INTERNATIONAL JOURNAL OF
ADVANCED ROBOTIC SYSTEMS

*International Journal of Advanced
Robotic Systems*
May-June 2017: 1–16
© The Author(s) 2017
DOI: 10.1177/1729881417710795
journals.sagepub.com/home/arx
SAGE

A voting method for stereo egomotion estimation

Hugo Silva[1], Alexandre Bernardino[2] and Eduardo Silva[1]

Abstract

The development of vision-based navigation systems for mobile robotics applications in outdoor scenarios is a very challenging problem due to frequent changes in contrast and illumination, image blur, pixel noise, lack of image texture, low image overlap and other effects that lead to ambiguity in the interpretation of motion from image data. To mitigate the problems arising from multiple possible interpretations of the data in outdoor stereo egomotion, we present a fully probabilistic method denoted as probabilistic stereo egomotion transform. Our method is capable of computing 6-degree of freedom motion parameters solely based on probabilistic correspondences without the need to track or commit key point matches between two consecutive frames. The use of probabilistic correspondence methods allows to maintain several match hypothesis for each point, which is an advantage when ambiguous matches occur (which is the rule in image feature correspondence problems), because no commitment is made before analysing all image information. Experimental validation is performed in simulated and real outdoor scenarios in the presence of image noise and image blur. Comparison with other current state-of-the-art visual motion estimation method is also provided. Our method is capable of significant reduction of estimation errors mainly in harsh conditions of noise and blur.

Keywords

Stereo vision, egomotion, visual odometry

Date received: 5 December 2016; accepted: 18 April 2017

Topic: Vision Systems
Topic Editor: Antonio Fernandez-Caballero
Associate Editor: Tiziana D'Orazio

Introduction

In this article, we focus on the inference of robot self-motion (egomotion) based on visual observations of the environment. Although egomotion can be estimated without visual information using sensors such as inertial measurement units (IMUs) or global positioning systems (GPSs), the use of visual information plays an important role specially in MU/GPS denied environments, for example, crowded urban areas or other environments where there are challenging imaging conditions such as aerial and underwater scenarios. In Figure 1, we present some examples of mobile robotic platforms equipped with vision sensors, spanning applications in land, sea, air and underwater (courtesy of INESC TEC).

Egomotion estimation from outdoors imagery is extremely challenging due to multiple factors that generate blur, ambiguities and low signal-to-noise ratio in images. In land robots, camera vibration produces significant motion

[1] INESTEC Centro de Robótica e Sistemas Autónomos, Instituto Superior de Engenharia do Porto, Porto, Portugal
[2] ISR-Lisboa, LARSyS, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

Corresponding author:
Hugo Silva, INESC TEC Centro de Robótica e Sistemas Autónomos, Instituto Superior de Engenharia do Porto, Rua Dr. António Bernardino de Almeida 431, 4200-072 Porto, Portugal.
Email: hugo.m.silva@inesctec.pt

Figure 1. INESC-TEC mobile robotics platforms on land, sea and air application scenarios. All robotic platforms are equipped with one or more visual sensors to perform visual navigation or other complementary tasks.

blur. In sea and underwater robots, repetitive image patterns and low texture generate serious matching ambiguities. In all cases, low lighting conditions, shadows and other illumination artefacts lead to unfavourable signal-to-noise ratios. It is thus essential to develop robust algorithms capable of mitigating some of the aforementioned effects.

This article is an extension of the work by Silva et al.[1] where we have introduced the probabilistic stereo egomotion transform (PSET), a fully probabilistic algorithm for the computation of image motion from stereo vision systems that provides better estimates than alternative approaches. This article provides a deeper explanation, analysis and performance evaluation of PSET. In particular, it focuses on PSET advantages in images with severe amounts of noise and blur that often characterize outdoors operating conditions.

The article outline is as follows: in the following section, the related work is presented. We then make a brief introduction to the probabilistic egomotion estimation problem and an outline of the rationale of the method. Afterwards, we present in detail the steps of the probabilistic stereo egomotion approach, and the obtained results in both synthetic and real image data sets with emphasis in the results obtained under extreme image conditions (presence of image noise and blur). In the final section, we present the conclusions and future work.

Related work

In robotics applications, egomotion estimation is directly linked to visual odometry (VO) applications as described by Scaramuzza.[2] The use of VO methods for estimating robot motion has been a subject of research by the robotics community in recent years. One way of performing VO is by determining instantaneous camera displacement on consecutive frames and integrating over time the estimated linear and angular velocities. The need to develop such applications urged from the increase use of mobile robots on modern world tasks in different application scenarios. Robots need to extend their perception capabilities to be able to navigate in complex scenarios where typical inertial navigation system information cannot be used, for example, urban areas or underwater GPS-denied environments.

Visual motion perception is achieved by measuring image point displacement on consecutive frames. In monocular egomotion estimation, there is translation scale ambiguity, that is, in the absence of other sources of information, only the linear velocity direction can be measured in a reliable manner. Whenever a calibrated stereo setup is used, the full angular and translational velocity components can be extracted, which is denoted by stereo VO.

Most of the work on stereo VO methods started by Maimone et al.[3] and Maimone et al.[4] on the famous Mars Rover Project. The proposed method was able to determine all 6-degree of freedom (DOF) of the rover (x, y, z, roll, pitch and yaw) by tracking 2D image key points between stereo image pairs and obtain their 3D coordinates by triangulation. Concerning the way image motion information is obtained, the method employs a key point detector using[5,6] corner detector combined with a grid scheme to sample key points over the image. After 3D point position is triangulated using stereo correspondence, a fixed number of points is used within an RANSAC[7] framework to obtain an initial motion estimation using least squares. Subsequently, a maximum likelihood estimation (batch estimation) procedure uses the rotation matrix and translation vector obtained by least squares as well as the 'inlier' points to produce a more accurate motion estimation.

The stereo VO method implemented in the Mars Rover Project was inspired by Olson et al.[8] At the time, VO methods appear as replacements for wheel odometry dead reckoning methods to overcome long distance limitations. To avoid large drift in robot position over time, Olson method combined a primitive form of stereo egomotion estimation procedure also used by Maimone et al.[3] with absolute orientation (AO) sensor information.

The taxonomy adopted by the robotics and computer vision community classifies stereo VO methods into two categories based on either feature detection scheme or pose estimation procedure. The most utilized methods for pose estimation are 3D AO methods and perspective-n-point (PnP) methods.

The AO method consists of 3D points triangulation for every stereo pair and then motion estimation is solved using point alignment algorithms, for example, procrustes method,[9] the AO using unit quaternions method by Horn,[10] iterative-closest-point method[11] or the one utilized by Milella and Siegwart[12] for estimating motion of an all-terrain rover.

In the study by Alismail et al.,[13] a benchmark study is performed to evaluate both AO and PnP techniques for robot pose estimation using stereo VO methods. The authors concluded that PnP methods perform better than AO methods due to stereo triangulation uncertainty, especially in the presence of small stereo rig baselines.

The influential work of Nister et al.[14] was one of the first PnP method implementations. It utilized the perspective-three-point method (P3P) developed by Haralick et al.[15] combined with an outlier rejection scheme (RANSAC). Despite the fact of having instantaneous 3D information from a stereo camera setup, the authors use a P3P method instead of a more easily implementable AO method. The authors concluded that P3P pose estimation deals better with depth estimation ambiguity, which corroborates the conclusions drawn by Alismail et al.[13]

In a similar line of work, and in order to avoid having a great dependency of feature matching and tracking algorithms, Kai and Dellaert[16] tested both three-point and one-point stereo VO implementations using a quadrifocal setting within an RANSAC framework. Later on, Ni et al.[17] decouple the rotation and translation estimation into two different estimation problems. The method starts with the computation of a stereo putative matching, followed by a classification of features based on their disparity. Afterwards, distant points are used to compute the rotation using a two-point RANSAC method. The underlying idea is to reduce the problem of the rotation estimation to the monocular case. The closer points with a disparity above a given threshold are used together with the estimated rotation to compute the one-point RANSAC translation.

Recent efforts on stereo VO are being driven by novel intelligent vehicles and by automotive industry applications. One example is the work developed by Kitt et al.[18] The proposed method is available as an open-source VO library named LIBVISO. The stereo egomotion estimation approach is based on image triples and online estimation of the trifocal tensor.[19] It uses rectified stereo image sequences and outputs a 6D vector with linear and angular velocity estimation using an iterative extended Kalman filter. Comport et al.[20] also developed a stereo VO method based on the quadrifocal tensor.[19]

Other recent developments on VO have been achieved by the extensive research conducted at the Autonomous System Laboratory of ETH Zurich University.[21–25] The work developed by Scaramuzza and Fraundorfer[26] and Scaramuzza et al.[21] takes advantages of motion constraints (planar motion) to reduce model complexity and allow a much faster estimation. Also, since the camera is installed on a non-holonomic wheeled vehicle, motion complexity can be further reduced to a single-point correspondence. More recently, the work of Kneip et al.[27] introduced a novel parameterization for the P3P PnP. The method differs from standard algebraic solutions for the P3P estimation problem[15] by computing the aligning transformation directly in a single stage without the intermediate derivation of the points in the camera frame. This pose estimation method combined with key point detectors[28–30] and with IMU information was used to estimate monocular VO[22] and stereo VO by Voigt et al.[23] In the study by He et al.,[31] a visual-inertial egomotion estimation method is used to estimate an arbitrary body motion in indoor environment. Vision is used to estimate the camera motion from a sequence of feature correspondence using bundle adjustment while the inertial estimation outputs the orientation using adaptive-gain orientation filter.

Most of the previously mentioned state-of-the-art algorithms use deterministic methods to find matches between images and then compute the motion. Our approach, on the contrary, takes full advantage of not defining the correspondence at an early stage but keep multiple correspondence hypothesis that together will contribute to a more accurate egomotion estimation, especially when image conditions contain many ambiguous and unreliable correspondences due to non-ideal imaging conditions.

Probabilistic monocular egomotion estimation

The seminal work of Domke and Aloimonos[32] has introduced the notion of probabilistic correspondence in the context of the single camera egomotion estimation problem. The authors introduced the term probabilistic (which is actually a belief) to code the distance between Gabor filters using an exponential transformation. In this setting, it is possible to compute the angular velocity of the vehicle and the direction of the linear velocity (5-DOF) overall, but it is not possible to determine the amplitude (scale) of the linear velocity.

In this section, we briefly describe Domke and Aloimonos'[32] approach and introduce the notation required for the remaining sections.

