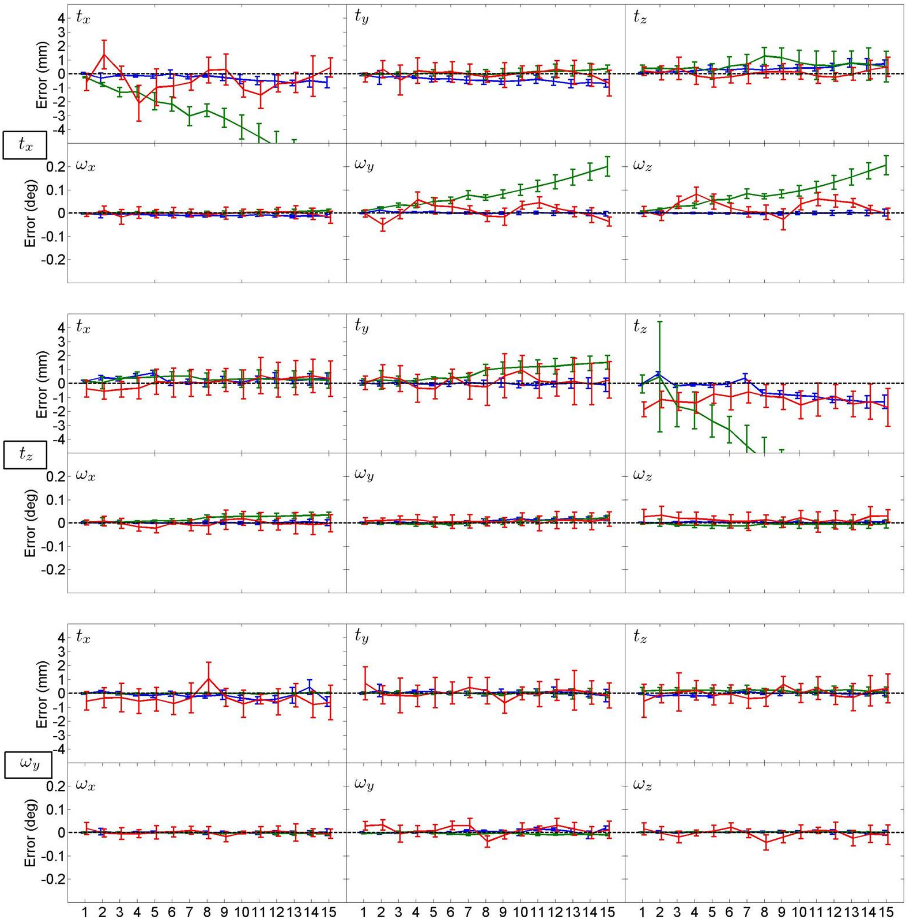
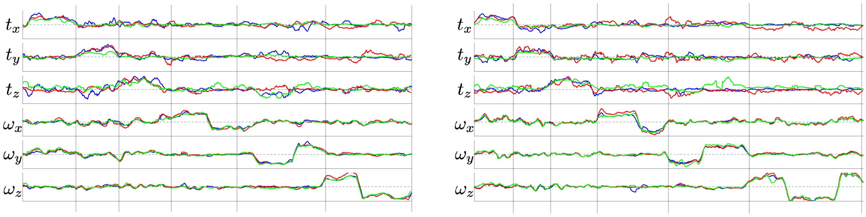


Figure 2: Results on laboratory dataset. Top-to-bottom are grouped error plots as actual egomotion is purely t_x, t_z and ω_y, resp. Subplots show error mean and standard deviation for indicated parameters along the ordinate as speed increases along the abscissa. Blue, green and red denote results for **SOE** (proposed), **DC** and **KGL**, resp. See text for details.



(a)

(b)

Figure 3: Estimated egomotion parameter values vs. time for indoor (left) and outdoor (right) datasets. Algorithm colour coding as in Fig. 2. See text for details.

