Probabilistic Stereo Egomotion Transform

Hugo Silva, Eduardo Silva
INESC TEC
School of Engineering, Polytechnic Institute of Porto
Email:hugo.m.silva,eduardo.silva@inescporto.pt

Alexandre Bernardino
Institute for Systems and Robotics
IST Lisbon, Portugal
Email: alex@isr.ist.utl.pt

***Abstract*—In this paper we propose a novel fully probabilistic solution to the stereo egomotion estimation problem. We extend the notion of probabilistic correspondence to the stereo case which allow us to compute the whole 6D motion information in a probabilistic way. We compare the developed approach against other known state-of-the-art methods for stereo egomotion estimation, and the obtained results compare favorably both for the linear and angular velocities estimation.**

I. Introduction

The use of mobile robots on modern world tasks is increasing rapidly as well as their application scenarios. One of the robots most complex task is navigation where typically GPS-IMU sensor information is used. Some of these scenarios(e.g urban areas, underwater GPS denied environments) are prone to GPS-IMU failures, making it necessary to use other alternative or complementary sensors such as vision cameras for the robot to perceive and navigate through out the environment. When using visual sensors (cameras), robots must determine motion measuring their displacement relative to static key points in the environment, process which is usually denoted as Visual Odometry(VO). The use of VO methods for obtaining robot motion has been continuously subject of research by the robotics and automotive industry over the past years. One way of performing VO estimation is by determining instantaneous camera displacement on consecutive frames, a process denoted as visual egomotion estimation, and integrating over time the obtained rotational and translational velocities. In monocular egomotion estimation there is translation scale ambiguity,, i.e. in the absence of other sources of information, only the translational velocity direction is possible to measure reliably. Therefore, whenever possible two cameras are used to have a full velocity estimation, usually denoted as stereo egomotion estimation e.g([1], [2], [3], [4]).

Most approaches to the stereo egomotion estimation problem, rely on non-probabilistic correspondences methods. Common approaches try to detect, match, and track key points between images on adjacent time frames and afterwards use the largest subset of point correspondences that yield a consistent motion. In probabilistic correspondence methods matches are not fully committed during the initial phases of the algorithm and multiple matching hypotheses are accounted for. Our previous work in egomotion estimation (6DP) [1] [2], has shown the probabilistic correspondence methods were a viable way to estimate egomotion with advantages in precision over classical feature based methods. Nevertheless 6DP method was unable to estimate the translation scale factor based only on probabilistic approaches, and required a mixed approach to be able to recover all motion parameters.

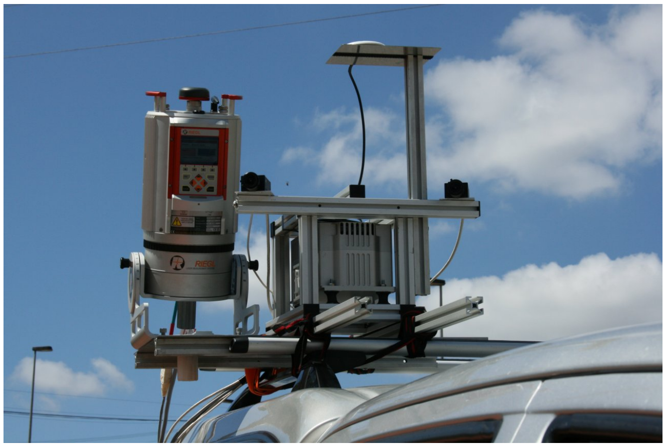

Fig. 1: Acquisition Setup for a vehicle-like robot, with the use of stereo cameras for providing estimates of vehicle angular and linear velocities.

In this paper, we developed a novel probabilistic stereo egomotion method(PSET) capable of computing 6-DOF motion parameters solely based on probabilistic correspondence approaches, and without the need to track or commit key point matches between consecutive frames. The use of probabilistic correspondence methods allows to maintain several match hypothesis for each point, which is an advantage when there are ambiguous matches (which is the rule in image feature correspondences problems), because no commitment is made before analyzing all image information. Another advantage is that a full probabilistic distribution of motion provides a better sensor fusion with other sensors, e.g. inertial.

Our proposed approach improves the work conducted in [1], [2] and propose a fully probabilistic algorithm to perform stereo egomotion estimation, which we denote as probabilistic stereo egomotion transform(PSET). While in 6DP [1], a probabilistic and deterministic approach was used to estimate rotation and translation parameters, PSET only employs probabilistic correspondences. The rotation estimation is achieved the same way as in 6DP(with a 5D search over the motion space based on the notion of epipolar constraint), yet the translation scale factor is obtained by exploiting an accumulator array voting scheme based also on epipolar stereo geometry combined with probabilistic distribution hypotheses between the two adjacent stereo image pairs. The obtained results demonstrate a clear performance improvement in the estimation of the linear and angular velocities over current state-of-the-art stereo egomotion estimation methods, when compared to Inertial Measurement Unit ground-truth information. Furthermore, since real-time is a concern in today modern mobile robotics applications the algorithm is easily paralellizable.

This paper is organized as follows: In section II some related work is presented. In section III, we discuss the monocular approach to probabilistic egomotion estimation. Afterwards in section IV we extend this approach to the stereo egomotion estimation case. In section V results of the PSET Transform compared with other state of the art stereo egomotion estimation methods are presented. Finally, section VI contains the conclusions with final remarks and future work.

