Noname manuscript No.
(will be inserted by the editor)

Probabilistic Egomotion for Stereo Visual Odometry

H. Silva · A. Bernardino · E. Silva

Received: date / Accepted: date

Abstract We present a novel approach of probabilistic egomotion methods for Stereo Visual Odometry using vehicles equipped with calibrated stereo cameras. We combine a dense probabilistic 5D egomotion estimation method with a sparse keypoint based stereo approach to provide high quality estimates of vehicle's angular and linear velocities. To validate our approach, we perform two sets of experiments with a well known benchmarking dataset. First, we assess the quality of the raw velocity estimates in comparison to classical pose estimation algorithms. Second, we cascade our method's instantaneous velocity estimates with an Extended Kalman Filter and compare its performance results with a well known open source stereo Visual Odometry library. The presented results compare favorably with state-of-the-art approaches, mainly in the estimation of the angular velocities, where significant improvements are achieved.

1 Introduction

Visual Navigation systems [2], have been subject of important developments by the robotics research community in the last decade. The use of low-cost visual sensors (cameras) together with Inertial Measurement Units (IMU) are becoming ubiquitous on today's modern mobile robots and pushing research on high-performance algorithms for robot navigation.

The use of vision based methods in navigation systems is justified by their ability to ground perception to static features in the environment and measure the robot relative displacement with respect to those features. Therefore, vision based methods are, in principle, less prone to bias and drifts common in other rather common navigation sensory modalities like IMU's and wheel odometers.

This paper is an extension of the work conducted by Silva *et al.* [30], with the addition of a novel parametrization of the mixed sparse and dense egomotion method (here on denoted as 6DP), as well as extended results comparison with other well known state-of-the-art methods. In [29] VO was defined as the process of estimating a vehicle's egomotion by using vision cameras. Cameras work as linear and angular velocity sensors but, because they rely on the observation of fixed points in the environment, they typically provide measurements with less drift than IMU's and wheel odometers. Ultimately, the linear and angular velocities obtained from the egomotion estimation process are integrated along time to provide the relative pose of the robot with respect to some inertial frame. In this paper we focus on the visual egomotion estimation process, since it is the most critical component of a VO system.

Most of the research on VO employ sparse feature based methods. These methods have the advantage of being fast, since only a subset of image points is processed, but depend critically on the features to track between adjacent time frames and are often sensitive to noise and outliers. On the contrary, dense (pixel based) methods combined with probabilistic approaches have demonstrated higher robustness to those source of errors. Domke and Aloimonos [4] proposed a dense probabilistic egomotion estimation method based upon epipolar geometry for describing the motion of a cam-

H.Silva · E. Silva
INESC TEC (formerly INESC Porto)
Instituto Superior de Engenharia do Porto
Portugal
ORCID: http://orcid.org/0000-0002-8133-7216
ORCID: http://orcid.org/0000-0001-7166-3459
E-mail: hugo.m.silva, eduardo silva@inescporto.pt

A. Bernardino
Institute for Systems and Robotics
Instituto Superior Técnico (IST) Lisbon
Portugal
ORCID: http://orcid.org/0000-0003-3991-1269
E-mail: alex@isr.utl.pt

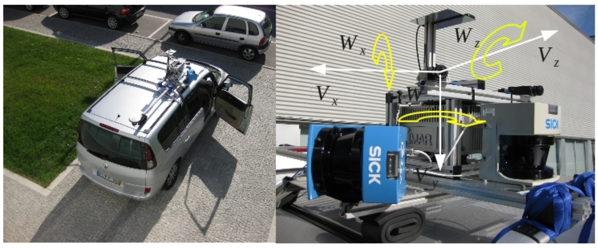

Fig. 1 Example of the INESC TEC camera calibrated setup for a vehicle-like robot, with the use of stereo vision cameras for providing egomotion estimates

era. The method does not commit feature matches between two images on adjacent time frames, but instead computes a probability distribution over all possible correspondences. By exploiting a larger amount of data, a better performance is achieved under noisy measurements. However, Domke method is more computationally expensive than standard feature based methods and only computes the direction of the linear motion, but not the translation scale factor (i.e., the amplitude of the linear motion).

To overcome such limitations, we use a dense probabilistic method such as the one developed by Domke and Aloimonos [4] but with three important contributions. First we add a sparse feature based method that provides stereo vision information needed to compute the translation scale factor. Second we implement a fast correspondence method based on recursive Zero Normalized Cross Correlation (ZNCC) scheme for computational efficiency. Third we integrate the obtained velocity estimates in an Extended Kalman Filter able to reduce the noise present in the instantaneous measurements. Our proposed approach to stereo egomotion estimation (6DP) combines a deterministic sparse feature based method for obtaining depth estimation, with a dense probabilistic egomotion approach that allows to recover camera rotation (R) and translation ($\tilde{t}$) up to a scale factor (α). For recovering the missing translation scale factor (α) we use a Procrustes Absolute Orientation method, that takes registered 3D point information from two adjacent time frames.

In this paper we compare our mixed deterministic and probabilistic egomotion estimation approach (6DP) against two well known state-of-the art egomotion estimation methods. First we evaluate 6DP linear and angular velocities raw estimates (without any type of filtering) against a 5-point algorithm. The use of a dense probabilistic approach allows to obtain better estimates of the rotation and translation up to scale factor when compared to the 5-point implementation, but exhibits an unfavorably performance in the translation scale factor (α) estimation. Afterwards we have implemented a filtering approach on top of the 6DP estimator and compared it with LIBVISO Visual Odometry Library [13], using a standard dataset from this library. The dataset also provides ground-truth information from the fusion of IMU and GPS measurements. Results show that our method presents significant improvements in the estimation of angular velocities and a similar performance for linear velocities.

This paper is organized as follows: In section 2 related work regarding stereo VO is presented. The 6DP algorithm implementation is detailed in section 3. Finally section 4 and section 5 contain the experimental results and conclusions with final remarks.