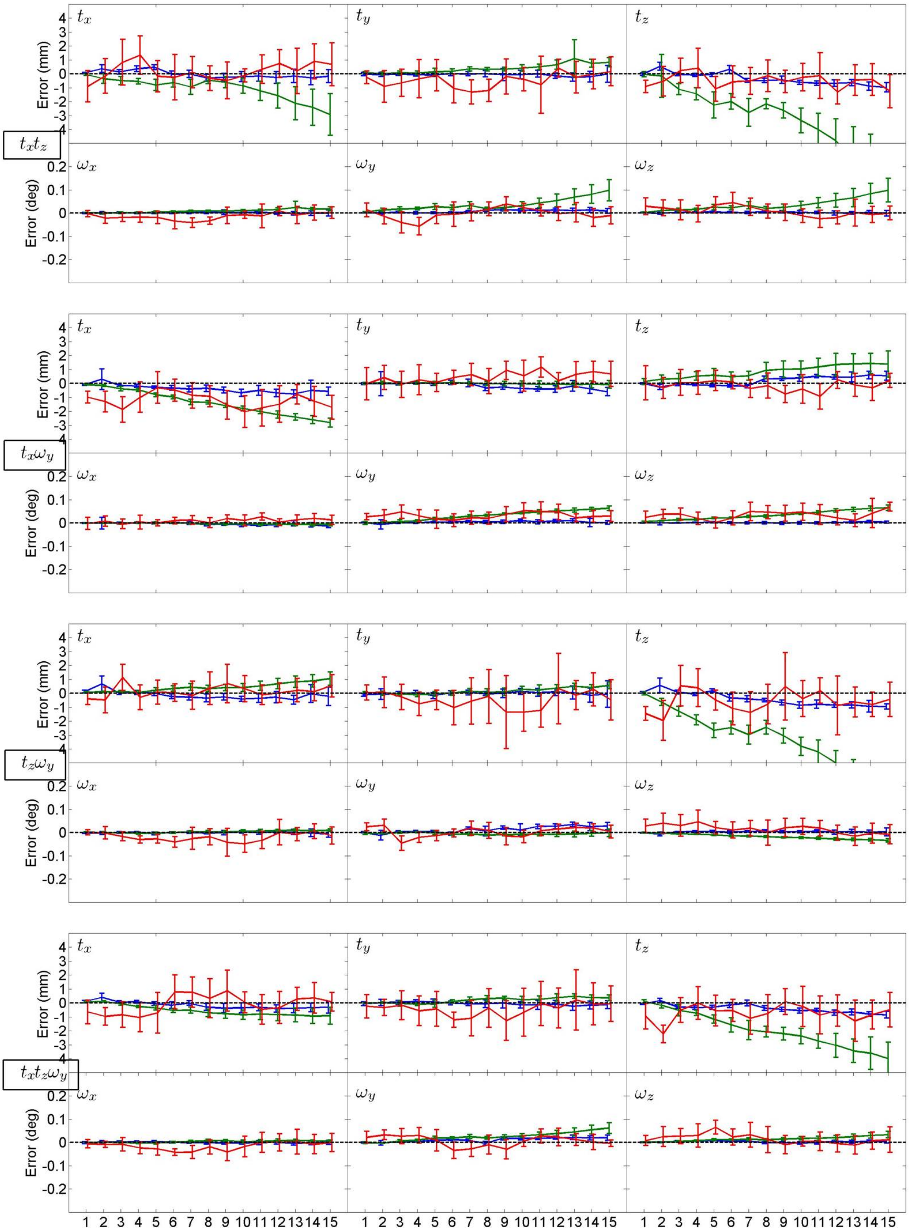


Figure 4: Results on laboratory dataset, part 2. Top-to-bottom are grouped error plots as actual egomotion is $t_x \& t_z, t_x \& \omega_y, t_z \& \omega_y, t_x \& t_z \& \omega_y$, resp. Format otherwise same as Fig. 2

is weakest, with error increasing at higher speeds for t_x, ω_y and ω_z. For the pure t_z case, all algorithms do well in yielding near 0 error for Ω. However, there are differences on **T**: **SOE** and **KGL** show similar small errors on the nonactuated t_x, t_y, but **SOE** has better performance on t_z until at highest speeds it is equaled by **KGL**; **DC** shows error increasing with speed. For pure ω_y, all algorithms show small errors, but **KGL** again oscillates. For combined t_x, t_z, **SOE** has smallest errors for all nonactuated parameters and t_x. At lower speeds, it also shows smallest errors for t_z, but is equaled by **KGL** at higher speeds. **KGL** generally is second smallest in error, but continues to oscillate. **DC** continues its trend of increased error with speed. Combined t_x, ω_y shows **SOE** with smallest error on all parameters, **KGL** second smallest (but still oscillating) and **DC** also with small errors, but larger than the alternatives. Combined t_z, ω_y shows **SOE** with generally smallest error rates, **KGL**'s tendency to oscillate about 0 error particularly pronounced (*e.g.* on t_y, t_z) and **DC** outperforming **KGL**, except on t_x and t_z. Finally, combined t_x, t_z, ω_y shows **SOE** with smallest error on all **T** parameters. **KGL** achieves similar error to **SOE** on Ω and on t_z at high speeds, but still is plagued by oscillation, especially on **T** errors. **DC** performs somewhat better than **KGL**, except on t_z.