II. Related Work

One way to perform stereo VO, is to estimate the instantaneous velocity (egomotion) of the vehicle or robot, from a sequence of images acquired using a stereo camera setup and integrate the velocity measurement in time. The main advantage of having a stereo camera configuration setup is the ability to easily recover the translation motion scale. Most of the known stereo VO methods use the 3D position of observed image key points that are computed by triangulating their relative position between stereo pair images. Afterwards, camera motion can be computed based on the alignment of the same 3D key points position between adjacent time steps. The research work on stereo VO estimation started with Olson *etal* [5], when visual sensors became an alternative to the dead reckoning estimation(e.g wheel odometry), that were unable to estimate robot motion over long distances, and accumulated large errors specially due to wheel slip in uneven terrain. However, the main boost in the development of stereo VO methods was given by Cheng *et al* [6] with the Mars Rover Project. The proposed method estimated all 6-DOF motion parameters, by tracking image key points between left and right image pairs in consecutive time frames, and determine the change of key points position and attitude using maximum likelihood estimation.

Currently stereo VO methods are classified according to either feature detection scheme or by the way motion estimation is performed [7]. In Alismail *et al* [8], a study for evaluating Absolute Orientation(AO) methods and Perspective-n-Point methods(PnP) for achieving robot pose estimation using only stereo visual odometry is conducted, and concluded using several outdoor and indoor datasets that PnP methods are more accurate than AO methods. The AO methods consist on triangulating 3D key points for every stereo pair. Then motion estimation is solved using key point alignment algorithms such as the Procrustes method [9], used in [1] or Iterative-Closest-Point(ICP) method [10], used in Milella and Siegwart [11] for estimating motion of an all-terrain rover. Nister *et al* [12], developed the first PnP algorithm (3D-2D camera pose estimation), computed in real-time with an outlier rejection scheme, that minimized the re-projection error.

What concerns image information, stereo VO methods may use sparse or dense approaches. One of the most relevant dense stereo VO applications was developed by Howard [13] for ground vehicle applications. The method makes no assumption of prior knowledge over camera motion, and so it can handle very large image translations. However, due to the absence of feature detectors invariant to rotation and scaling, it only works on low-speed applications with high frame-rate, and large motions around the optical axis result in poor performance. In [14] a sparse stereo VO method is presented. A closed form solution is derived for the incremental movement of the cameras and combines distinctive features invariant to rotation and scale (SIFT) [15] with sparse optical flow Lucas-Kanade Tracker [16]. Some authors like Ni *et al* [17], minimize dependencies on feature matching and tracking algorithms by simultaneously using an algorithm that computes feature displacement in both cameras, together with a quadrifocal setting within a RANSAC [18] framework. Later on, the same authors [19], decoupled the rotation and translation recovery into two different estimation problems. Instead of using the three-point method, they used a RANSAC two-point algorithm for rotation recovery and a one-point method for the translation recovery.

More recently the application focus of stereo VO methods has moved from planetary rover application to the development of novel intelligent vehicles by the automotive industry. Obdrzalek *et al* [20] developed a voting scheme strategy for egomotion estimation, where 6-DOF problem was divided into a four dimensions problem and then decomposed in two sub-problems for rotation and translation estimation. Another influential work, is the one developed by Kitt *et al* [3]. Their method, is available as an open-source visual odometry library named LIBVISO. Stereo egomotion estimation is based on image triples and the online estimation of the trifocal tensor [21]. It uses rectified stereo image sequences and produces an output 6D vector with estimated linear and angular velocities. Comport *et al* [4] also develop a stereo VO method based on the quadrifocal tensor. By using tensor notation, the authors can compute motion using 2D-2D image pixels matches, thus yielding a more precise motion estimation. Another way of developing stereo VO is to combine with other absolute sensor information. Rehder *et al* [22] developed a stereo visual odometry method that combined visual data with GPS and IMU information. The proposed method consistently fused stereo visual odometry information with inertial measurements and sparse GPS information into a single pose estimate in real-time. Kneip *et al* [23] also proposed an alternative tightly coupled approach with vision and IMU information. Their strategy for continuous robust pose computation is based on the triangulation of frame to frame point clouds when there is sufficient disparity among them. More recently Kazik *et al* [24] developed a framework that performed 6-DOF absolute scale motion with a stereo setup that copes with non-overlapping fields of view in indoor environments. It estimates monocular VO from each camera and afterwards scale is recovered by imposing the known stereo rig transformation between both cameras. Finally, there are already stereo egomotion estimation methods implemented on GPGPU, such as the one developed by Isvtan *et al* [25].

In our previous related work [1], [2] a probabilistic correspondence method developed by [26] was used to compute the stereo egomotion estimation. The results obtained show an improvement in the angular velocities estimation, but not in the linear velocities estimation where scale was obtained using deterministic matches instead of probabilistic correspondences. In this paper, we will extend the probabilistic correspondence approach to the linear velocities estimation case.

III. Probabilistic Egomotion Estimation

The seminal work of Domke *et al* [26], has introduced the notion of probabilistic correspondence in the context of the egomotion estimation problem. In this section we provide a brief description of their method and introduce the notation required for the remaining sections.