2 Related Work

Stereo VO consists of performing egomotion estimation from the sequence of images acquired from a stereo camera rig rigidly attached to the vehicle or robot. One of the advantages of performing VO estimation with a stereo camera configuration is the ability to recover translation motion scale. Classical stereo VO algorithms estimate the 3D position of observed image point features by using triangulation between the left and right images. Then, relative camera motion can be calculated through the alignment of 3D feature's position between consecutive image frames.

Most of the work on stereo visual odometry methods was driven by Matthies *et al.*[17][18] outstanding work on the famous Mars Rover Project. Their system was able to determine all 6-DOF of the rover (x,y,z,roll,pitch,yaw) by tracking the motion of 2D image keypoints between stereo image pairs, as well as their 3D world coordinates. Afterwards, a maximum likelihood estimation method was used to compute motion between consecutive image frames. Their method exploited robust methods for outlier rejection such as RANSAC [5]. The Stereo VO work performed on Mars Rover Project was somewhat inspired by Olson *et al.* [26]. The method was developed as a replacement for wheel odometry dead reckoning methods that were not able to correctly estimate robot motion over long distances. In order to avoid large drift in robot position over time, Olson's method combined stereo egomotion estimation with absolute orientation sensor information.

Among the different approaches to compute stereo VO, two main categories have emerged in the literature, either based on their feature detection scheme or by the way motion estimation is performed. Usually, motion estimation is computed using 3D Absolute Orientation (AO) methods or Perspective-n-Point (PnP) methods. Alismail *et al.*[1] conducted a study on evaluating both AO and PnP methods for achieving robot pose estimation using only stereo visual odometry, and concluded that PnP methods are more accurate than AO methods. The AO methods consist on 3D triangulated points estimation for every stereo pair. Then motion estimation is solved by using point alignment algorithms like the Procrustes method [6] or Iterative-Closest- Point (ICP) method [28], such as the one used by Milella and Siegwart [19] for estimating motion of an all-terrain rover. Nister *et al.*[24], were one of the first to develop a PnP algorithm (3D-2D camera pose estimation), that could be computed in real-time with an outlier rejection scheme. The authors argue that minimizing the re-projection error would benefit stereo VO method accuracy. Nister *et al.*[23] also developed a Visual Odometry system, based on a 5-point algorithm, that became the standard algorithm for comparison of Visual Odometry techniques. This algorithm can be used either in stereo or monocular vision approaches and consists on the use of several visual processing techniques, namely: feature detection and matching, tracking, stereo triangulation and RANSAC for pose estimation with iterative refinement.

Most of stereo VO methods differ on the way stereo information is acquired and computed: sparse or dense approaches. One of the most relevant dense stereo VO applications was developed by Howard [9] for ground vehicle applications. The method does not assume prior knowledge over camera motion and so can handle very large image translations. However, due to the absence of feature detectors invariant to rotation and scaling, only works on low-speed applications and with high frame-rate, since large motions around the optical axis result in poor performance. In [20] 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 (KLT) [16]. Some other authors like Ni *et al.*[11], 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 framework. Later on, the same authors [22], 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 automotive industry. Obdrzalek *et al.*[25] developed a voting scheme strategy for egomotion estimation, where 6-DOF problem was divided into a four dimensions problems and then decomposed in two sub-problems for rotation and translation estimation. Another influential work, is the one developed by Kitt *et al.* [13]. 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 [8]. It uses rectified stereo image sequences and produces an output 6D vector with estimated linear and angular velocities. Comport *et al.* [3] also develop a stereo VO method based on a different geometry estimation solution, 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.*[27] 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.*[14] 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.*[12] 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.

3 A mixed approach to stereo visual odometry: combining sparse and dense methods

Our method, denoted 6DP, combines sparse feature based methods with dense probabilistic methods [30]. While feature based methods are less computational expensive and are used in real-time applications, dense correlation methods tend to be computational intensive and used in more complex applications. However, when combined with probabilistic approaches, such methods are usually more robust and tend to produce more precise results. Therefore we developed a solution that tries to exploit the advantages of both methods.

Our 6DP method, as schematically illustrated in Fig 2, can be roughly divided into three main steps:

- **Keypoint Detection**
  The first step is to detect sparse features in an image stereo pair. To obtain such features a feature detector such as the well known Harris corner [7] or a SIFT detector [15] is used. By performing stereo triangulation 3D point clouds at time T_k and T_{k+1} are computed but correspondences between points are not resolved at this stage.
- **Correspondence and Egomotion estimation**
  In order to be able to estimate egomotion, first there is the need to compute correspondence information between images I_{Tk} and I_{Tk+1}, where Tk and $Tk+1$ are consecutive time instants. For egomotion estimation a variant of the dense probabilistic egomotion estimation method of [4] is used. By doing so, we establish a probabilistic correspondence between the left images at con-

secutive time steps, I^L_{Tk} and I^L_{Tk+1}, and estimate camera rotation (R) and translation ($\tilde{\mathbf{t}}$) up to a scale factor (α).

- **Scale Estimation**
  The missing translation scale factor (α), is obtained by using the sparse features in each of stereo pairs at consecutive time instants and, by triangulation, compute two 3D registered point sets in consecutive frames. Afterwards an AO method like the Procrustes method [6] is used to obtain the best alignment between the two sets of points and determine the value of the translation scale factor (α).