Probabilistic correspondence

Given two images taken at different times, I_k and I_{k+1}, the probabilistic correspondence between a point $\mathbf{s} \in \mathcal{R}^2$ in image I_k and point $\mathbf{q} \in \mathcal{R}^2$ in image I_{k+1} is defined as a belief image

$$\rho_{\mathbf{s}}(\mathbf{q}) = \text{match}(\mathbf{s}, \mathbf{q} | I_k, I_{k+1}) \tag{1}$$

The belief image $\rho_{\mathbf{s}}(\mathbf{q})$ contains in each pixel $\mathbf{q}$ a value between 0 and 1 expressing similarity of appearance between local neighbourhoods around $\mathbf{s}$ in I_k and $\mathbf{q}$ in I_{k+1}. In the study by Domke and Aloimonos,[32] the match function was implemented by the correlation of a Gabor filter bank response in the two points. In our work, we use the zero-mean normalized cross-correlation function (ZNCC)

$$\text{ZNCC}(\mathbf{s}, \mathbf{q}) = \frac{\sum_{\delta \in W} [I_k(\mathbf{s}+\delta) - \bar{I}_k][I_{k+1}(\mathbf{q}+\delta) - \bar{I}_{k+1}]}{\sqrt{\sum_{\delta \in W} [I_k(\mathbf{s}+\delta) - \bar{I}_k]^2} \sqrt{\sum_{\delta \in W} [I_{k+1}(\mathbf{q}+\delta) - \bar{I}_{k+1}]^2}} \tag{2}$$

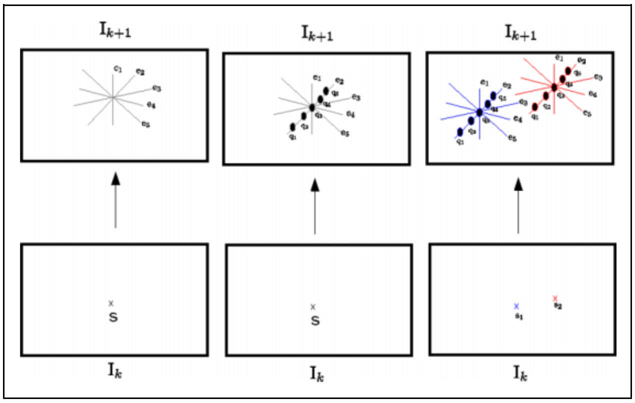

Figure 2. Left: a point **s** in image I_k generates epipolar lines e_i in image I_{k+1} corresponding to motion hypotheses represented by the epipolar matrices E_i, see equation (4). Centre: at each point **s**, motion hypothesis E_i are evaluated by computing the highest probabilistic correspondence at points q_i along epipolar line e_i, see equation (7). Right: the overall motion likelihood is computed by collecting the information from all considered points, see equation (8).

where $W \subset \mathcal{R}^2$ denotes a 2D window centred at the origin whose size defines the neighbourhood of analysis around points **s** and **q**, and $\bar{I}_k$ and $\bar{I}_{k+1}$ are the mean values of those patches. In practice, we use a fast recursive implementation of the ZNCC developed by Huang et al.[33] The probabilistic correspondence is then computed as

$$\rho_{\mathbf{s}}(\mathbf{q}) = \frac{\text{ZNCC}(\mathbf{s}, \mathbf{q}) + 1}{2} \tag{3}$$

So that, it maps to the range 0–1.

Probabilistic motion

Motion hypotheses are defined as a set of incremental rotation matrices R and translation directions $\hat{\mathbf{t}} = \frac{\mathbf{t}}{\|\mathbf{t}\|}$. The likelihood of a particular motion hypothesis $(R, \hat{\mathbf{t}})$ is evaluated by analysing the probabilistic correspondences $\rho_{\mathbf{s}}(\mathbf{q})$ along epipolar lines.[32] A correspondence **q** for point **s** must satisfy the epipolar constraint denoted by

$$\tilde{\mathbf{s}}^T E \tilde{\mathbf{q}} = 0 \tag{4}$$

where $\tilde{\mathbf{s}}$ and $\tilde{\mathbf{q}}$ are homogeneous representations of **s** and **q**, respectively, and E is the essential matrix,[19] a 3 × 3 matrix of rank 2- and 5-DOFs that encodes rigid camera motion

$$E = R[\hat{\mathbf{t}}]_\times \tag{5}$$

where $[\hat{\mathbf{t}}]_\times$ is the skew symmetric matrix

$$[\hat{\mathbf{t}}]_\times = \begin{bmatrix} 0 & -\hat{t}_z & \hat{t}_y \\ \hat{t}_z & 0 & -\hat{t}_x \\ -\hat{t}_y & \hat{t}_x & 0 \end{bmatrix} \tag{6}$$

In order to obtain an estimate of the essential matrix (E) from the probabilistic correspondences, Domke and Aloimonos[32] propose a maximum likelihood search on a probability distribution over the 5D space of essential matrices. Initially, likelihood values are measured on a grid where each dimension is divided into 10 bins, thus leading to 10^5 hypotheses E_i.

For each point **s** in image I_k, the likelihood of a motion hypothesis (E_i) is proportional to the belief of the best probabilistic correspondence along the epipolar constraint in I_{k+1}, generated by the essential matrix

$$\rho(E_i|\mathbf{s}) \propto \max_{(\tilde{\mathbf{q}})^T E_i \tilde{\mathbf{s}} = 0} \rho_{\mathbf{s}}(\mathbf{q}) \tag{7}$$

If one assumes statistical independence between the measurements obtained at each point **s**, the overall likelihood of a motion hypothesis is proportional to the product of the likelihoods for all points

$$\rho(E_i) \propto \prod_{\mathbf{s}} \rho(E_i|\mathbf{s}) \tag{8}$$

In Figure 2, an illustration of these steps is presented.

Finally, having computed all the motion hypotheses, an optimization method[34] is used to refine the motion estimate around the highest scoring samples E_i. The Nelder–Mead simplex method is a local search method for problems whose derivatives are not known. The method was already applied by Domke and Aloimonos[32] to search for the local maxima of likelihood around the top-ranked motion hypotheses

$$E_i^* = \arg \max_{E_i + \delta E} \rho(E_i + \delta E) \tag{9}$$

where δE is perturbations to the initial solution E_i computed by the Nelder–Mead optimization procedure.

Then, the output of the algorithm is the solution with the highest likelihood as defined

$$E^* = \text{argmax}_{E_i^*} \rho(E_i^*) \tag{10}$$

Probabilistic stereo egomotion estimation

Now we extend the notion of probabilistic correspondence and probabilistic egomotion estimation to the stereo case. This allow us to compute the whole 3D motion information in a probabilistic way. Let us consider images I_k^L, I_{k+1}^L, I_k^R and I_{k+1}^R, where superscripts L and R denote, respectively, the left and right images of the stereo pair. Probabilistic matches of a point **s** in I_k^L are now computed not only for points **q** in I_{k+1}^L but also for points **r** in I_k^R and **p** in I_{k+1}^R (see Figures 3 and 4)

$$\rho_{\mathbf{s}}(\mathbf{r}) = \frac{ZNCC(\mathbf{s}, \mathbf{r}) + 1}{2}, \quad \rho_{\mathbf{s}}(\mathbf{p}) = \frac{ZNCC(\mathbf{s}, \mathbf{p}) + 1}{2} \tag{11}$$

For the sake of computational efficiency, analysis can be limited to subregions of the images given prior knowledge about the geometry of the stereo system or bounds of the motion given by other sensors like IMU's. In particular, for

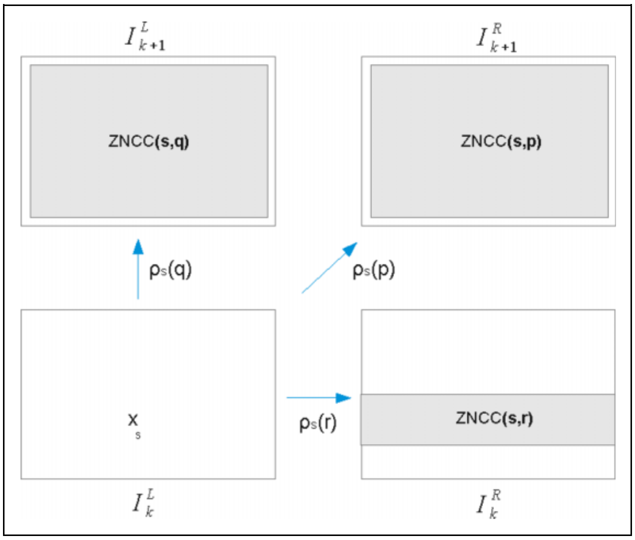

Figure 3. ZNCC matching used to compute the PSET transform. ZNCC: zero-mean normalized cross-correlation function; PSET: probabilistic stereo egomotion transform.

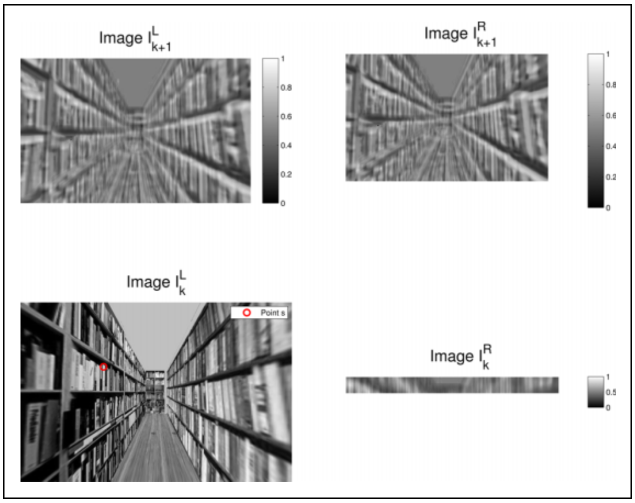

Figure 4. Example of probabilistic correspondence images ($\rho_{\mathbf{s}}(\mathbf{r})$, $\rho_{\mathbf{s}}(\mathbf{q})$, and $\rho_{\mathbf{s}}(\mathbf{p})$) obtained by ZNCC matching of a given point **s** from I^L_k in images, I^R_k, I^L_{k+1}, and I^R_{k+1}, respectively. The $\rho_{\mathbf{s}}(\mathbf{r})$ correspondence can be limited to a band, since the epipolar geometry is known by stereo calibration. ZNCC: zero-mean normalized cross-correlation function.

each point **s**, coordinates **r** can be limited to a band around the epipolar lines according to the fixed stereo setup epipolar geometry, as illustrated in Figure 3.

The geometry of stereo egomotion

In this section, we describe the geometry of the stereo egomotion problem, that is, will analyse how world points project in the four images acquired from the stereo setup in two consecutive instants of time according to its motion. This analysis is required to derive the expressions to compute the translational motion amplitude.