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

$$\rho_s(q) = \text{match}(s,q)$$

where the function match$(\cdot)$ outputs a value between 0 and 1 expressing similarity in the appearance of the two points in local neighborhoods.

In [26] that 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*) to measure point similarity:

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

where W denotes a $2D$ window centered at the origin whose size defines the neighborhood of analysis around points s and q.

In practice we use a fast recursive implementation of the ZNCC developed by Huang *et al* [27]. The probabilistic correspondence is then computed as:

$$\rho_s(q) = \frac{ZNCC(s,q) + 1}{2}$$

so that its values range from 0 to 1.

Afterwards, motion hypotheses defined by the incremental rotation matrix R and the translation velocity direction $\hat{t} = t/\|t\|$ are checked by analyzing the probabilistic correspondences $\rho_s(q)$ along the epipolar lines [26]. A correspondence q for point s must satisfy the epipolar constraint denoted by:

$$\tilde{s}^T E \tilde{q} = 0 \tag{1}$$

where $\tilde{s}$ and $\tilde{q}$ are the homogeneous representations of s and q, respectively. Matrix E is the so called Essential Matrix, a 3×3 matrix of rank 2 and 5 degrees of freedom that encodes the rigid camera motion:

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

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

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

In order to obtain the Essential Matrix (E) from the probabilistic correspondences [26] proposes the computation of a probability distribution over the 5-dimensional space of essential matrices. Each dimension of the space is discretized in 10 bins, thus leading to 100000 hypotheses E_i. For each point s the likelihood of these hypotheses is evaluated by:

$$\rho(E_i|s) \propto \max_{(\tilde{\mathbf{q}})^T E_i \tilde{\mathbf{s}} = 0} \rho_s(q) \tag{2}$$

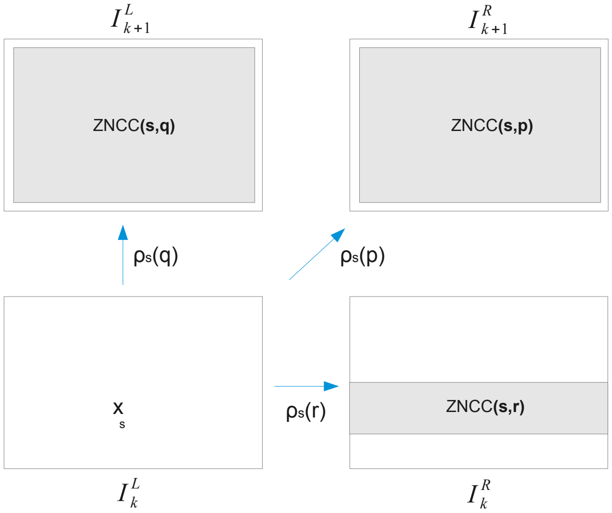

Fig. 2: ZNCC matching used to compute the PSET transform

Therefore for a single point s in image I^L_k, the likelihood of a motion hypothesis (E_i) is proportional to the likelihood of the best match obtained along the epipolar line generated by the essential matrix. If one assumes statistical independence between the measurements obtained at each point the overall likelihood of a motion hypothesis is proportional to the product of the likelihoods for all points:

$$\rho(E_i) \propto \prod_s \rho(E_i|s) \tag{3}$$

Finally, given the top ranked motion hypotheses, a Nelder-Mead simplex method [28] is used to refine the motion estimate, and obtain the best motion up to a scale factor that represents the image movement between the two frames.

IV. Probabilistic Stereo Egomotion Estimation

In this work we extend the notion of probabilistic correspondence to the stereo case which allow us to compute the whole 6D motion information in a probabilistic way. In a stereo setup we consider images I^L_k, I^L_{k+1}, I^R_k and I^R_{k+1}, where superscripts L and R denote respectively the left and right images of the stereo pair. Probabilistic matches of a point s in I^L_k are now computed not only for points q in I^L_{k+1} but also for points r in I^R_k and p in I^R_{k+1} (see Fig. 2):

$$\rho_s(r) = \frac{ZNCC(s,r) + 1}{2}$$

$$\rho_s(p) = \frac{ZNCC(s,p) + 1}{2}$$

For the sake of computational efficiency, analysis can be limited to sub-regions of the images given prior knowledge about the geometry of the stereo system or the motion given by other sensors like IMU's. In particular, for each point s, coordinates r can be limited to a band around the epipolar lines according to the stereo setup epipolar geometry.

A. *The Geometry of Stereo Egomotion*

In this section we describe the geometry of the stereo egomotion problem, i.e. will analyze 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 scale factor.