3.1 Probabilistic Correspondence

The key to the proposed method relies on a robust probabilistic computation of the epipolar geometry relating the camera's relative pose on consecutive time steps. This will speed-up and simplify the search for 3D matches on the subsequent phases of the algorithm. Usual methods for motion estimation consider a match function M that associates coordinates of points $\mathbf{x} = (x,y)$ in image 1 (I^L_{Tk}) to points $\mathbf{x}' = (x',y')$ in image 2 (I^L_{Tk+1}) :

$$M(\mathbf{x}) = \mathbf{x}' \tag{1}$$

Instead, the probabilistic correspondence method defines a probability distribution over the points in image 2 for all points in image 1:

$$\rho_{\mathbf{x}}(\mathbf{x}') = \rho(\mathbf{x}'|\mathbf{x}) \tag{2}$$

Thus, all points $\mathbf{x}'$ in image 2 are candidates for matching with point $\mathbf{x}$ in image 1 with a likelihood proportional to $\rho_{\mathbf{x}}(\mathbf{x}')$. One can consider $\rho_{\mathbf{x}}$ as images (one per each pixel in image 1) whose value in $\mathbf{x}'$ is proportional to the likelihood of $\mathbf{x}'$ matching with $\mathbf{x}$. In Fig.4, we can observe the likelihood results of a point $\mathbf{x}$ in image I^L_{Tk} with all matching candidates $\mathbf{x}'$ in I^L_{Tk+1}. For the sake of computational cost, likelihoods are not computed for the whole range in image 2 but just on windows around $\mathbf{x}$, or suitable predictions based on prior information (see Fig. 3).

In [4] the probabilistic correspondence images were computed via the normalized product, over a filter bank of Gabor filters with different orientation and scales, of the exponential of the negative differences between the angle of the Gabor filter responses in $\mathbf{x}$ and $\mathbf{x}'$. The motivation for using a Gabor filter bank is its robustness to changes in the brightness and contrast of the image. However, it demands a significant computational effort, thus we propose to perform the computations with the well known Zero Mean Normalized Cross Correlation function (ZNCC):

Algorithm 1: 6DP Method

```
Input: 2 stereo Image pairs $(I^L_{Tk}, I^R_{Tk})$ and $(I^L_{Tk+1}, I^R_{Tk+1})$,
Output: (Velocities) V, $\Omega$
Step 1. Using a Feature Detector obtain corresponding inliers
    This sparse feature detector is only used to obtain the translation scale estimation. We conducted experiments using both Harris corners and Scale Invariant Features (SIFT)
    $F_k \leftarrow Feature\ Detector(I^L_{Tk}, I^R_{Tk})$
Step 2. For a new stereo image pair we keep the feature detection points that match in the stereo image pair $I^L_{Tk+1}$ and $I^R_{Tk+1}$.
    Therefore a correlation procedure is conducted over the epipolar line to match corresponding key points in both images. The points $P2_{k+1}$ which are the matching key points of the stereo pair (inliers) are kept.
    $F_{k+1} \leftarrow Feature\ Detector(I^L_{Tk+1}, I^R_{Tk+1})$
    $P2_{k+1} \leftarrow Epipolar\ Correlation(F_{k+1})$
Step 3. Initiate the dense part of the 6DP method and estimate the motion.
Step 4. Compute the probabilistic correspondences between images $I^L_{Tk}$ and $I^L_{Tk+1}$ .
    $\rho_{\mathbf{x}}(\mathbf{x}') \leftarrow ZNCC(\mathbf{x},\mathbf{x}'; I^L_k, I^L_{k+1}); \mathbf{x}' \in W(I^L_{k+1})$; W (pre-defined window size)
Step 5. Compute using probabilistic approaches the Motion $(E)$
    The Essential Matrix $(E)$ is obtained from the from the probabilistic correspondences over the 5-dimensional space of essential matrices $E_i$. For a single point $x$ in image $I^L_{Tk}$, 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.
    $\rho(E_i) \propto \prod_x \rho(E_i|x)$
Step 6. Perform stereo triangulation to obtain 3D points in both $I_{Tk}$ and $I_{Tk+1}$ time instants, only the matches of $P2_k$, $P2_{k+1}$ that are correlated by the epipolar line given by the obtained motion $(E)$ are utilized
    $P3_{k,k+1} \leftarrow Stereo\ Triangulation\ (P2_k, P2_{k+1})$
Step 7. Perform a robust estimation step (using RANSAC) to disregard 3D points that are outliers. For avoiding biased samples a bucketting estimation procedure is used
    $X_{k+1} = RX_k + T\ Ransac\ (P3_{k,k+1})$
Step 8. Perform Translation scale estimation using an Absolute Orientation method (Procrustes) The Procrustes method allows to recover rigid body motion between frames through the use of 3D point matches. This transformation can be represented as:
    $\mathbf{Y_i} = R'\mathbf{X_i} + \mathbf{t}'$
    $\alpha \leftarrow Procrustes(P3_{k,k+1})$
Step 9. Estimate Linear and Angular Velocities
    $\mathbf{V} \leftarrow \frac{\alpha\tilde{\mathbf{t}}}{\Delta T}$
    $\mathbf{r} \leftarrow R(r_x, r_y, r_z)$
    $\Omega \leftarrow \frac{\mathbf{r}}{\Delta T}$
Step 10. Extended Kalman Filtering
Repeat from Step 2
```