Let us consider the 4 × 4 rototranslations T^R_L and M^{k+1}_k that describe, respectively, the rigid transformation between the left and right cameras of the stereo setup and the transformation describing the motion of the left camera from time k to $k+1$ as described by

$$T^R_L = \begin{bmatrix} R_{\text{calib}} & \mathbf{t}_{\text{calib}} \\ 0 & 1 \end{bmatrix} \quad M^{k+1}_k = \begin{bmatrix} R & \mathbf{t} \\ 0 & 1 \end{bmatrix} \tag{12}$$

We factorize the translational motion $\mathbf{t}$ in its direction $\hat{\mathbf{t}}$ and amplitude α

$$\mathbf{t} = \alpha\hat{\mathbf{t}} \tag{13}$$

The rotational motion R and translation direction $\hat{\mathbf{t}}$ can be computed by Silva et al.[35] The computation of α is thus the main objective at this stage.

Let us consider an arbitrary 3D point $\mathbf{X} = (X_x, X_y, X_z)^T$ expressed in the left camera reference frame at time k. Considering normalized intrinsic parameters (unit focal distance $f = 1$, zero central point $c_x = c_y = 0$, no skew), the homogeneous coordinates of the projection of $\mathbf{X}$ in the four images are given by

$$\begin{cases} \tilde{\mathbf{s}} = \mathbf{X} \\ \tilde{\mathbf{r}} = R_{\text{calib}}\mathbf{X} + \mathbf{t}_{\text{calib}} \\ \tilde{\mathbf{q}} = R\mathbf{X} + \alpha\hat{\mathbf{t}} \\ \tilde{\mathbf{p}} = R_{\text{calib}}R\mathbf{X} + \alpha R_{\text{calib}}\hat{\mathbf{t}} + \mathbf{t}_{\text{calib}} \end{cases} \tag{14}$$

To illustrate the solution, let us consider the particular case of parallel stereo. This will allow us to obtain the form of the solution with simple equations but does not compromise generality because the procedure to obtain the solution in the non-parallel case is analogous. In parallel stereo, the cameras are displaced laterally with no rotation. The rotation component is 3 × 3 identity ($R_{\text{calib}} = I_{3\times 3}$) and the translation vector is an offset (baseline b) along the x coordinate, $\mathbf{t}_{\text{calib}} = (b, 0, 0)^T$. Solving the first two equations, in coordinates, we obtain solutions for $\mathbf{s} = (s_x, s_y)^T$ and $\mathbf{r} = (r_x, r_y)^T$

$$\mathbf{s} = \left(\frac{X_x}{X_z}, \frac{X_y}{X_z}\right)^T \quad \mathbf{r} = \left(\frac{X_x + b}{X_z}, \frac{X_y}{X_z}\right)^T \tag{15}$$

Introducing the disparity d as $d = r_x - s_x$, we have $d = \frac{b}{X_z}$ and we can reconstruct the 3D coordinates of point $\mathbf{X}$ as function of image coordinates $\mathbf{r}$ and $\mathbf{s}$ and baseline value b

$$\mathbf{X} = \left(\frac{s_x b}{d} \quad \frac{s_y b}{d} \quad \frac{b}{d}\right)^T = \tilde{\mathbf{s}}\frac{b}{d} \tag{16}$$

Replacing equation (16) in the last two equations of equation (14), we obtain

$$\begin{cases} \tilde{\mathbf{q}} = R\tilde{\mathbf{s}}\dfrac{b}{d} + \alpha\hat{\mathbf{t}} \\ \tilde{\mathbf{p}} = R\tilde{\mathbf{s}}\dfrac{b}{d} + \alpha\hat{\mathbf{t}} + \mathbf{t}_{\text{calib}} \end{cases} \tag{17}$$

Let us write R in its constituent rows $R = [\mathbf{r_1}, \mathbf{r_2}, \mathbf{r_3}]^T$ and $\hat{\mathbf{t}}$ in coordinates $\hat{\mathbf{t}} = (\hat{t}_x, \hat{t}_y, \hat{t}_z)$. Computing the coordinates of $\mathbf{p} = (p_x, p_y)$ and $\mathbf{q} = (q_x, q_y)$, we get

$$q_x = \frac{\mathbf{r_1^T}\tilde{\mathbf{s}}b + \alpha\hat{t}_x d}{\mathbf{r_3^T}\tilde{\mathbf{s}}b + \alpha\hat{t}_z d} \tag{18}$$

$$q_y = \frac{\mathbf{r_2^T}\tilde{\mathbf{s}}b + \alpha\hat{t}_x d}{\mathbf{r_3^T}\tilde{\mathbf{s}}b + \alpha\hat{t}_z d} \tag{19}$$

$$p_x = \frac{(\mathbf{r_1^T}\tilde{\mathbf{s}} + d)b + \alpha\hat{t}_x d}{\mathbf{r_3^T}\tilde{\mathbf{s}}b + \alpha\hat{t}_z d} \tag{20}$$

$$p_y = \frac{\mathbf{r_2^T}\tilde{\mathbf{s}}b + \alpha\hat{t}_x d}{\mathbf{r_3^T}\tilde{\mathbf{s}}b + \alpha\hat{t}_z d} \tag{21}$$

Solving for α each of the previous equations, we get four possible solutions

$$\alpha^{(1)} = \frac{(\mathbf{r_1^T} - q_x\mathbf{r_3^T})\tilde{\mathbf{s}}}{q_x\hat{t}_z - \hat{t}_x}\frac{b}{d} \tag{22}$$

$$\alpha^{(2)} = \frac{(\mathbf{r_2^T} - q_y\mathbf{r_3^T})\tilde{\mathbf{s}}}{q_y\hat{t}_z - \hat{t}_y}\frac{b}{d} \tag{23}$$

$$\alpha^{(3)} = \frac{(\mathbf{r_1^T} - p_x\mathbf{r_3^T})\tilde{\mathbf{s}}}{p_x\hat{t}_z - \hat{t}_x}\frac{b}{d} \tag{24}$$

$$\alpha^{(4)} = \frac{(\mathbf{r_2^T} - p_y\mathbf{r_3^T})\tilde{\mathbf{s}}}{p_y\hat{t}_z - \hat{t}_y}\frac{b}{d} \tag{25}$$

Solutions exist whenever disparity d is not null, that is, the corresponding 3D point is not at infinity. The other potential singular case is

$$\begin{cases} q_x\hat{t}_z - \hat{t}_x & = 0 \\ q_y\hat{t}_z - \hat{t}_y & = 0 \\ p_x\hat{t}_z - \hat{t}_x & = 0 \\ p_y\hat{t}_z - \hat{t}_y & = 0 \end{cases} \tag{26}$$

This corresponds to the case when $\mathbf{q}$ and $\mathbf{p}$ are simultaneously aligned with the translation direction. However, for finite fields of view, this only happens when $\mathbf{q} = \mathbf{p}$ which again corresponds to zero disparity. If, for a certain combination of points $\mathbf{r}, \mathbf{s}, \mathbf{p}, \mathbf{q}$, all denominators are low (due to very low disparity or close to degenerate motion), that combination is not used for the estimation. In our implementation, we empirically set a predetermined minimal value for the disparity d, $d \geq d_{\min}$. If all disparities are very small, then all observed points are very far and it is impossible to determine the linear velocity scale factor. Therefore, because our method uses all available image information, if at least one point has enough disparity, we will have a solution for α. In practice, to prevent numerical errors, we choose the solution with the largest denominator.

One special case to take into account is when the translational component of motion is zero. When this happens, the value of $\hat{\mathbf{t}}$ computed by the monocular egomotion estimation process is arbitrary but non null, so does not bring any singularity to the problem. The computation of α can be made with the same expressions as before and should result in values very close to zero.

Probabilistic scale estimation

In the previous section, we demonstrated how to estimate the translation scale factor α from the observation of a single static point $\mathbf{s}$, if point correspondences $\mathbf{r}$, $\mathbf{q}$ and $\mathbf{p}$ are known and disparity is non null. In practice, two major problems arise: (i) it is hard to determine what are the static points in the environment given that the cameras are also moving and (ii) it is very hard to obtain reliable matches due to the noise and ambiguities present in natural images. Therefore, using a single point to perform, this estimation is doomed to failure. We must therefore use multiple points and apply robust methodologies to discard outliers.

Previously in Silva et al.,[35] this was achieved by computing the rigid transformation between point clouds obtained from stereo reconstruction at times k and $k+1$. Point correspondences were deterministically assigned by searching for the best matches along epipolar lines in space (from camera L to camera R) and time (from time k to time $k+1$).

Instead in this article, we apply the notion of probabilistic correspondence to the stereo case. Instead of committing matches in space and time, we create a probabilistic observation model for possible matches

$$P_{\text{match}}(\mathbf{s}, \mathbf{r}, \mathbf{p}, \mathbf{q}) = \rho_{\mathbf{s}}(\mathbf{r})\rho_{\mathbf{s}}(\mathbf{q})\rho_{\mathbf{s}}(\mathbf{p}) \tag{27}$$

where we assume statistical independence in the measurements obtained in the pairwise probabilistic correspondence functions $\rho_{\mathbf{s}}(\cdot)$. An example is shown in Figure 5 for the $\rho_{\mathbf{s}}(\mathbf{r})$ case.

From the pairwise probabilistic correspondence, we obtain all possible combination of corresponding matches. Then, because each four-tuple $(\mathbf{s}, \mathbf{r}, \mathbf{p}, \mathbf{q})$ will correspond to a given hypothesis value of α, we create an accumulator of α hypotheses weighted by $P_{\text{match}}(\mathbf{s}, \mathbf{r}, \mathbf{p}, \mathbf{q})$. Searching for global maxima in the accumulator will provide a robust (most agreed) value for α.

PSET accumulator

For computing the accumulator, we assume E has been computed by the previously described methods and the system calibration is known.