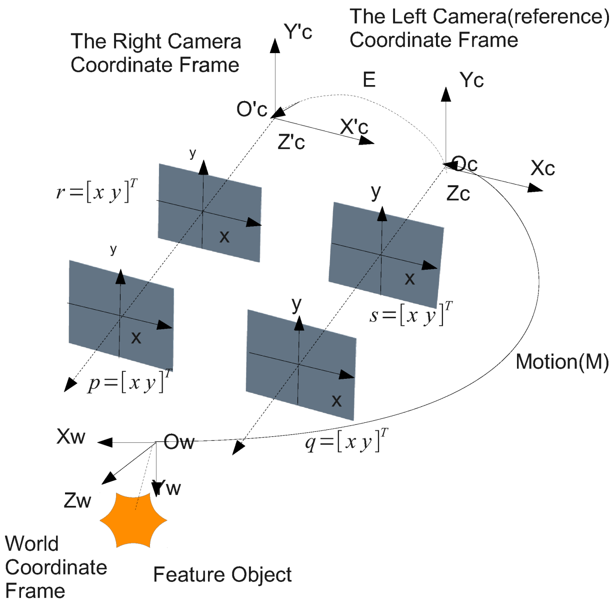

Fig. 3: Stereo Egomotion Geometry

Let us consider the 4×4 rototranslations T_L^R and M_k^{k+1} 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 $k+1$:

$$T_L^R = \begin{bmatrix} R_L^R & t_L^R \\ 0 & 1 \end{bmatrix} \quad M_k^{k+1} = \begin{bmatrix} R_k^{k+1} & t_k^{k+1} \\ 0 & 1 \end{bmatrix}$$

where $R_\cdot^\cdot$ and $t_\cdot^\cdot$ denote the rotational and translational components. We factorize the translational motion t_k^{k+1} in its direction $\hat{t}$ and amplitude α:

$$t_k^{k+1} = \alpha \hat{t}$$

Given that rotational motion and translation direction are computed by the method described in the previous section, the computation of α is the objective to pursue.

Let us consider an arbitrary 3D point $X = (X_x, X_y, X_z)^T$ expressed in the reference frame of the left camera 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 X in the 4 images is given by:

$$\begin{cases} \tilde{s} = X \\ \tilde{r} = R_L^R X + t_L^R \\ \tilde{q} = R_k^{k+1} X + \alpha \hat{t} \\ \tilde{p} = R_L^R R_k^{k+1} X + \alpha R_L^R \hat{t} + t_L^R \end{cases} \tag{4}$$

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 the 3×3 identity ($R_L^R = I_{3\times 3}$) and the translation vector is an offset (baseline b) along the x coordinate, $t_L^R = (b, 0, 0)^T$. In this case, expanding the equations for $s = (s_x, s_y)^T$ and $r = (r_x, r_y)^T$ we obtain:

$$\begin{cases} s_x = \frac{X_x}{X_z} \\ s_y = r_y = \frac{X_y}{X_z} \\ r_x = \frac{(X_x + b)}{X_z} \end{cases}$$

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 X as a function of the image coordinates r and s and the known baseline value b:

$$X = \left(\frac{s_x b}{d} \quad \frac{s_y b}{d} \quad \frac{b}{d}\right)^T$$

Replacing this value now in (4) we obtain:

$$r = \begin{bmatrix} \frac{(\frac{s_x b}{d} + b)d}{b} \\ s_y \end{bmatrix} \tag{5}$$

$$q = \begin{bmatrix} \frac{r11 s_x b + r12 s_y b + r13 b + \alpha t_x d}{r31 s_x b + r32 s_y b + r33 b + \alpha t_z d} \\ \\ \frac{r21 s_x b + r22 s_y b + r23 b + \alpha t_y d}{r31 s_x b + r32 s_y b + r33 b + \alpha t_z d} \end{bmatrix} \tag{6}$$

$$p = \begin{bmatrix} \frac{r11 s_x b + r12 s_y b + r13 b + \alpha t_x d + bd}{r31 s_x b + r32 s_y b + r33 b + \alpha t_z d} \\ \\ \frac{r21 s_x b + r22 s_y b + r23 b + \alpha t_y d}{r31 s_x b + r32 s_y b + r33 b + \alpha t_z d} \end{bmatrix} \tag{7}$$

We determine the translation scale factor α, using (6) by:

$$q_x = \frac{\overbrace{r11 s_x b + r12 s_y b + r13 b}^{\text{A}} + \alpha t_x d}{\underbrace{r31 s_x b + r32 s_y b + r33 b}_{\text{C}} + \alpha t_z d} \tag{8}$$

being α given by:

$$\alpha = \frac{A - q_x C}{q_x t_z d - t_x d} \tag{9}$$

The same procedure is applied to q_y:

$$q_y = \frac{\overbrace{r21 s_x b + r22 s_y b + r23 b}^{\text{B}} + \alpha t_y d}{\underbrace{r31 s_x b + r32 s_y b + r33 b}_{\text{C}} + \alpha t_z d} \tag{10}$$

being α given by:

$$\alpha = \frac{B - q_y C}{q_y t_z d - t_y d} \tag{11}$$

The translation scale factor α, can also be determined using the same procedure applied to (7). Therefore, being α an over-determined parameter since there are four equations to one unknown, we choose the α with the highest denominator to minimize the effect of numerical errors. In case both denominators are low due to very low disparity or degenerate motions, this particular point can not be used for the estimation.

B. Translational Scale Estimation

In the previous section we have seen that is is possible to estimate the translational scale α from the observation of a single static point s, if its point correspondences r, q and p are known and there are no degeneracies. 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.