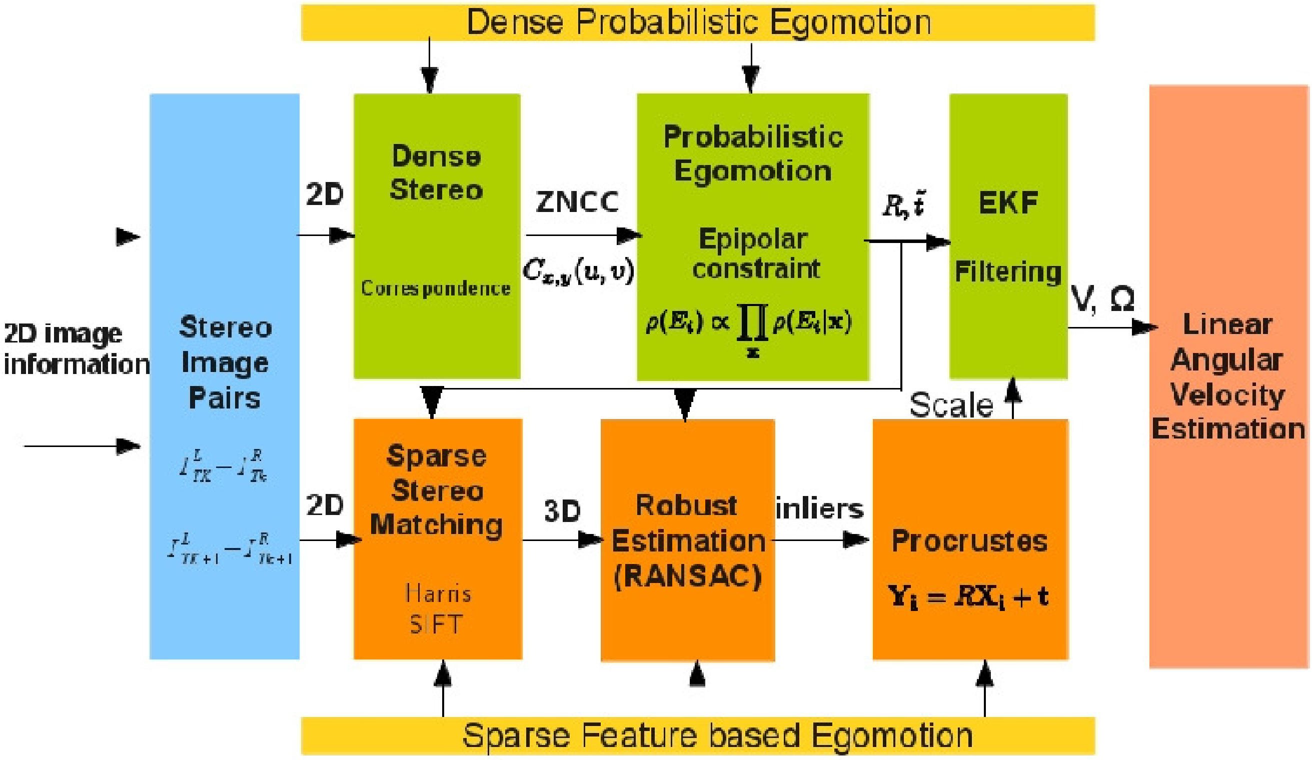

Fig. 2 6DP architecture

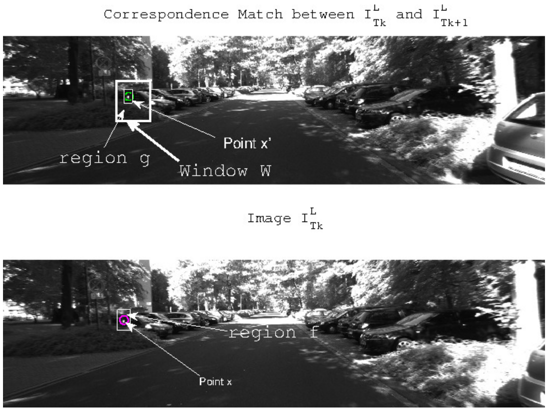

Fig. 3 Image feature point correspondence for ZNCC matching, with searching regions(f,g) defined with window size W between points $\mathbf{x}$ and point $\mathbf{x}'$ represented in red and green respectively

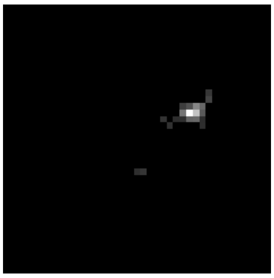

Fig. 4 Likelihood of a point $\mathbf{x}$ in image I^L_{Tk} with all matching candidates $\mathbf{x}'$ in I^L_{Tk+1}, for the case of Fig. 3. Points with high likelihood are represented in lighter colour

$$C_{x,y}(u,v) = \frac{\sum\limits_{x,y \in W} (f(x,y) - \bar{f})(g(x+u,y+v) - \bar{g})}{\sqrt{\sum\limits_{x,y \in W} (f(x,y) - \bar{f})^2} \sqrt{\sum\limits_{x,y \in W} (g(x+u,y+v) - \bar{g})^2}} \tag{3}$$

The ZNCC method allows to compute the correlation factor $C_{x,y}(u,v)$ between regions of two images f and g by using a correlation window around pixel $\mathbf{x} = (x,y)$ in image f and pixel **x'** = **x**+(u,v) in image g, being the correlation window size $N_W = 20$. Having $\bar{f}$ and $\bar{g}$ as the mean values of the images in the regions delimited by the window size. This correlation factor is then transformed into a likelihood match between f and g:

$$\rho_{\mathbf{x}}(\mathbf{x}') = \frac{C_{x,y}(u,v)}{2} + 0.5 \tag{4}$$

The ZNCC function is known to be robust to brightness and contrast changes and recent efficient recursive schemes developed by Huang *et al.* [10] render it suitable to real-time implementations. That method is faster to compute and yields the same quality as the method of Domke. For more detail about the recursive scheme, see the Appendix.