First, a large set of points $\mathbf{s}_j, j = 1 \cdots J$ is selected in image I_k^L. Selection can be random, uniform or based on key points, for example, Harris corner[6] or scale-invariant features.[28]

Point-wise computations. For each point $\mathbf{s}_j$, the epipolar lines $E_{\text{calib}} = \tilde{\mathbf{s}}_j^T S$ (being S given by stereo calibration) and $E_{sq} = \tilde{\mathbf{s}}_j^T E$ are sampled at L_j points $\mathbf{r}_j^{l_j}$, $l_j = 1 \cdots L_j$ and M_j points $\mathbf{q}_j^{m_j}$, $m_j = 1 \cdots M_j$, in images I_k^R and I_{k+1}^L,

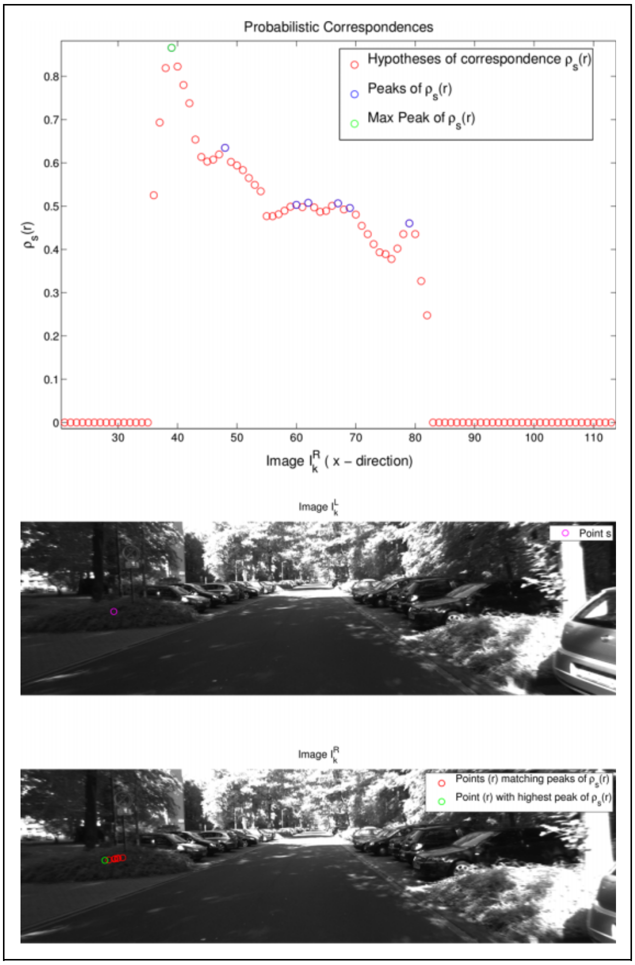

Figure 5. Probabilistic correspondence $\rho_s(r)$ for a point **s** along a section of the epipolar line E_{sr}. On the top figure, we show the probabilistic correspondence values. In red, we have all points of the distribution (non-normalized). In green, we show the global maximum. In blue, we show all other local maxima of $\rho_s(r)$ (blue). On the bottom figure, the sample point **s** in I_k^L and the local maxima peaks in I_k^R are displayed. Sampling is performed in a neighbourhood of the points on the epipolar line.

respectively. Again sample point selection can be uniform along the epipolar lines or based on match quality. In our implementation, we compute local maxima over the epipolar lines. For each triple $(\mathbf{s}_j, \mathbf{r}_j^{l_j}, \mathbf{q}_j^{m_j})$, the geometric solution of $\mathbf{p}$ becomes uniquely determined and is denoted as $\mathbf{p}_j^{l_j m_j}$.

After all probabilistic correspondences have been computed for a point $\mathbf{s_j}$, we create a 2D table $H_j(l_j, m_j)$ to store disparity, likelihood and α values. Each entry (l_j, m_j) of table H_j corresponds to a tuple $(\mathbf{s}_j, \mathbf{r}_j^{l_j}, \mathbf{q}_j^{m_j}, \mathbf{p}_j^{l_j m_j})$ from which it is computed the disparity value $d_j^{l_j}$, the scale value $\alpha_j^{l_j m_j}$ determined by equations (24) and (25), and the match likelihood (27) $\lambda_j^{l_j m_j}$

$$\lambda_j^{l_j m_j} = \rho_{\mathbf{s}_j}\left(\mathbf{r}_j^{l_j}\right)\rho_{\mathbf{s}_j}\left(\mathbf{q}_j^{m_j}\right)\rho_{\mathbf{s}_j}\left(\mathbf{p}_j^{l_j m_j}\right) \qquad (28)$$

Finally, for a particular $\mathbf{s}_j$, we compute the global maximum of $\lambda^j_{l_j m_j}$, which will indicate the best match hypothesis

$$\left(l_j^*, m_j^*\right) = \operatorname{argmax}_{l_j, m_j} \lambda^j_{l_j m_j} \qquad (29)$$

From this best match, we retrieve from H_j, the solution α voted by this point

$$\alpha_j = \alpha_j^{l_j^* m_j^*} \qquad (30)$$

and associated likelihood

$$\lambda_j = \lambda_j^{l_j^* m_j^*} \qquad (31)$$

Thus, each point $\mathbf{s}_j$ votes for a certain motion scale factor α_j, according to the confidence λ_j collected from the probabilistic correspondences in the other images. As a side product, we also get the best disparity hypothesis at that point

$$d_j = d_j^{l_j^*} \qquad (32)$$

Image-wise computations. In the previous section, we described how each point $\mathbf{s}_j$ votes for a translation amplitude α_j with weight λ_j. We collect all these values in sets $A = \{\alpha_j\}$ and $\Lambda = \{\lambda_j\}$, $j = 1 \cdots J$ and use a kernel smoothing method for estimating the highest density of α votes,[36] as described by

$$\hat{f}_h(\alpha) = \frac{1}{jh}\sum_{i=1}^{j} K\left(\frac{\lambda_i(\alpha_i - \alpha)}{h}\right) \qquad (33)$$

being K a Gaussian kernel function, and h the interval bandwidth.

Dealing with calibration errors

A common source of errors in a stereo setup is the uncertainty in the calibration parameters. Both intrinsic and extrinsic parameter errors will deviate the epipolar lines from their nominal values and influence the computed correspondence probability values. To minimize these effects, we modify the correspondence probability function when evaluating sample points such that a neighbourhood of the point is analysed, instead of using only the exact coordinate of the sample point

$$\rho_s'(q) = \max_{q' \in \mathcal{N}(q)} \left[\rho_s(q') \exp\frac{(q - q')^2}{2\sigma^2}\right] \qquad (34)$$

where $N(q)$ denotes a neighbourhood of the sample point q which, in our experiments, is defined as a 7 × 7 window.

Another method used to diminish the uncertainty of the correspondence probability function when performing ZNCC is to use subpixel refinement methods, for example, parabola fitting and Gaussian fitting as presented by Debella-Gilo and Kaab.[37]

Algorithm 1. PSET.

```
Input: Two stereo image pairs $(I_k^L, I_k^R)$ and $(I_{k+1}^L, I_{k+1}^R)$, $E_{rig}$ (stereo
    calibration)
Output: (Velocities) $V$, $W$
Step 1. Use a feature based method to select a set of initial points
    or use the all image.
Step 2. Compute the probabilistic correspondences between
    images $I_k^L$ and $I_{k+1}^L$, $\rho_s(q)$. Equations (1) to (3).
Step 3. Compute probabilistic egomotion, E. Equations (7) to (10).
Step 4. Compute probabilistic correspondences for the stereo
    case, $I_k^L$ and $I_k^R, I_{k+1}^R$, $\rho_s(r), \rho_s(p)$ equation (11).
Step 5. Obtain the probabilistic observation model $P_{\text{match}}$ using
    $\rho_s(r)\rho_s(q)\rho_s(p)$ to relate all possible four-tuple $(s, r, q, p)$
    matches equation (27).
Step 6. Create an accumulator array H for each point $s_j$, and
    perform pointwise computations for obtaining the translation
    scale $\alpha$ and the associated likelihood $\lambda$ for each point,
    equations (28) to (31).
Step 7. Compute the imagewise computations and obtain the final
    translation scale factor $\alpha_{\max}$ using Weighted Kernel density
    estimation (33)
Step 8. Estimate Linear and Angular Velocities, V and W. (35) to (37)
Step 9. Constant Velocity Kalman Filtering. Equations (38) and (39)
```

Velocities estimation

The linear and angular velocities are then estimated, using the same procedure applied by Silva et al.[35] After having obtained the rotation (R), translation direction ($\hat{\mathbf{t}}$) and translation scale factor (α), the linear and angular velocities are computed by

$$V = \frac{\alpha\tilde{\mathbf{t}}}{\Delta T} \tag{35}$$

where ΔT is the sampling interval

$$\Delta T = T_{k+1} - T_k \tag{36}$$

Likewise, the angular velocity is computed by

$$W = \frac{r}{\Delta T} \tag{37}$$

where $r = \theta u$, the angle-axis representation of the incremental rotation R, as defined by Craig.[38]

Kalman filter

In order to achieve a more smooth estimation, we filter the linear and angular velocities estimates using a Kalman filter with a constant velocity model. The state transition model with zero-mean stochastic acceleration is given by

$$X_k = FX_{k-1} + \xi_k \tag{38}$$

where the state transition matrix is the identity matrix, $F = I_{6x6}$, and the stochastic acceleration vector ξ_k is distributed according to a multivariate zero-mean Gaussian distribution with covariance matrix Q, $\xi_k \sim \mathcal{N}(0,Q)$.

The observation model considers state observations with additive noise

$$Y_k = HX_k + \eta_k \tag{39}$$

where the observation matrix H is identity, $H = I_{6x6}$, and the η_k measurement noise is zero-mean Gaussian with covariance R.

We set the covariance matrices Q and R empirically, according to our experiences, to

$$Q = \operatorname{diag}(q_1, \cdots, q_6) \tag{40}$$

$$R = \operatorname{diag}(r_1, \cdots, r_6) \tag{41}$$

where $q_i = 10^{-3}, i = 1, \cdots, 6$, $r_3 = 10^{-3}$ and $r_i = 10^{-4}$, $i \neq 3$.

The r_3 differs from the other measurement noises values, because it corresponds to the translation on the z-axis which is inherently noisier due to the uncertainty of the t_z estimates in the stereo triangulation step. A brief summary of the aforementioned PSET method is described in algorithm 1.

Results

In order to evaluate the accuracy of PSET, we performed evaluation tests with synthetic and real image data. For comparison purposes, we used LIBVISO[18] as a state-the-art deterministic egomotion estimation method. The choice was based on the fact that it is an open source 6D VO library with a filtering step equivalent to ours (constant velocity model).

Synthetic images results

As a first test for evaluating the egomotion estimation accuracy of the PSET method, we utilized a sequence of synthetic stereo images. The sequence was created using a VRML-based simulator and implemented a quite difficult scene (see Figure 6) in which the images contain a great deal of repetitive structure that cause ambiguity in image point correspondence. The sequence is composed by four linear tracks (see Figure 7), as we are more interested in evaluating the performance of the method in the estimation of the translation scale factor.

We assume a stereo camera pair calibrated setup with a 10-cm baseline, 576 × 380 image resolution, with ZNCC window $N_w = 7$. For computational reasons, we used 1000 uniform selected points $\mathbf{s}_j$ for the dense probabilistic egomotion estimation and a subgroup of 100 points $\mathbf{r}_j^{l_j}$ and $\mathbf{q}_j^{l_j}$ ($J = 1000$, $L_j = 100$, $M_j = 100$, $\forall j$). The experiments conducted to compute the PSET were performed using an Intel I5 Dual Core 3.2 GHz which took about 20 s. The code was written in MATLAB as a proof of concept without any kind of optimization. The processing time is spent nearly 70% on the 5D estimation part and on the probabilistic correspondence step, the voting scheme takes the remaining 30%.