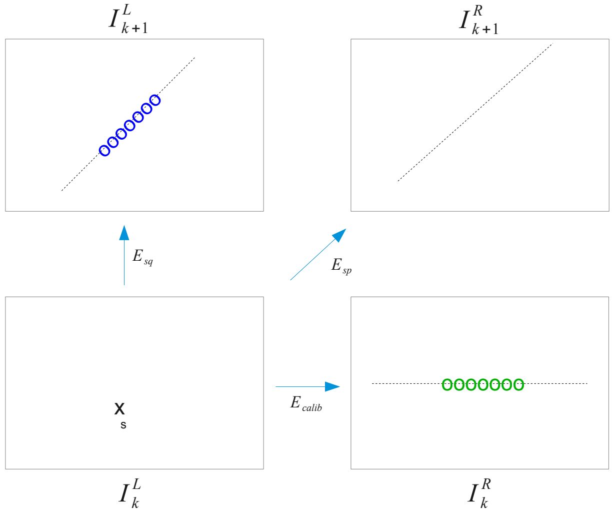

Fig. 4: Point correspondence hypotheses along the epipolar lines

In our previous work [1], this was achieved by computing the rigid transformation between point clouds obtained from stereo reconstruction at times k and $k+1$ with a robust method RANSAC [18]. 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 $k+1$) see Fig.4.

In the current work, we extend the probabilistic notion of correspondence to the stereo case. Instead of deterministically committing to matches in space and time, we create a probabilistic observation model for possible matches:

$$P_{match}(s,r,p,q) = \rho_s(r)\rho_s(q)\rho_s(p)$$

where we assume statistical independence in the measurements obtained in the pairwise probabilistic correspondence functions $\rho_s(\cdot)$. Then, because each possible match (s,r,p,q) will correspond to a value of α, we will create an accumulator of α hypotheses, weighted by $P_{match}(s,r,p,q)$. Searching for peaks in the accumulator will provide us the best (most agreed) hypothesis for α given all the information in the images.

C. The Probabilistic Stereo Egomotion Transform

Here we detail how the method is implemented computationally. We assume E has been computed by the methods described previously and the system calibration is known.

First a large set of points $s_j, j = 1\cdots J$ is selected. Selection can be random, uniform or based on key points, e.g. Harris corners [29] or Scale-Invariant features [15]. In our current implementation SIFT features are used to detect salient points.

For each point s_j, the epipolar lines $E_{calib} = \tilde{s}_j^T S$ and $E_{sq} = \tilde{s}_j^T E$ are sampled at points r_l and q_m, in images I^R_k and I^L_{k+1}, respectively. Again sample point selection can be random along the epipolar lines or based on match quality. In our implementation we compute local maxima of match quality over the epipolar lines.

At this point we create tables that associate to each triplet (s_j, r_l, q_m) a disparity value d_{jl} and a scale value α_{jlm}, determined by either (9) or (11). Given this information the value of p becomes uniquely determined by (7) and is stored as p_{jlm}. The likelihood of this triplet is then computed by:

$$\lambda_{jlm} = \rho_{s_j}(r_l)\rho_{s_j}(q_m)\rho_{s_j}(p_{jlm})$$

Finally, all computed α_{jlm} values and associated weights λ_{jlm} are fit to a Gaussian Mixture Model using the Expectation Maximization algorithm. The final estimate of α is computed as the mean of the highest likelihood component. In our experiments we use three components in the mixture.

D. Dealing with calibration errors

A common source of errors in a stereo setup is 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 neighborhood of the point is analyzed and not only the exact coordinate of the sample point:

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

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

E. EKF Linear and Angular velocities estimation

Having determined the translation scale factor(α), and the results of the rotation (R) and translation ($\hat{t}$), is possible to compute the linear and angular velocities between T_k and T_{k+1}. The instantaneous linear velocity is given by:

$$V = \frac{\alpha\hat{t}}{\Delta T} \tag{12}$$

where

$$\Delta T = T_{k+1} - T_k \tag{13}$$

Likewise, the angular velocity is computed by:

$$\Omega = \frac{r}{\Delta T} \tag{14}$$

where r contains the incremental roll, pitch and yaw angles computed from R.

In order to achieve a more robust estimation and, therefore, disregard erroneous instantaneous measurements, an Extended Kalman filter, whose filter dynamics follows a constant velocity model was used to integrate the velocity estimates.

$$\begin{bmatrix} V_{k+1} \\ \Omega_{k+1} \\ t_{k+1} \\ R_{k+1} \end{bmatrix} = \begin{bmatrix} V_k + \Delta T a \\ \Omega_k + \Delta T n \\ t_k + R_k V_k \Delta T \\ R_k R_I \end{bmatrix} \tag{15}$$

where a and n are the linear and angular accelerations respectively, taken as independent white noise sequences. As for R_I, it is parametrized as an incremental rotation using Rodrigues formula:

$$R_I = I + \frac{\theta}{||\theta||}\times sin(||\theta||) + \left(\Delta T^2 \Omega_k \Omega_k^T - I\right)\times(1-cos(||\theta||) \tag{16}$$

where θ is given by:

$$\theta = \Omega_k \Delta T \tag{17}$$

Only the linear and angular velocities (V,Ω) are observed by the Extended Kalman Filter, thus the EKF observation model is given by:

$$\begin{bmatrix} V \\ \Omega \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{bmatrix}\begin{bmatrix} V_k \\ \Omega_k \\ t_k \\ R_k \end{bmatrix} + \eta \tag{18}$$

where η is the observation noise taken as a zero mean independent process.