3.2 Probabilistic Egomotion Estimation

From two images of the same camera, one can recover its motion up to the translation scale factor. Given the camera motion, image motion can be represented by the epipolar constraint which, in homogeneous normalized coordinates, can be written as:

$$(\tilde{\mathbf{x}}')^T E \tilde{\mathbf{x}} = 0 \tag{5}$$

where E is the so called Essential Matrix [8], a $3X3$ matrix with rank 2 and 5 degrees-of-freedom and $\tilde{\mathbf{x}}, \tilde{\mathbf{x}}'$ the homogeneous coordinate representations of points $\mathbf{x}$ and $\mathbf{x}'$. Given a point $\tilde{\mathbf{x}}$ in image 1, this expression constraints the points $\tilde{\mathbf{x}}'$ in image 2 to lie on line $E\tilde{\mathbf{x}}$, thus it expresses the loci in image 2 that should be searched for matches of points in image 1. It can be factored by:

$$E = R \left[\tilde{\mathbf{t}}\right]_{\times} \tag{6}$$

where R and $\tilde{\mathbf{t}}$ are, respectively, the rotation and translation direction of the camera between the two frames, with $\tilde{\mathbf{t}}_{\times}$ the skew symmetric representation of $\tilde{\mathbf{t}}$, as defined in the following expression:

$$\tilde{t}_{\times} = \begin{bmatrix} 0 & -\tilde{t}_z & \tilde{t}_y \\ \tilde{t}_z & 0 & -\tilde{t}_x \\ -\tilde{t}_y & \tilde{t}_x & 0 \end{bmatrix} \tag{7}$$

To obtain the Essential matrix from the probabilistic correspondences, [4] 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 $\mathbf{x}$ the likelihood of these hypotheses is evaluated by:

$$\rho(E_i|\mathbf{x}) \propto \max_{(\tilde{\mathbf{x}}')^T E_i \tilde{\mathbf{x}}=0} \rho_{\mathbf{x}}(\mathbf{x}') \tag{8}$$

Intuitively, for a single point $\mathbf{x}$ in image 1, the likelihood of a motion hypothesis is proportional to the likelihood of the best match obtained along the epipolar line generated by the essential matrix. Assuming 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:

$$\rho(E_i) \propto \prod_{\mathbf{x}} \rho(E_i|\mathbf{x}) \tag{9}$$

After the dense correspondence probability distribution has been computed for all points, the method [4] computes a probability distribution over motion hypotheses represented by the epipolar constraint. Finally, having computed all the motion hypotheses, a Nelder-Mead simplex method [21] is used to refine the motion estimate using a pre-defined number of the highest scoring samples E_i. The idea behind this approach already applied in [4] is to robust the final solution in a way that missing the global maximum becomes highly unlikely. The motion sample returned by the optimization method that has the highest probability is considered to be the final solution.

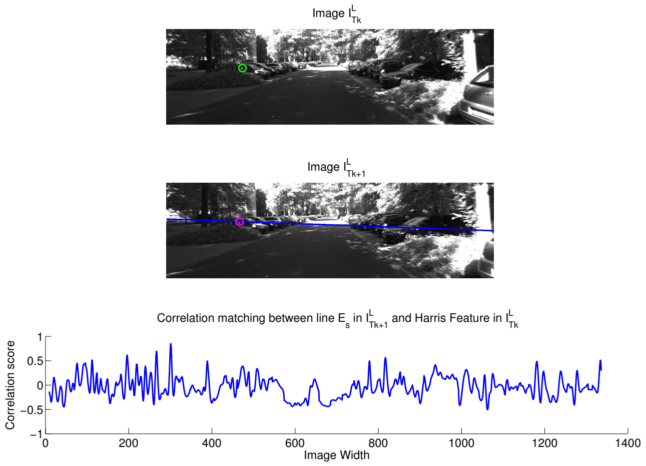

Fig. 5 Image feature point marked in colour green in image I^L_{Tk} lies in the epipolar line (blue) estimated between I_{Tk} to I_{Tk+1}. The point with higher correlation score, marked in red in image I^L_{Tk+1} is chosen as the matching feature point.

3.3 Scale Estimation

By using the previous method, we compute the 5D transformation $(R, \tilde{\mathbf{t}})$ between the camera frames at times T_k and T_{k+1}. However, translation $\tilde{\mathbf{t}}$ component does not contain translation scale information. This type of information, will be calculated by an Absolute Orientation(AO) method like the Procrustes method.

Once the essential matrix between images I^L_{Tk} and I^L_{Tk+1} has been computed by the method described in the previous section, we search along the epipolar lines for matches in I^L_{Tk+1} to the features computed in I^L_{Tk}, as displayed in Fig. 5.

Finally, the matches in I^L_{Tk+1} are propagated to I^R_{Tk+1} by searching along horizontal stereo epipolar lines. From this step we compute 3D point clouds at time T_{k+1} corresponding to the ones obtained for T_k. Points whose matches are unreliable or were not found are discarded from the point clouds.

3.3.1 Procrustes Analysis and Scale Factor Recovery

The Procrustes method allows to recover rigid body motion between frames through the use of 3D point matches. Let the set of 3D keypoints computed by triangulation of the Harris Corners in instant T_k be described by $\{\mathbf{X_i}\}_{\mathbf{T_k}}$. At time T_{k+1} they move to a new position and orientation. Let the new set be described by $\{\mathbf{Y_i}\}_{\mathbf{T_{k+1}}}$. This transformation can be represented as:

$$\mathbf{Y_i} = R'\mathbf{X_i} + \mathbf{t}' \tag{10}$$

In order to estimate the motion $[R', \mathbf{t}']$, a cost function that measures the sum of squared distances between corresponding points is used.

$$c^2 = \sum_i^n \left\| \mathbf{Y_i} - (R'\mathbf{X_i} + \mathbf{t}') \right\|^2 \tag{11}$$

Performing minimization of equation (11) is possible to estimate $[R', \mathbf{t}']$. However these estimates are only used to obtain the missing translation scale factor α, since rotation (R) and translation direction $(\tilde{\mathbf{t}})$ were already obtained by the probabilistic method. Although conceptually simple, some aspects regarding the practical implementation of the Procrustes method must be taken into consideration. Namely, since this method is very sensible to data noise, obtained results tend to vary in the presence of outliers. To overcome this difficulty, RANSAC [5] is used to discard possible outliers within the set of matching points.