Indoor and outdoor naturalistic datasets were captured with the stereo camera handheld. An attempt was made to move sequentially along each of the egomotion parameters, in order $t_x, t_y, t_z, \omega_x, \omega_y, \omega_z$, to yield 6 temporal epochs within a single video. Results are shown in Fig. 3; vertical lines delineate the temporal epochs in each plot. Indoors, all algorithms sequentially increase/decrease their estimates reasonably as the **T** parameters are actuated/deactuated. For Ω, qualitatively correct estimates also are shown, as rotation is performed about each axis first in one direction and then back. A similar pattern of results is shown for outdoors. In both cases, all algorithms tend to show slight nonzero responses to parameters that the camera operator attempted not to actuate. The Supp. Video confirms there was slight motion along these axes, due to the difficulty of actuating one motion at a time. Still, **SOE** shows more stable estimates across time in accord with the video than the alternatives, especially in outdoors. See, *e.g.***KGL**'s greater tendency to oscillate inappropriately about 0 for **T** during Ω actuation as well as variation in its **T** estimates during **T** actuation and **DC**'s tendency to relatively pronounced responses to t_z during Ω actuation, whereas **SOE** tends to more consistent responses throughout. Further, when **SOE** deviates from smoothness the Supp. Video suggests its estimates follow the actual egomotion (*e.g.* t_z responses during t_y actuation indoors, where the operator inadvertently also actuated t_z).

Overall, the results in comparison to groundtruth show that **SOE** exhibits best performance. **KGL** is second best, but tends for its error rates to oscillate with increased velocity. **DC** shows weakest performance, especially at higher speeds. These tendencies may underline the difficulty of establishing reliable temporal correspondences as egomotion (and hence image displacement) increases, a challenge **SOE** avoids by not requiring correspondences across time. Note, in particular, that **DC** employs the same disparity and feature selection as **SOE**; so, differences along those lines do not account for their relative levels of performance. Results on the natural imagery indicate that all algorithms perform qualitatively correctly, with **SOE** showing somewhat more consistent estimates across time. Temporal consistency may result from the benefits of using spatiotemporal orientation analysis, which integrates more temporal information at a given instant (*e.g.* due to underlying filter support).

Finally, **SOE** run-time in unoptimized C++ executed on a 3.4GHz processor with 16GB RAM is $\approx$ 84 ms./frame for 512×384 images, beyond the time required for SOE filtering and stereo matching. Significantly, previous research has shown that both SOE filtering and stereo matching can be done in real-time, *e.g.*, [34]. Thus, the overall approach has potential for real-time applications.

4 Conclusions

This paper has presented a novel algorithm for egomotion estimation based on binocularly matched spatiotemporal oriented energy distributions (SOEs). Basing the estimation on oriented energy measurements made it possible to recover egomotion without the need to establish temporal correspondences or convert disparity into 3D world coordinates. A key to these developments was an analysis that explicitly parameterizes binocularly matched 3D orientation in visual spacetime, (x,y,t), in terms of egomotion parameters. In empirical evaluation, it has been shown that the developed approach is competitive with and even exceeds the accuracy of representative alternative algorithms. An interesting direction for future research would be to embed the developed egomotion estimator in a system for visual odometry.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America. A, Optics and Image Science*, 2(2): 284–99, 1985.

[2] J. Aulinas, Y. Petillot, J. Salvi, and X. Lladó. The SLAM problem: A survey. In *Proceedings of the Conference on Artificial Intelligence Research and Development*, pages 363–371, 2008.

[3] R. Azuma and Y. Baillot. Recent advances in augmented reality. *IEEE Computer Graphics and Applications*, 21:34–47, 2001.

[4] H. Badino. A robust approach for ego-motion estimation using a mobile stereo platform. *Complex Motion*, pages 198–208, 2007.

[5] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the European Conference on Computer Vision*, pages 237–252, 1992.