V. Results

A. Experimental Setup

In order to evaluate the PSET results, we utilized one of the sequences of the Karshule dataset, and compared performance against LIBVISO [3], and with our previous implementation of the 6DP [1]. We also compared our performance against Inertial Measurement Unit (RTK-GPS information) using part of one of Kitt *et al* [3] Karlsruhe dataset sequences.

1) Computational requirements: The code used to compute the PSET transform was written in MATLAB as a prove of concept, without using any kind of optimization. The experiments conducted to compute the PSET transform, were performed using an Intel I5 Dual Core 3.2 GHz. The dataset images have resolution of 1344 × 391, which consumes a considerable amount of computational and memory resources making unfeasible the computation of all image points using standard CPU hardware. The PSET transform results were obtained using only 1000 points to estimate the motion. It computes at around 20 sec per image pair. Most of time is consumed in the first stage of the implementation, with the dense probabilistic correspondences and the motion up to a scale factor estimates. Even so, the approach is feasible and can be implemented in real-time for use on mobile robotics applications. The main option is to develop a GPGPU version of the PSET transform implementation since the method copes with multiple hypothesis of correspondences, as well as generated motion hypothesis, making it suitable to be implemented into parallel hardware.

B. Stereo Egomotion results

In this section stereo egomotion estimation results are presented. We compared the PSET estimation results with other state-of-the-art estimation results namely 6DP [1] and LIBVISO [3].

In Fig.5, PSET and LIBVISO estimation of the angular velocities are presented together with IMU ground-truth information. The 6DP results are not presented in the linear and angular velocities figures, due to the fact that for angular velocities case PSET and 6DP use the same method of computation and thus obtain identical results. One can observe that PSET has a much closer performance to IMU than LIBVISO. This is stated in Table 1, where the RMS error for the LIBVISO method is about twice the error of the PSET/6DP method.

In Fig.6, one can observe the behavior of both methods in the linear velocity estimation case. Both LIBVISO and PSET present similar results for the linear velocity estimation case, but PSET has about 50 % less overall RMS error, as can be checked in Table 1. The results confirm that estimating the translation scale using probabilistic approaches produces better results than using deterministic correspondences, as displayed in table 1.

VI. Conclusions and Future Works

The PSET methodology described in this paper has proven to be an accurate method of computing stereo egomotion. The proposed approach is very interesting 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 full dense probabilistic method to compute stereo egomotion. The results demonstrate that PSET is more accurate then other SOA 3D egomotion estimation methods, improving the overall accuracy in about 50 % in angular velocity estimation then LIBVISO and 50 % better accuracy performance in linear velocity, over both LIBVISO and 6DP previous methods.

In future work the main task will be to develop a GPGPU implementation of the PSET transform and test it on difficult and image structure repetitive scenarios where common feature based approaches are more prone to failures. A fused implementation with inertial measurements will also be developed, in order to use the inertial information as motion prior reducing the number of motion hypotheses needed to find the most probable motion.

Acknowledgment

This work is financed by project SEAGULL (SI IDT 34063), funded by the Portuguese government through FEDER funds, and also by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701, and under grant SFRH / BD / 47468 / 2008

References

[1] H. Silva, A. Bernardino, and E. Silva, "Combining sparse and dense methods for 6d visual odometry," *13th IEEE International Conference on Autonomous Robot Systems and Competitions, Lisbon Portugal*, April 2013.

[2] ——, "6d visual odometry with dense probabilistic egomotion estimation," *8th International Conference on Computer Vision Theory and Applications, Barcelona Spain*, February 2013.

[3] G. A. Kitt, B. and H. Lategahn, "Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme," in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2010, pp. 486–492.

[4] A. Comport, E. Malis, and P. Rives, "Real-time quadrifocal visual odometry," *The International Journal of Robotics Research*, vol. 29, no. 2-3, pp. 245–266, Jan. 2010.

[5] C. Olson, L. Matthies, M. Schoppers, and M. Maimone, "Rover navigation using stereo ego-motion," *Robotics and Autonomous Systems*, vol. 43, pp. 215–229, Feb. 2003.

[6] M. Maimone, L. Matthies, and Y. Cheng, "Visual Odometry on the Mars Exploration Rovers," in *IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2005, pp. 903–910.

[7] F. F. Scaramuzza, D., "Visual odometry tutorial," *Robotics Automation Magazine, IEEE*, vol. 18, no. 4, pp. 80 –92, dec 2011.

[8] H. Alismail, B. Browning, and M. B. Dias, "Evaluating pose estimation methods for stereo visual odometry on robots," in *In proceedings of the 11th International Conference on Intelligent Autonomous Systems (IAS-11)*, 2010.

[9] C. Goodall, "Procrustes Methods in the Statistical Analysis of Shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53, no. 2, pp. 285–339, 1991.

[10] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Third International Conference on 3D Digital Imaging and Modeling (3DIM)*, Jun. 2001.

[11] A. Milella and R. Siegwart, "Stereo-based ego-motion estimation using pixel tracking and iterative closest point," in *in IEEE International Conference on Computer Vision Systems*, 2006, p. 21.

[12] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry for ground vehicle applications," *Journal of Field Robotics*, vol. 23, no. 1, pp. 3–20, 2006.