3.3.2 Bucketing

For a correct motion scale estimation, it is necessary to have a proper spatial feature distribution through out the image. For instance, if the Procrustes method uses all obtained image feature points without having their image spatial distribution into consideration, the obtained motion estimation $[R, \mathbf{t}]$ between two consecutive images could turn out biased. Given these facts, to avoid having biased samples in the RANSAC phase of the algorithm a bucketing technique [31] is implemented to assure a balanced image feature distribution sample. In Fig. 6 a possible division of the image is displayed. The image region is divided into $L_x \times L_y$ buckets, based on the minimum and maximum image width of the feature points. Afterwards, image feature points are classified as belonging to one of the buckets. In case a bucket does not contain any feature, it will be disregarded. The bucket size must be previously defined: in our case we divided the image into a 8 × 8 buckets. Assuming we have l buckets, the interval between [0...1] is divided into l intervals such that the width of the bucket interval is defined as $n_i / \sum_i n_i$, where n_i is the number of matches assigned to each bucket. Based on this computation the RANSAC sample points are chosen based on point bucket probability. Buckets that contain higher number of points have higher probability, the method allows a faster RANSAC convergence and avoids having biased point samples.

3.4 Linear and Angular Velocity Estimation

To sum up the foregoing, we determine camera motion estimation up to a scale factor using a probabilistic method, and by adding stereo vision combined with Procrustes estimation method, we are able to determine missing motion scale α:

$$\alpha = \frac{||\mathbf{t}'||}{||\tilde{\mathbf{t}}||} \tag{12}$$

Then, the instantaneous linear velocity is given by:

$$V = \frac{\alpha \times \tilde{\mathbf{t}}}{\Delta T} \tag{13}$$

where

$$\Delta T = T_{k+1} - T_k \tag{14}$$

Likewise, the angular velocity is computed by:

$$\Omega = \frac{r}{\Delta T} \tag{15}$$

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

Thus, using motion scale information given by the Procrustes method, we can estimate vehicle linear velocity between time instants T_k and T_{k+1}. The AO orientation method is only used for linear velocity estimation (motion scale). For the angular velocity estimation we use the rotation matrix R calculated by Domke's probabilistic method, that is more accurate than the rotation obtained by the AO method.

In order to achieve a more robust estimation, we also develop an Extended Kalman filter approach. The Extended Kalman filter is used to integrate the velocity estimates and thus obtain vehicle pose algorithm. The EKF filter dynamics follows a constant velocity model given by:

$$\begin{bmatrix} V_{k+1} \\ \Omega_{k+1} \\ t_{k+1} \\ R_{k+1} \end{bmatrix} = \begin{bmatrix} V_k + \Delta T a \\ \Omega_k + \Delta T n \\ t_k + R_k V_k \Delta T \\ R_k R_I \end{bmatrix} \tag{16}$$

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:

$$R_I = I + \frac{\theta}{||\theta||} \times sin(||\theta||) + \left(\Delta T^2 \Omega_k \Omega_k^T - I\right) \times (1 - cos(||\theta||) \tag{17}$$

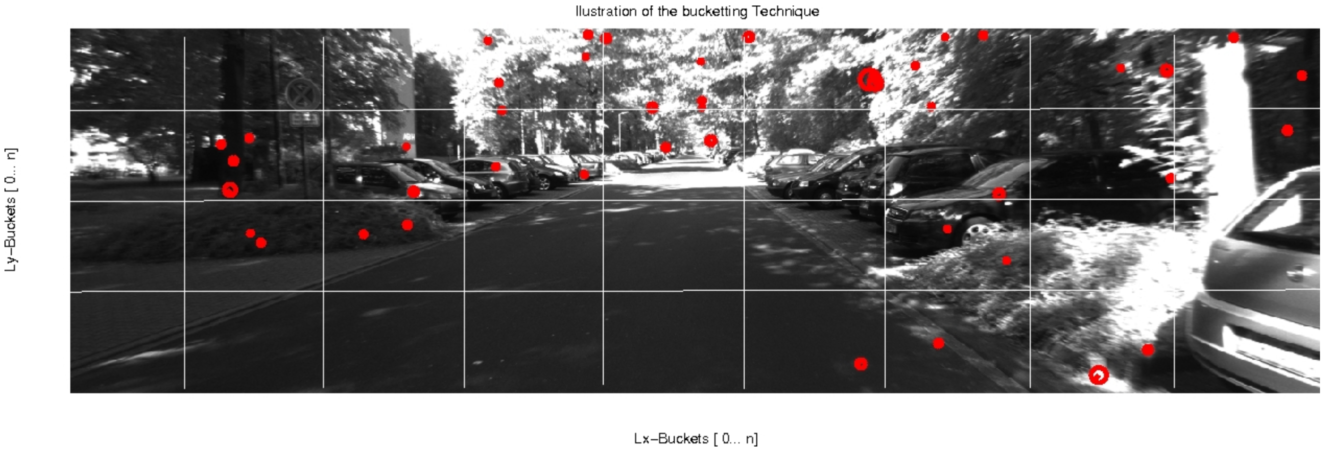

Fig. 6 Feature detection bucketing technique used to avoid biased samples in the RANSAC method stage. The image is divided in buckets where feature points are assigned to and pulled according to the bucket probability.

where θ is given by:

$$\theta = \Omega_k \Delta T \tag{18}$$

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

$$\begin{bmatrix} V \\ \Omega \end{bmatrix} = \begin{bmatrix} I\ 0\ 0\ 0 \\ 0\ I\ 0\ 0 \end{bmatrix} \begin{bmatrix} V_k \\ \Omega_k \\ t_k \\ R_k \end{bmatrix} + \eta \tag{19}$$

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

4 Results

4.1 Computational requirements

The code used to compute 6DP was written in MATLAB as a prove of concept, without using any kind of code optimization. The experiments conducted to compute the 6DP, 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 6DP results were obtained using only 1000 points to estimate the motion. It computes at around 12 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. The recursive ZNCC approach allowed to reduce Domke Gabor Filter processing time by 20 %.