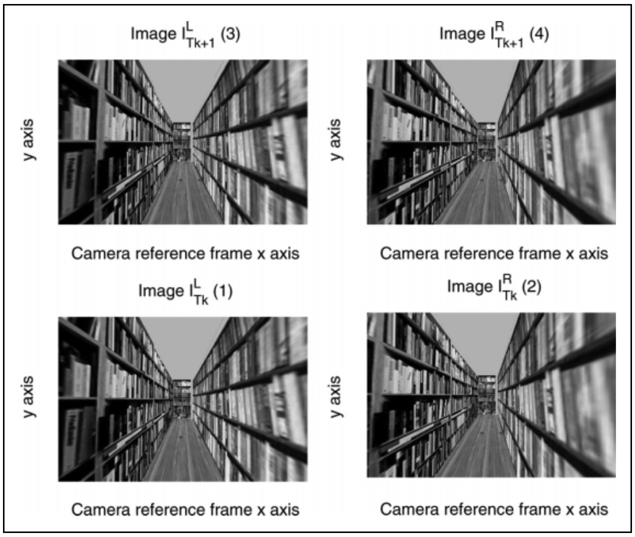

Figure 6. Synthetic images stereo pairs for translation scale motion estimation

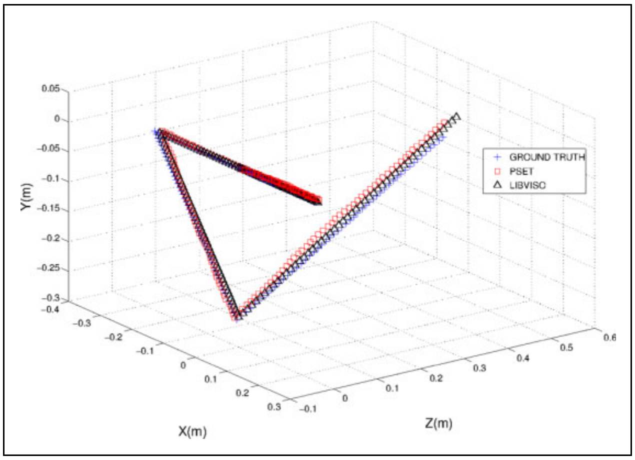

Figure 7. Generated and estimated trajectories in the synthetic image experiment.

Table 1. Comparison of the standard mean squared error between PSET and LIBVISO.

	V_x m/frame	V_y m/frame	V_z m/frame	$\|\|V\|\|$ m/frame
LIBVISO	0.000690	0.000456	0.0011	0.0022
PSET	0.000336	0.000420	0.000487	0.0012

PSET: probabilistic stereo egomotion transform.

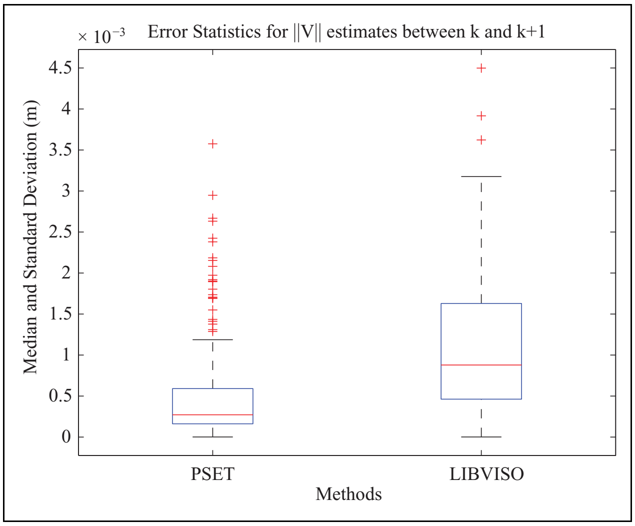

Figure 8. Error distribution $||V||$ obtained by PSET and LIBVISO. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. PSET: probabilistic stereo egomotion transform.

In Figure 7, one can observe the generated and the estimated trajectories obtained using PSET and LIBVISO.

From Table 1, we can observe that PSET obtains a more accurate egomotion estimation, having less root mean square (RMS) error than LIBVISO in all velocity components. This turns out to be more evident in the computation of the velocity norm over the global motion trajectory, where PSET results are almost 50% more accurate than the ones displayed by LIBVISO.

In this experiment, we focused on the evaluation of translational motion estimation, since the angular velocity case was already demonstrated by Silva et al.[35]

In Figure 8, we can observe a box plot of the instantaneous linear velocity error distribution during the sequence. It is clear better performance of PSET both in terms of the mean, median and variance of the error. Figure 9 shows the same information discriminated by coordinate axis where the same tendency is observed, especially for the X and Z components.

Real image sequences

The evaluation of PSET in real images was performed using KITTI data set[39] composed of stereo image sequences. The KITTI data set uses a car-vehicle robot in different road scenarios (urban street, countryside and highways) providing stereo image sequences in colour or greyscale format at 10 fps, 1.4-MP image resolution (1334 × 391) with IMU/GPS (OXTS RT 3003) information to act as external validation. In the study by Silva et al.,[35] PSET and LIBVISO were already compared using that data set. Results show that PSET outperforms LIBVISO in both linear and angular velocity estimation. In this work, we perform novel experiments with added Gaussian noise and image blur. Not only we want to evaluate the egomotion estimation accuracy in normal conditions but also with unfavourable image characteristics, typical of outdoor scenarios. The main argument we want to validate is that probabilistic methods, although requiring additional computations, can be more effective in robotic scenarios where

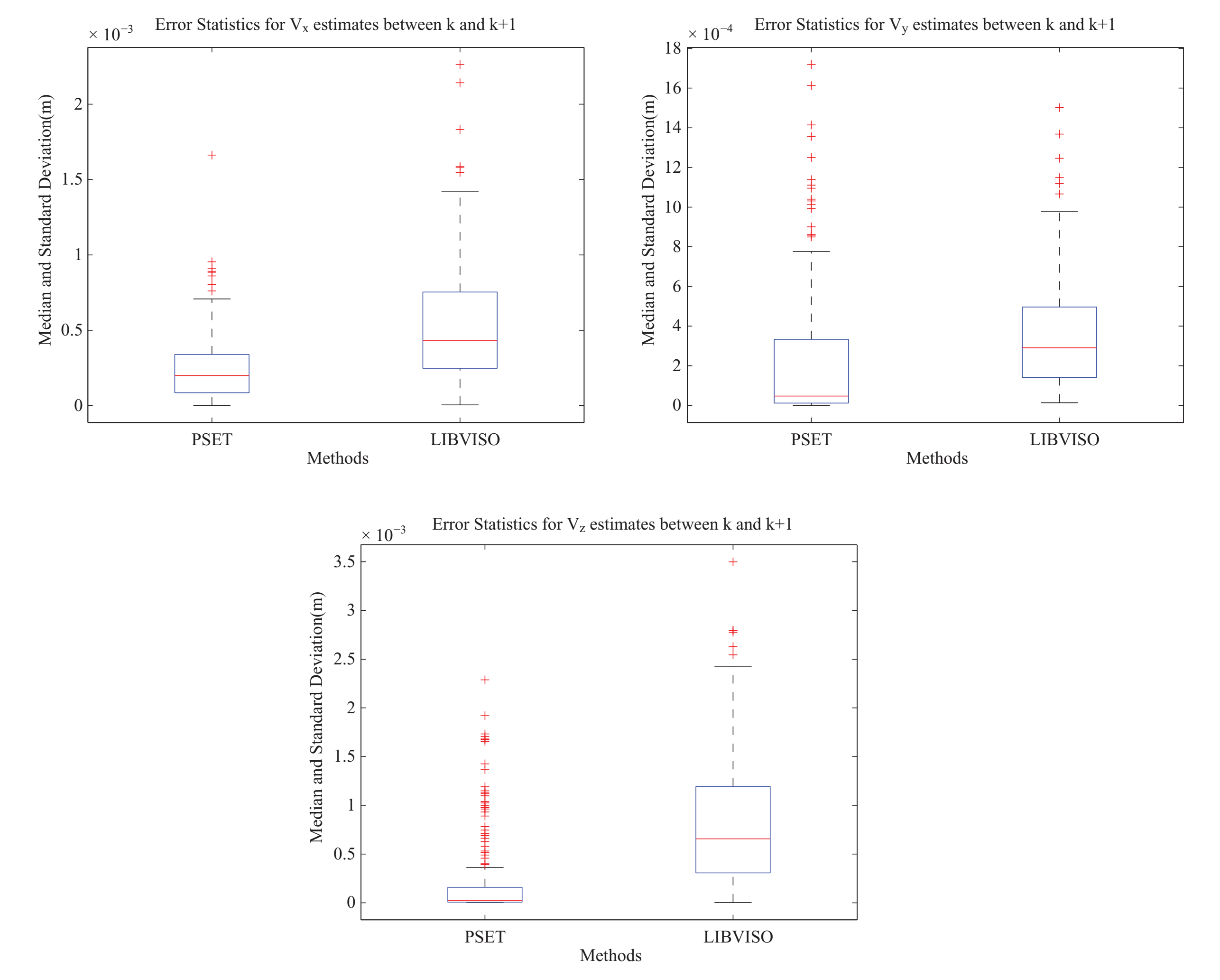

Figure 9. Error distribution of estimated linear velocities obtained by PSET and LIBVISO in all three axes (V_x, V_y, V_z). PSET: probabilistic stereo egomotion transform.

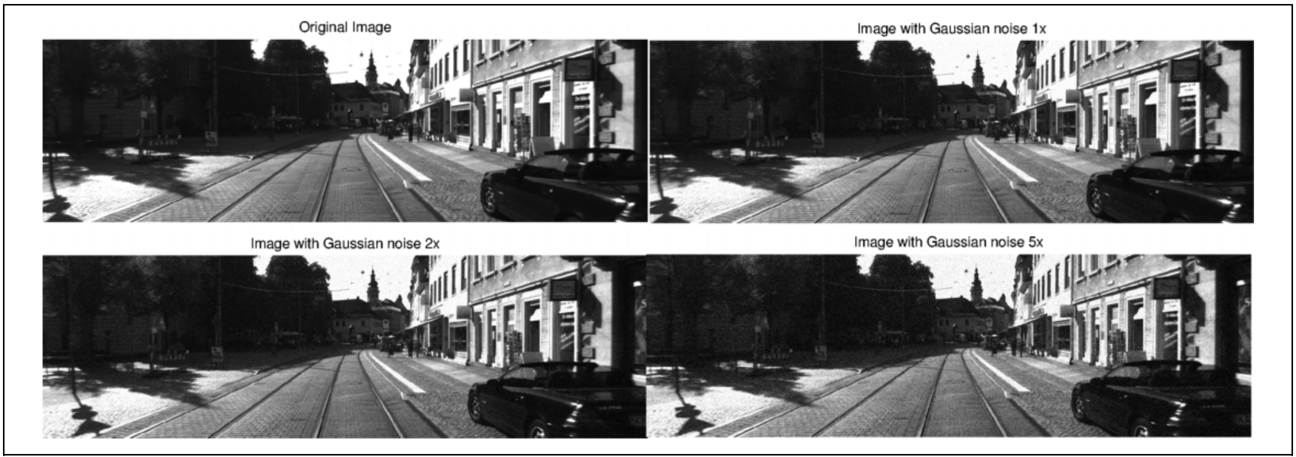
Original Image
Image with Gaussian noise 1x
Image with Gaussian noise 2x
Image with Gaussian noise 5x

Figure 10. Original image from KITTI data set drive 2011-09-26-0091, and corrupted versions with white Gaussian noise of variance 0.001, 0.002 and 0.005 in an image grey level range between 0 and 255.