[13] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2008*. Ieee, sep 2008, pp. 3946–3952.

[14] F. Moreno, J. Blanco, and J. González, "An efficient closed-form solution to probabilistic 6D visual odometry for a stereo camera," in *Proceedings of the 9th International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer-Verlag, 2007, pp. 932–942.

[15] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," 1981, pp. 674–679.

[17] N. Kai and F. Dellaert, "Stereo tracking and three-point/one-point algorithms - a robust approach," in *Visual Odometry, In Intl. Conf. on Image Processing (ICIP*, 2006, pp. 2777–2780.

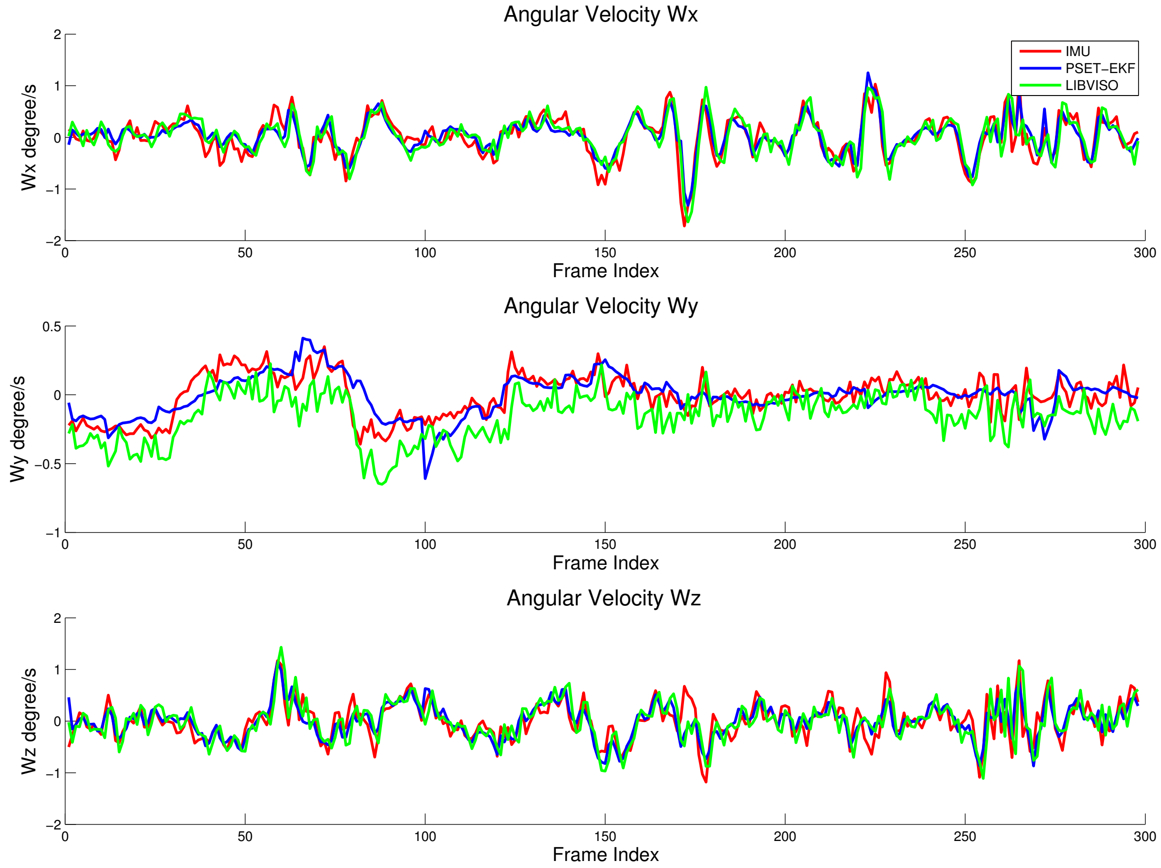

Fig. 5: Results for the angular velocities estimation of 300 frames: ground truth(GPS-IMU information), filtered PSET measurements (PST-EKF) and 6D Visual Odometry Library (LIBVISO). Even though all exhibit similar behaviors the filtered implementation PSET-EKF is the one which is closer to GT(GPS-IMU)(see also table 1).

TABLE I: Comparison of the standard mean squared error between IMU and stereo egomotion estimation methods(LIBVISO, 6DP, and PSET).

	V_x	V_y	V_z	Ω_x	Ω_y	Ω_z	$\|\|V\|\|$	$\|\|\Omega\|\|$
LIBVISO	0.0674	0.7353	0.3186	0.0127	0.0059	0.0117	1.1213	0.0303
6DP	0.0884	0.0748	0.7789	0.0049	0.0021	0.0056	0.9421	0.0126
PSET	0.0700	0.0703	0.3686	0.0034	0.0019	0.0055	0.5089	0.0108

[18] B. C. Fischler, M.A., “Random sample consensus a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[19] K. Ni, F. Dellaert, and M. Kaess, “Flow separation for fast and robust stereo odometry,” in *IEEE International Conference on Robotics and Automation, ICRA 2009*, vol. 1, 2009, pp. 3539–3544.

[20] S. Obdrzalek and J. Matas, “A voting strategy for visual ego-motion from stereo,” in *2010 IEEE Intelligent Vehicles Symposium*, pp. 382–387.