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 method since the method copes with multiple hypothesis of correspondences, as well as generated motion hypothesis, making it suitable to be implemented into parallel hardware.

4.2 6DP-RAW vs 5-point

In this section, one can observe results comparing our approach versus the 5-point RANSAC algorithm. Linear and angular velocities estimation results are presented in the camera reference frame. Results were obtained using only 1000 keypoints in I^L_{Tk}.

In Fig. 7, one can observe the angular velocity estimation between our 6DP method versus IMU and 5-point RANSAC information. We also show the Inertial Navigation System data (IMU/GPS OXTS RT 3003), which is considered as ground-truth information. The displayed results demonstrate a high degree of similarity between performance obtained using 6DP and IMU/GPS information. Results obtained by 6DP were performed without using any type of filtering technique, thus the display of one or two clear outliers. Most importantly, when it comes to angular velocities estimation, the 6DP method performance is better than the performance exhibited by the 5-point RANSAC algorithm.

However, for linear velocities as displayed in Fig. 8, the 5-point RANSAC algorithm implementation performance is smoother than our 6DP approach, especially in Z axis T_z. As shown in Fig. 10, the 5-point algorithm contains more image features when performing Procrustes Absolute Orientation method (after RANSAC) which may also explain the higher robustness in motion scale estimation in Fig. 9, where the 5-point algorithm displays a constant translation scale value.

The results here displayed demonstrate complementary performances, one more suitable for linear motion estimation and the other more suitable for angular motion estimation.

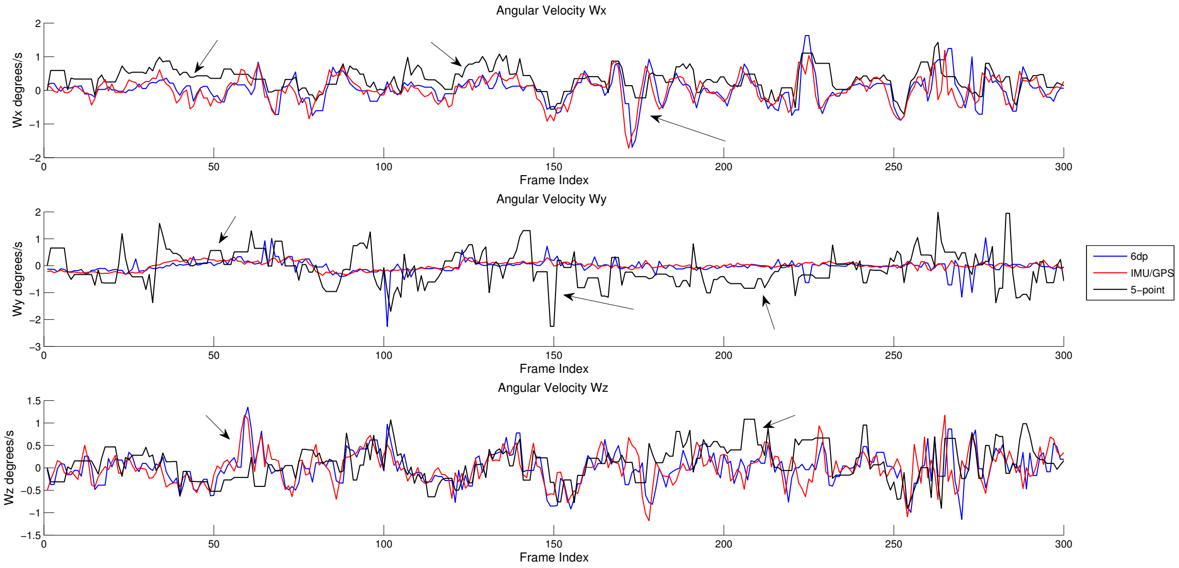

Fig. 7 Comparison of angular velocity estimation results between IMS/GPU(red), raw 6DP measurements (blue) and a native 5-point implementation (black). The obtained 6DP raw measurements are similar to the data estimated by the IMU/GPS, contrary to the 5-point implementation that has some periods of large errors (e.g. the regions indicated with arrows in the plots).

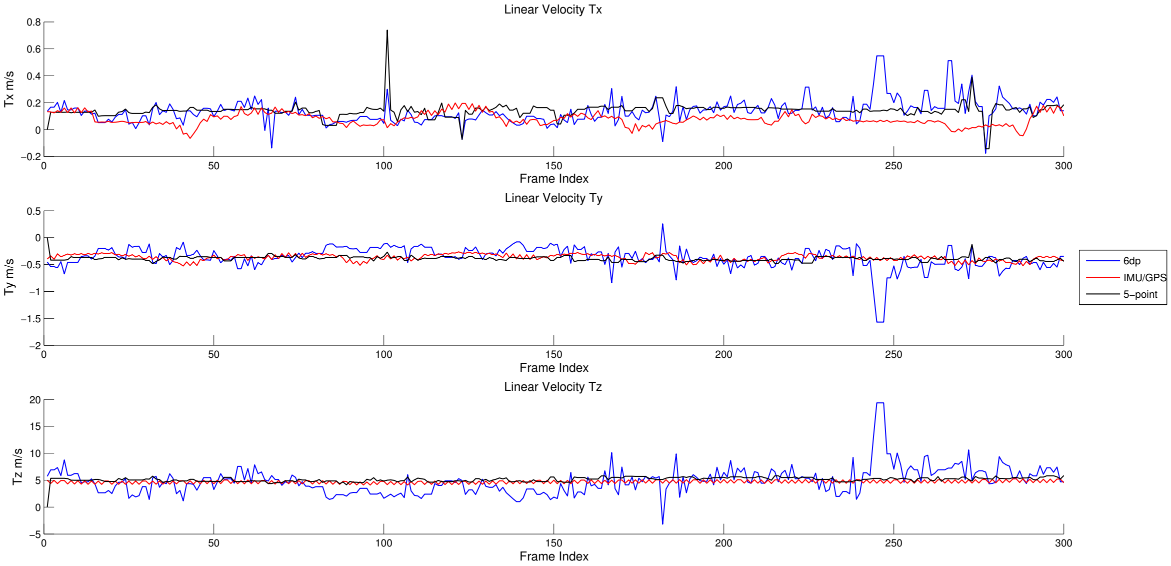

Fig. 8 Comparison of linear velocity estimation results, where the 5-point implementation (black) exhibits a closer match to the IMU/GPS information (red). The 6DP method (blue) displays some highlighted outliers due to the use of the Harris feature detection matching in the sparse method stage.