Table 2. RMS for PSET and LIBVISO under different values of image Gaussian noise.

Gaussian noise	0×		1× (0.001)		2× (0.002)		5× (0.005)	
Egomotion	$\|\|V\|\|$	$\|\|W\|\|$	$\|\|V\|\|$	$\|\|W\|\|$	$\|\|V\|\|$	$\|\|W\|\|$	$\|\|V\|\|$	$\|\|W\|\|$
PSET	0.4170	0.9400	0.4436	0.9400	0.4556	0.9400	0.4899	0.9405
LIBVISO	0.4444	0.9605	0.5210	1.0068	0.5535	1.0600	0.6332	1.2712
Improvement	≈6%	≈2%	≈15%	≈7%	≈18%	≈12%	≈23%	≈26%

RMS: root mean square; PSET: probabilistic stereo egomotion transform.

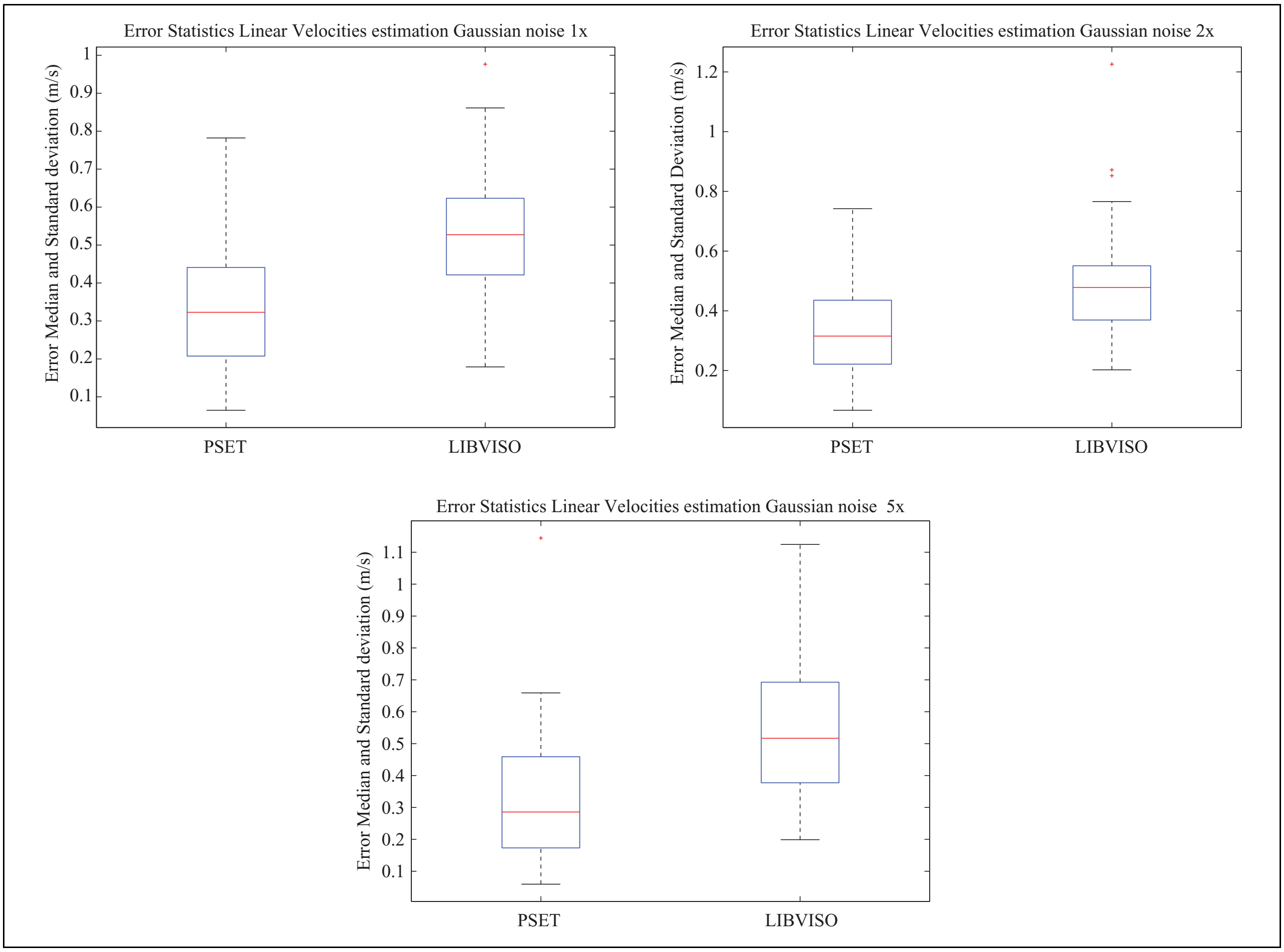

Figure 11. Error distribution of the magnitude of linear velocity computed by PSET and LIBVISO, images corrupted with Gaussian noise with variance 0.001, 0.002, 0.005, denoted, respectively, 1×, 2×, 5×. PSET: probabilistic stereo egomotion transform.

image conditions are far from ideal and deterministic egomotion estimation methods tend to fail.

Experiment with added Gaussian noise. In principle, probabilistic egomotion estimation methods are less sensible to image noise than their deterministic counterparts. In outdoor mobile robotic scenarios, images are frequently corrupted due to factors such as sensor noise, bad scene illumination, highlights, specular reflections and other optical artefacts. In order to validate probabilistic methods as more robust to image noise than deterministic methods, a set of experimental trials was performed. The experimental procedure consisted on adding white Gaussian noise to all the images (346 stereo pairs) in the KITTI stereo image data set sequence (drive 2011-09-26-0091),[39] since it is a scene that contains high contrast images and shadows and comparing egomotion estimation accuracy of both PSET and LIBVISO methods.

In Figure 10, we show an example of an image of the KITTI data set and the corresponding corrupted images with different values of Gaussian noise. Table 2 shows results for PSET and LIBVISO with three different noise powers: 0.001, 0.002 and 0.005 variance in grey level units in the 0–255 range.

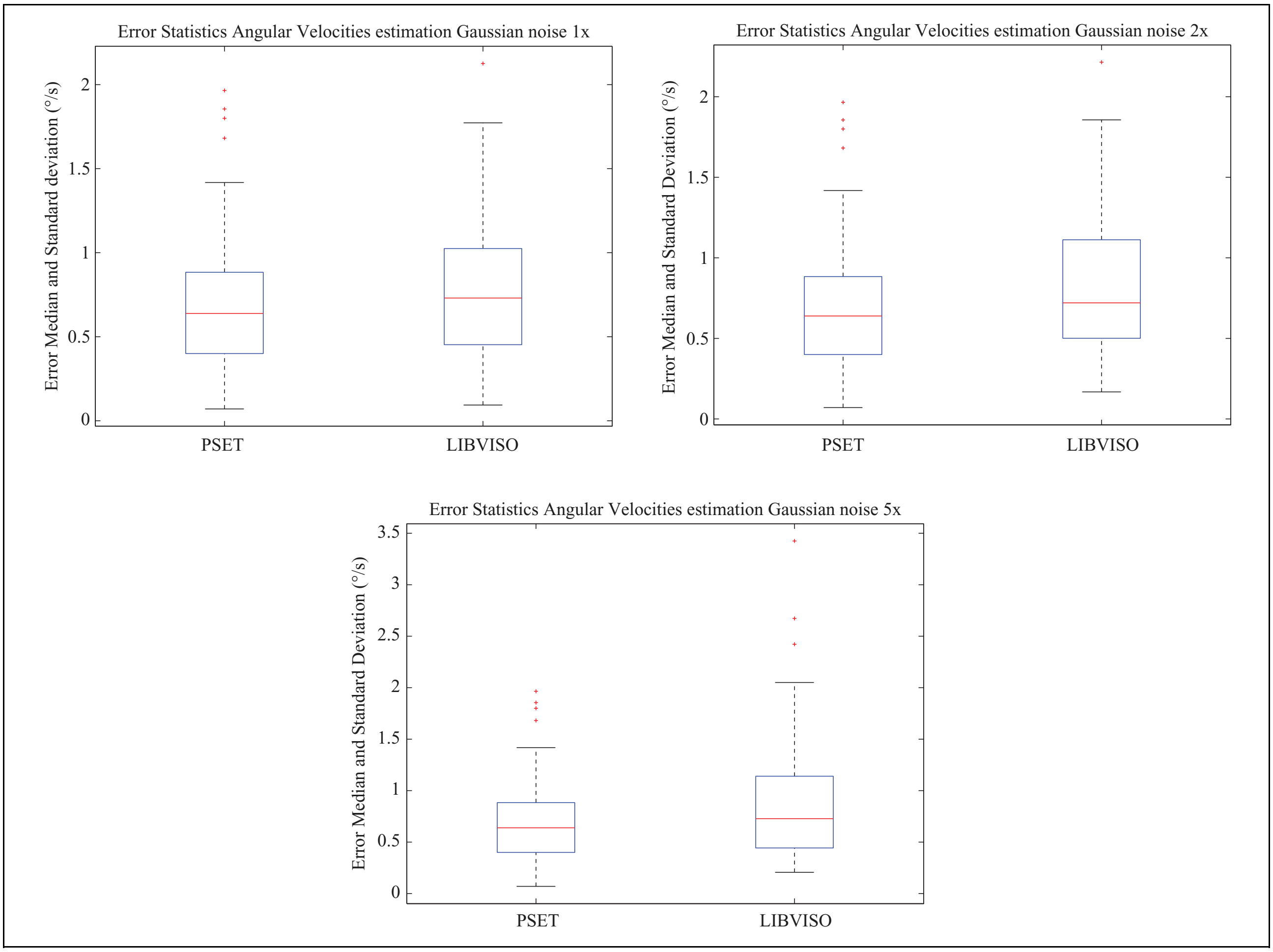

Figure 12. Error distribution of the magnitude of angular velocities computed by PSET and LIBVISO images corrupted with Gaussian noise with variance 0.001, 0.002, 0.005, denoted, respectively, 1×, 2×, 5×. PSET: probabilistic stereo egomotion transform.