[21] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[22] J. Rehder, K. Gupta, S. T. Nuske, and S. Singh, “Global pose estimation with limited gps and long range visual odometry,” in *IEEE Conference on Robotics and Automation*, May 2012.

[23] L. Kneip, M. Chli, and R. Siegwart, “Robust real-time visual odometry with a single camera and an imu,” in *Proc. of the British Machine Vision Conference (BMVC)*, 2011.

[24] T. Kazik, L. Kneip, J. Nikolic, M. Pollefeys, and R. Siegwart, “Real-time 6d stereo visual odometry with non-overlapping fields of view,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, june 2012, pp. 1529 –1536.

[25] S. Istvan, C. Golban, and S. Nedevschi, “Fast vision based ego-motion estimation from stereo sequences - a gpu approach,” in *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, 2011, pp. 538–543.

[26] J. Domke and Y. Aloimonos, “A Probabilistic Notion of Correspondence and the Epipolar Constraint,” in *Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT’06)*. IEEE, Jun. 2006, pp. 41–48.

[27] J. Huang, T. Zhu, X. Pan, L. Qin, X. Peng, C. Xiong, and J. Fang, “A high-efficiency digital image correlation method based on a fast recursive scheme,” *Measurement Science and Technology*, vol. 21, no. 3, 2011.

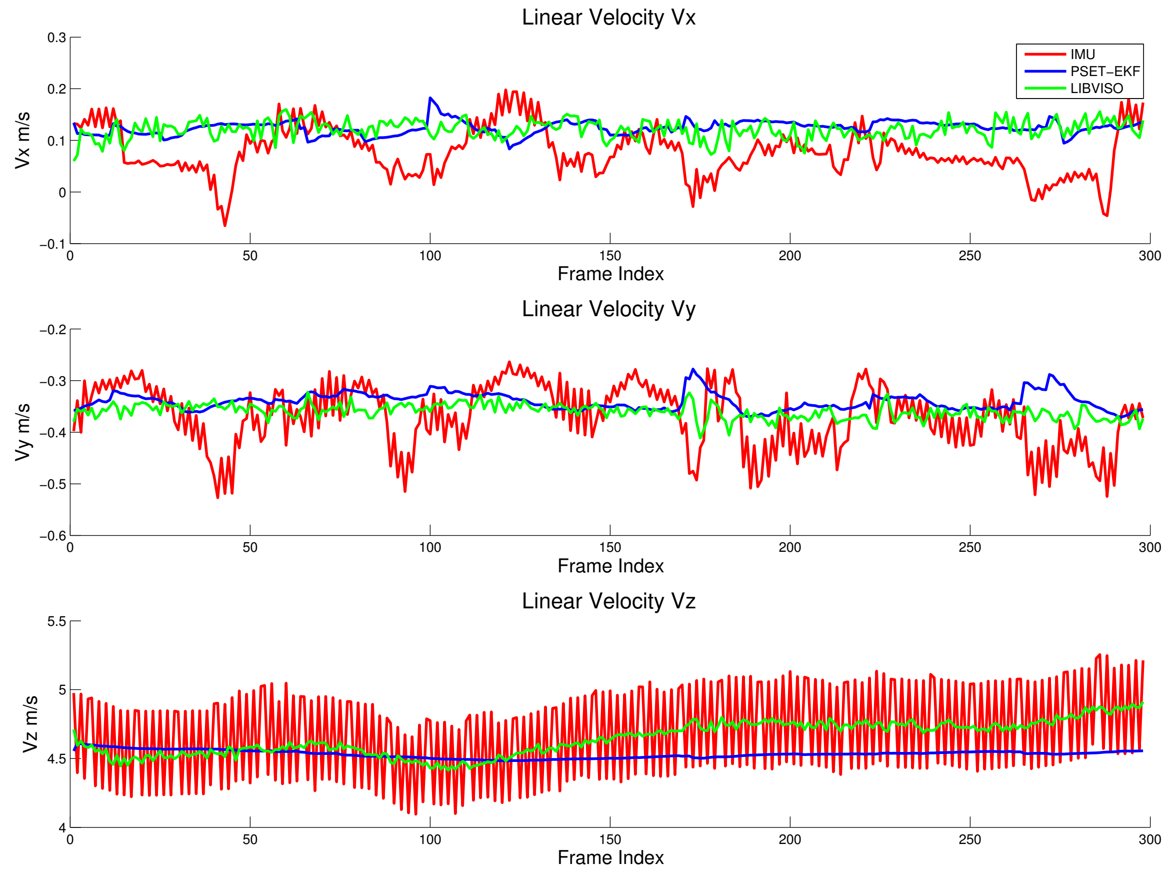

Fig. 6: Estimated linear velocities of 300 frames estimation. The PSET transform exhibits a better performance in V_y compared to LIBVISO, and the opposite occurs in V_z estimation (see Table 1). However in overall linear velocities estimation the PSET is about 50 % better, see Table 1

[28] J. A. Nelder and R. Mead, “A simplex method for function minimization,” *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[29] C. Harris and M. Stephens, “A combined corner and edge detection,” in *Proceedings of The Fourth Alvey Vision Conference*, 1988, pp. 147–151.