[6] J. Borenstein, L. Feng, and H. R. Everett. *Navigating Mobile Robots: Systems and Techniques*. A. K. Peters, Ltd., 1996.

[7] M. Brown, D. Burschka, and G. Hager. Advance in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003.

[8] P. Burt and E. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31:532–540, 1983.

[9] K. Cannons, J. Gryn, and R. Wildes. Visual tracking using a pixelwise spatiotemporal oriented energy representation. In *Proceedings of the European Conference on Computer Vision*, pages 511–524, 2010.

[10] G. Dahlquist and A. Bjorck. *Numerical Methods*. Dover, 1974.

[11] D. Demirdjian and T. Darrell. Motion Estimation from Disparity Images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 213–218.

[12] K. Derpanis and P. Chang. Closed-form linear solution to motion estimation in disparity space. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 268–275, 2006.

[13] G. Egnal, M. Mintz, and R. P. Wildes. A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image and Vision Computing*, 22(12):943–957, 2004.

[14] W. Freeman and E. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991.

[15] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[16] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer, Boston, 1995.

[17] K. Hanna and N. Okamoto. Combining stereo and motion analysis for direct estimation of scene structure. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 357–365, 1993.

[18] D. Heeger. Model for the extraction of image flow. *Journal of the Optical Society of America*, 4:1455–1471, 1987.

[19] A. Hogue and M. Jenkin. Development of an underwater vision sensor for 3D reef mapping. In *In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5351–5356, 2006.

[20] B.K.P. Horn. *Robot vision*. MIT Press, Cambridge, MA, 1986.

[21] A. Howard. Real-time stereo visual odometry for autonomous ground vehicles. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3946–3952, 2008.

[22] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 486–492, 2010.

[23] K. Konolige, M. Agrawal, and J. Sola. Large-scale visual odometry for rough terrain. *Robotics Research*, pages 201–212, 2011.

[24] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, volume 2, pages 674–679, 1981.

[25] M. Maimone, Y. Cheng, and L. Matthies. Two years of visual odometry on the mars exploration rovers. *Journal of Field Robotics*, 24:169–186, 2007.

[26] A. Mallet, S. Lacroix, and L. Gallo. Position estimation in outdoor environments using pixel tracking and stereovision. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3519–3524.

[27] R. Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of egomotion and structure from motion and stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 544–550, 1999.

[28] A. Milella and R. Siegwart. Stereo-based ego-motion estimation using pixel tracking and iterative closest point. In *Proceedings of the International Conference on Computer Vision Systems*, 2006.

[29] D. Nistér, O. Naroditsky, and J. Bergen. Visual odometry for ground vehicle applications. *Journal of Field Robotics*, 23(1):3–20, 2006.

[30] J. Oliensis. A Critique of Structure-from-Motion Algorithms. *Computer Vision and Image Understanding*, 80(2):172–214, 2000.

[31] C. Olson, L. Matthies, H. Schoppers, and M. Maimone. Robust stereo ego-motion for long distance navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 453–458, 2000.

[32] OpenCV, 2013. URL http://opencv.org.

[33] F. Raudies and H. Neumann. A review and evaluation of methods estimating egomotion. *Computer Vision and Image Understanding*, 116(5):606–633, 2012.

[34] M. Sizintsev and R. Wildes. Spatiotemporal oriented energies for spacetime stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, 2011.

[35] M. Sizintsev and R.P. Wildes. Spatiotemporal stereo and scene flow via stequel matching. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 34(6):1206–19, 2012.

[36] G.P. Stein and A. Shashua. Model-based brightness constraints: on direct estimation of structure and motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):992–1015, 2000.

[37] N. Sünderhauf and P. Protzel. Stereo Odometry - A Review of Approaches. Technical report, Chemnitz University of Technology, 2007.

[38] I. Vis. Survey of research in the design and control of automated guided vehicle systems. *European Journal of Operational Research*, 170(3):677–709, 2006.

[39] J. Weng, P. Cohen, and N. Rebibo. Motion and structure estimation from stereo image sequences. *IEEE Transactions on Robotics and Automation*, 8(3):362–382, June 1992.

[40] Z. Zhang and O. Faugeras. Estimation of displacements from two frames obtained from stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(12): 1141–1156, 1992.

[41] H. Zhong. Egomotion dataset, 2013. URL http://www.cse.yorku.ca/vision/research/egomotion-data/.