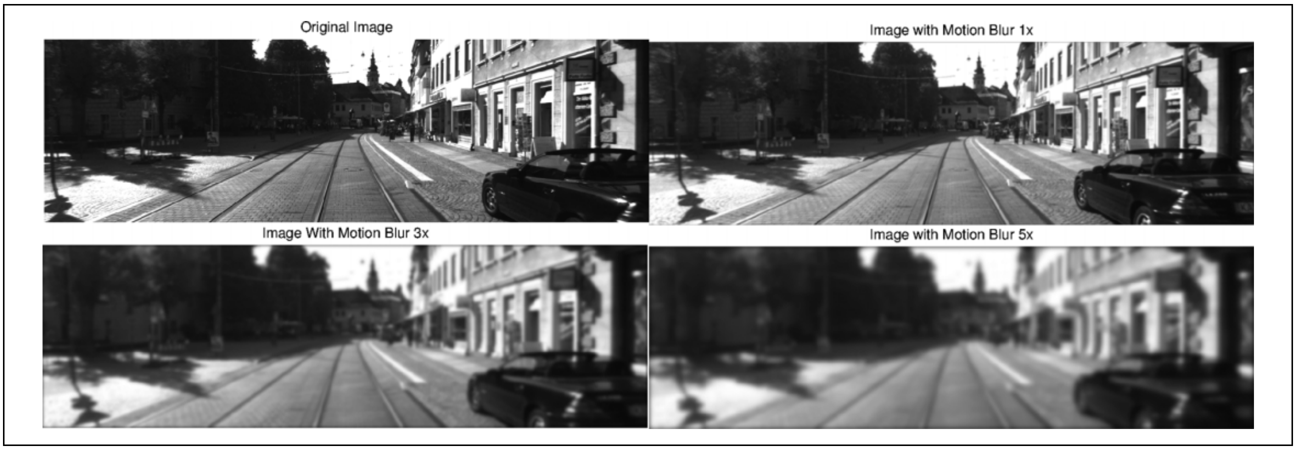
Original Image
Image with Motion Blur 1x
Image With Motion Blur 3x
Image with Motion Blur 5x

Figure 13. Original image from KITTI data set drive 2011-09-26-0091, and corrupted versions with blur 1, 3, 5 pixels standard deviation denoted as 1×, 2×, 5×.

We measure accuracy as the RMS error between the estimates and the IMU/GPS information. Quantitative results are shown in Table 2 and Figures 11 and 12. The obtained results show higher accuracy of the PSET method under all values of added Gaussian noise compared to LIBVISO. Furthermore, as the noise power grows, the PSET method shows bigger improvements. For the largest noise power tested, PSET reduces the

Table 3. RMS error for PSET and LIBVISO under different values of blur.

Image blur	0×		1× (1.0)		3× (3.0)		5× (5.0)	
Egomotion	$\|\|V\|\|$	$\|\|W\|\|$	$\|\|V\|\|$	$\|\|W\|\|$	$\|\|V\|\|$	$\|\|W\|\|$	$\|\|V\|\|$	$\|\|W\|\|$
PSET	0.4170	0.9400	0.4176	0.9400	0.4490	0.9400	0.4820	1.2965
LIBVISO	0.4444	0.9605	0.4475	1.0092	0.5535	1.1163	0.9400	1.9194
Improvement	≈6%	≈2%	≈7%	≈7%	≈19%	≈15%	≈48%	≈32%

RMS: root mean square; PSET: probabilistic stereo egomotion transform.

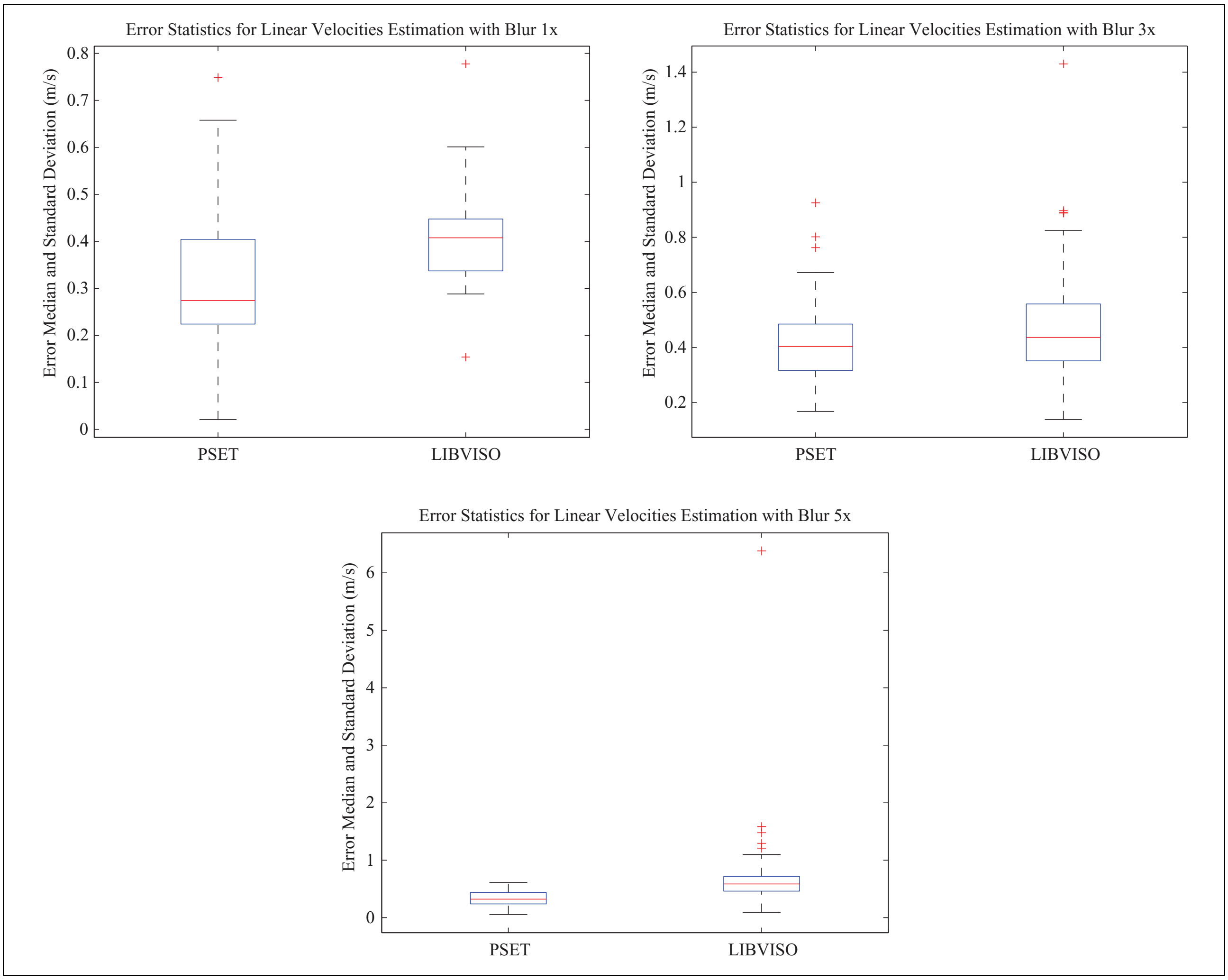

Figure 14. Error distribution of the magnitude of the linear velocities PSET and LIBVISO, images corrupted with a Gaussian blur filter with 1×, 2×, 5× pixels standard deviation. PSET: probabilistic stereo egomotion transform.

error in 23% for the linear velocities and 26% for the angular velocities.

In Figures 11 and 12, we show the error distribution of the linear and angular velocity magnitude computed by PSET and LIBVISO, for all tested error powers. The accuracy in the egomotion estimation obtained by PSET is higher, since it displays lower median error when compared to LIBVISO for all cases.

Experiment with added blur. In outdoor robotics scenarios, the presence of blur is somewhat frequent. The use of visual egomotion estimation in those scenarios was limited due to the fact that deterministic egomotion methods tend to fail in the presence of image blur. One of the reasons that justifies the use of probabilistic egomotion estimation methods is precisely the higher robustness exhibited by this type of approach when compared to deterministic methods in the presence of image blur. To validate such claim, we conducted another experiment using both PSET and LIBVISO in the same KITTI data set sequence (2011-09-

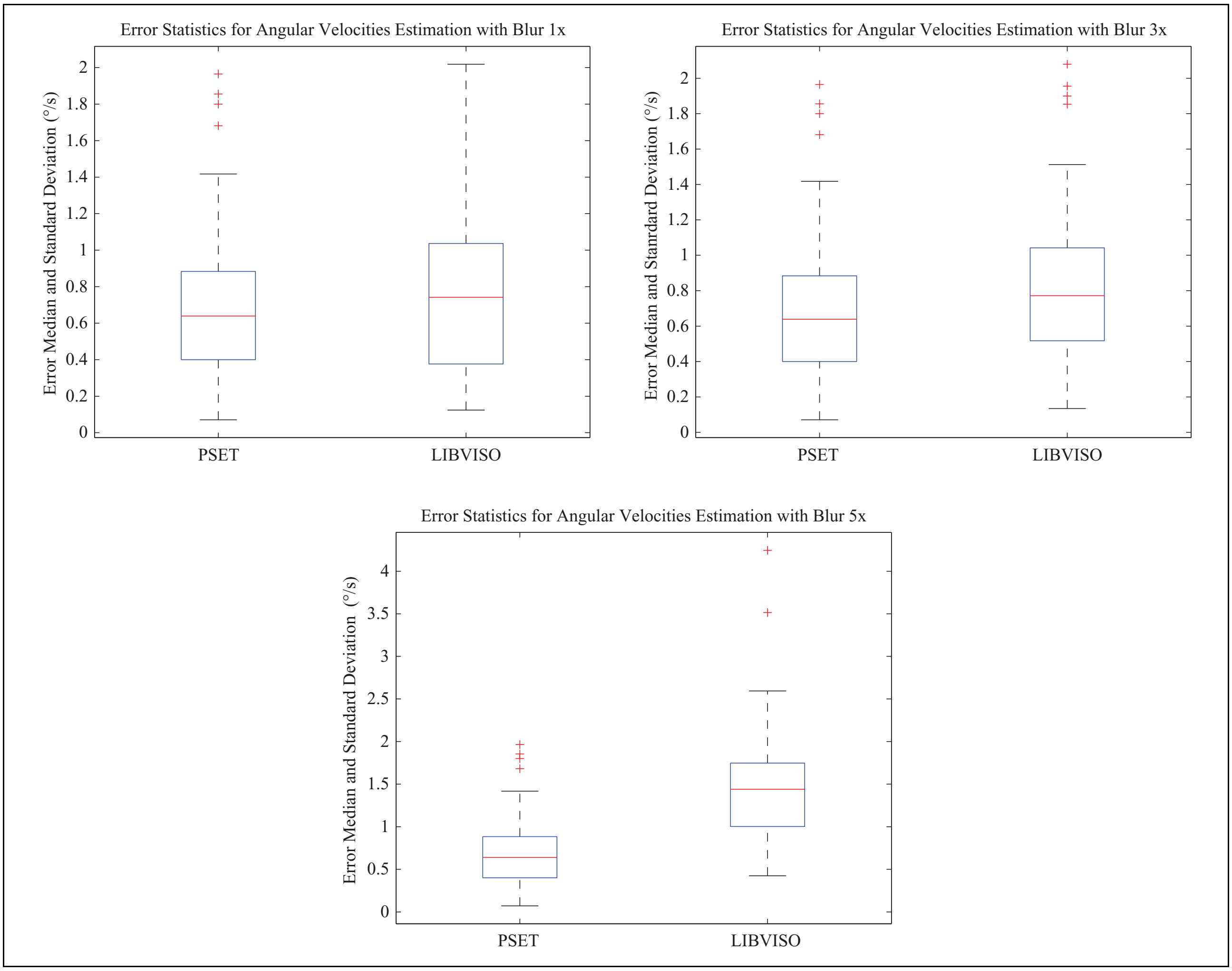

Figure 15. Error distribution of the magnitude of the angular velocities PSET and LIBVISO, images corrupted with a Gaussian blur filter with 1×, 2×, 5× pixels standard deviation. PSET: probabilistic stereo egomotion transform.

26-0091), but this time using different values of blur, as illustrated in Figure 13.

In Table 3, we show the RMS error using PSET and LIBVISO with different types of blur (1.0) when compared to IMU/GPS information. The corrupted images were created by adding a low-pass Gaussian filter of the image size with 1×, 3×, 5× standard deviation.

Again results show that PSET is more accurate than LIBVISO in the presence of higher quantities of blur. The error difference between PSET and LIBVISO increases from 7% for low values of blur (1×) to 48% and 32% in the linear and angular velocity estimation for high values of blur (5×).

In Figures 14 and 15, we can see the error distributions of linear and angular velocities both for PSET and LIBVISO. Again, PSET exhibits lower median error when compared to LIBVISO for all values of image blur.

The difference in the obtained accuracy from PSET and LIBVISO is bigger for the synthetic image scenario than in real image data sets. This fact maybe due to two causes. First the ground-truth information used in both experiments is different. In the synthetic image scenario, we generate the VRML simulation and therefore the world points have a precise position that provides a reliable egomotion trajectory verification. On the contrary for the real image data set, an IMU/GPS information is used. The IMU/GPS information is subject to bias and noise, and therefore, it can only be considered as weak ground truth information. Secondly, the KITTI sequence does not contain such image repetitive structure, and therefore, point correspondence ambiguity is lower.

Conclusions and future work

The probabilistic approach for stereo visual egomotion estimation described in this work has proven to be an accurate method of computing stereo egomotion. The proposed approach is very robust because no explicit matching or feature tracking is necessary to compute the vehicle motion. To the best of our knowledge, this is the first implementation of a fully dense probabilistic method to compute stereo egomotion. The results demonstrate that

PSET is more accurate than other state-of-the-art 3D egomotion estimation methods, significantly improving the overall accuracy in linear and angular velocity estimation. We have shown improvements up to 50% in a highly repetitive texture synthetic image scenario with ground truth information and above 20% in real images with large amounts of blur and noise with respect to IMU/GPS reference. One of the main advantage of probabilistic egomotion estimation methods is their higher robustness in difficult imaging scenarios, for example, in the presence of image noise or blur. In the experiments, conducted PSET achieved a better performance than LIBVISO and the improvement (error difference) between both methods increased in the presence of higher values of image noise and blur. Despite the clear advantages over other state-of-the-art methods, its effectiveness and usefulness in mobile robotics scenarios requires further improvements on the computational implementations in order to have real-time functionality. Given the highly parallel nature of the algorithm, composed of many independent operations, in future work, we plan to develop a PSET GPU implementation to achieve real-time performance. Another objective is to pursue further validation of the PSET algorithm in other heterogeneous mobile robotics scenarios, especially in aerial and underwater robotics, where the lack of image texture combined with high matching ambiguity provides an ideal scenario for further accessing the robustness of the proposed methodology.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is financed by the Coral project NORTE-01-0145-FEDER-000036, and by National Funds through the FCT -Fundacao para a Ciencia e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013 and project UID/EEA/50009/2013.

References

1. Silva H, Bernardino A and Silva E. Probabilistic stereo egomotion transform. In: *IEEE international conference on robotics and automation*, Hong Kong, May 31–7 June 2014.
2. Scaramuzza D. Performance evaluation of 1-point-RANSAC visual odometry. *J Field Robot* 2011; 28(5): 792–811.
3. Maimone M, Matthies L and Cheng Y. Visual odometry on the Mars Exploration Rovers. In: *IEEE international conference on systems, man and cybernetics*, Hawaii, November 2005.
4. Maimone M, Matthies L and Cheng Y. Two years of visual odometry on the mars exploration rovers: field reports. *J Field Robot* 2007; 24(3): 169–186.
5. Förstner W and Gülch E. A Fast Operator for Detection and Precise Location of Distinct Points, Corners and Centres of Circular Features, ISPRS Intercommission Workshop, 1987.
6. Harris C and Stephens M. A combined corner and edge detection. In: *Proceedings of the fourth Alvey vision conference*, 1988, pp. 147–151.
7. Fischler MA and Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communicat ACM* 1981; 24(6): 381–395.
8. Olson C, Matthies L, Schoppers M, et al. Rover navigation using stereo ego-motion. *Robot Autonom Syst* 2003; 43; 215–229.
9. Goodall C. Procrustes methods in the statistical analysis of shape. *J Roy Stat Soc Ser B (Methodological)* 1991; 53(2): 285–339.
10. Horn BKP. Closed-form solution of absolute orientation using unit quaternions. *J Opt Soc Am A* 1987; 4(4): 629–642.
11. Rusinkiewicz S and Levoy M. Efficient variants of the ICP algorithm. In: *Proceedings of the Third international conference on 3D digital imaging and modeling (3DIM)*, 2001, pp. 145–152.
12. Milella A and Siegwart R. Stereo-based ego-motion estimation using pixel tracking and iterative closest point. In *Fourth IEEE International Conference on Computer Vision Systems (ICVS'06)*, New York, 4–7 January 2006.
13. Alismail H, Browning B and Dias MB. Evaluating pose estimation methods for stereo visual odometry on robots. In: *Proceedings of the 11th international conference on intelligent autonomous systems (IAS-11)*, Ottawa, Canada, January 2011.
14. Nister D, Naroditsky O and Bergen J. Visual odometry for ground vehicle applications. *J Field Robot* 2006; 23(1): 3–20.
15. Haralick RM, Lee CN, Ottenberg K, et al. Review and analysis of solutions of the three point perspective pose estimation problem. *Int J Comput Vision* 1994; 13: 331.
16. Kai N and Dellaert F. Stereo tracking and three-point/one-point algorithms - a robust approach, visual odometry. In: *International conference on image processing (ICIP)*, Atlanta, USA, October 2006.
17. Ni K, Dellaert F and Kaess M. Flow separation for fast and robust stereo odometry. In: *IEEE international conference on robotics and automation*, Kobe, Japan, 12–17 May 2009.
18. Kitt B, Geiger A and Lategahn H. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In: *IEEE intelligent vehicles symposium (IV)*, San Diego, USA, June 2010.
19. Hartley RI and Zisserman A. *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press, 2004. ISBN: 0521540518.
20. Comport A, Malis E and Rives P. Real-time quadrifocal visual odometry. *Int J Robot Res SAGE Publicat* 2010; 29: 245–266.
21. Scaramuzza D, Fraundorfer F and Siegwart R. Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC. *IEEE Int Conf Robot Autom*, Kobe, Japan, 12–17 May 2009.
22. Kneip L, Chli M and Siegwart R. Robust real-time visual odometry with a single camera and an IMU. In: *Proceedings*

of the British machine vision conference (BMVC), September 2011, pp. 16.1–16.11. BMVA Press.

23. Voigt R, Nikolic J, Huerzeler C, et al. Robust embedded egomotion estimation. In: *Proceeding of the IEEE RSJ international conference on intelligent robots and systems (IROS)*, San Francisco, 25–30 September 2011.
24. Rehder J, Gupta K, Nuske S, et al. Global pose estimation with limited GPS and long range visual odometry. *IEEE Conf Robot Autom*, St Paul, USA, 14–18 May 2012.
25. Kazik T, Kneip L, Nikolic J, et al. Real-time 6D stereo visual odometry with non-overlapping fields of view. In: *IEEE International conference on computer vision and pattern recognition*, Providence RI, USA, 16–21 June 2012.
26. Scaramuzza D and Fraundorfer F. Visual odometry tutorial. *IEEE Robot Autom Magaz* 2011; 18(4): 80–92.
27. Kneip L, Scaramuzza D and Siegwart R. A novel parameterization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In: *Proceedings CVPR '11 Proceedings of the 2011 IEEE conference on computer vision and pattern recognition (CVPR)*, 2011, pp. 2969–2976.
28. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vision* 2004; 60: 91.
29. Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF). *Comput Vision Image Understand Elsevier Sci Inc* 2008; 110: 346–359.
30. Calonder M, Lepetit V, Strecha C, et al. Brief: binary robust independent elementary features. In: *European conference on computer vision*, Crete, Greece, 5–11 September 2010.
31. He H, Li Y, Guan Y, et al. “Wearable ego-motion tracking for blind navigation in indoor environments,” *IEEE Trans Autom Sci Eng* 2015; 12(4): 1181–1190.
32. Domke J and Aloimonos Y. A Probabilistic notion of correspondence and the epipolar constraint. In: *Proceeding 3DPVT ‘06 Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT‘06)*, 2006, pp. 41–48.
33. Huang J, Zhu T, Pan X, et al. A high-efficiency digital image correlation method based on a fast recursive scheme. *Measur Sci Technol* 2011; 21(3): 35101–35112.
34. Nelder JA and Mead RA. Simplex method for function minimization. *Comput J* 1965; 7(4): 308–313.
35. Silva H, Bernardino A and Silva E. Probabilistic egomotion for stereo visual odometry. *Int J Intell Robot Syst* 2015; 77(2): 265–280.
36. Wand MP and Jones MC. Kernel smoothing. In: *Chapman Hall CRC Monographs on Statistics Applied Probability*, 1994.
37. Debella-Gilo M and Kaab A. Sub-pixel precision image matching for measuring surface displacements on mass movements using normalized cross-correlation. *Remote Sens Environ* 2011; 115: 130–142.
38. Craig JJ. *Introduction to robotics: mechanics and control*. Harlow: Addison-Wesley Longman Publishing Co, Inc, 1989.
39. Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset. *Int J Robot Res (IJRR)* 2013; 32(11): 1–6.