4.3 6DP-EKF vs LIBVISO

Based on the previous results for translation scale (α) estimation, we modified the 6DP method and replaced the Harris corner feature detector [7] for the more robust and invariant to rotation and scale SIFT detector [15] (see Fig.10) and added a EKF filter to the 6DP method. To illustrate the performance of our method, we compared our system performance against LIBVISO [13], which is a standard library for computing 6 DOF motion. We also compared our performance against INS information of the previous experiment(IMU/GPS information) acting as ground truth info using the same Kitt *et al.*[13] Karlsruhe dataset sequences.

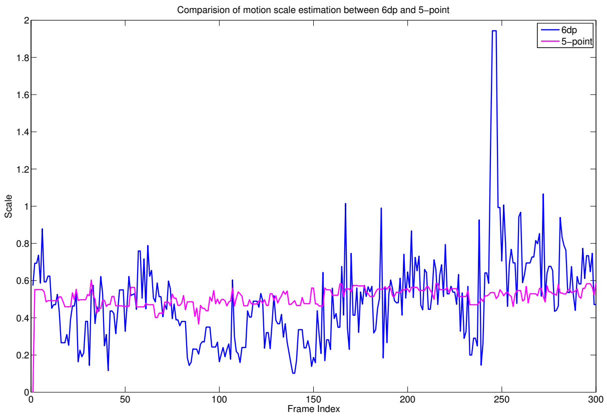

Fig. 9 Translation scale factor comparison between 5-point and 6DP, where the 5-point method exhibits a more constant behavior for the translation scale factor estimation.

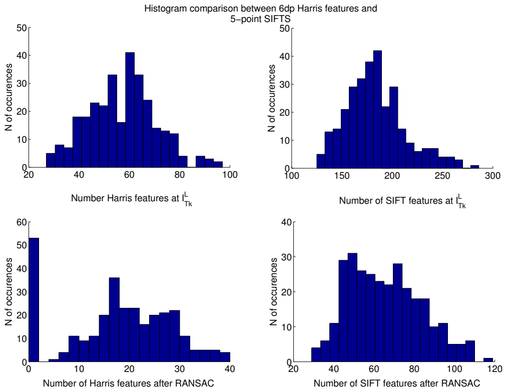

Fig. 10 Number of Features at different steps of 6DP and 5-point. SIFT features display a more robust matching behavior between images, contrary to Harris Corners, most of the SIFTS are not eliminated in the RANSAC stage.

In Fig. 11 one can observe angular velocity estimation from both IMU/GPS and LIBVISO, together with 6DP-RAW and EKF filtered measurements. All vision approaches obtained results consistent with the IMU/GPS, but the 6DP-EKF displays a better performance in what respects the angular velocities. These results are stated in Table 1, where root mean square error between IMU/GPS, LIBVISO and 6DP-EKF estimation are displayed. Both methods display considerable low standard deviation results, but 6DP-EKF shows 50% lower error than LIBVISO for the angular velocities estimation.

Although not as good as for the angular velocities, the 6DP-EKF method also displays a competitive performance in obtaining linear velocity estimates using the sparse feature approach based on SIFT features combined with Procrustes Absolute Orientation method, as displayed in Fig. 12.

5 Conclusions and Future Work

In this work, we developed a novel method of stereo visual odometry using sparse and dense egomotion estimation methods. We utilized dense egomotion estimation methods for estimating the rotation and translation up to scale and then complement the method with the use of a sparse feature approach for recovering the scale factor.

First, we compared the raw estimates of our 6DP algorithm against a native 5-point implementation without any type of filtering. The results obtained proved that 6DP performed better in the angular velocities estimation but compared unfavorably in the linear velocities estimation due to lack of robustness in the translation scale factor(α) estimation. On a second implementation, we replaced the Harris feature detector with the more robust SIFT detector, implemented EKF filtering on top of the raw estimates and tested the proposed algorithm against a state-of-the-art 6D visual Odometry Library such as LIBVISO. The presented results demonstrate that 6DP performs accurately when compared to other techniques for stereo VO estimation, yielding robust motion estimation results, mainly in the angular velocities.

The benefits of using dense probabilistic approaches are thus tested and validated in a real world scenario with practical significance. Even though more computational intensive, dense methods produce more accurate results than feature based methods and are a competitive alternative to stereo egomotion computation.

To overcome increased computational cost one should, in future work, explore their potential implementation in parallel hardware such as a GPU.

Acknowledgements This work is financed by Project "NORTE-07-0124-FEDER-000060" of the North Portugal Regional Operational Programme (ON.2 O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia within project FCT project [PEst-OE/EEI/LA0009/2013]" and under grant SFRH / BD / 47468 / 2008

Appendix

The global objective of the ZNCC method is to compare a reference subset (the correlation window sampled in the reference image) to a corresponding template in another image. The method developed by Huang *et al.*[10] uses a recursive scheme for calculating the numerator of (20) and a global sum-table approach for the denominator, thus saving significant computation time.

In summary, the method has two distinctive parts one for calculating ZNCC numerator and other for the denominator calculation. The ZNCC equation (3) can be described in the

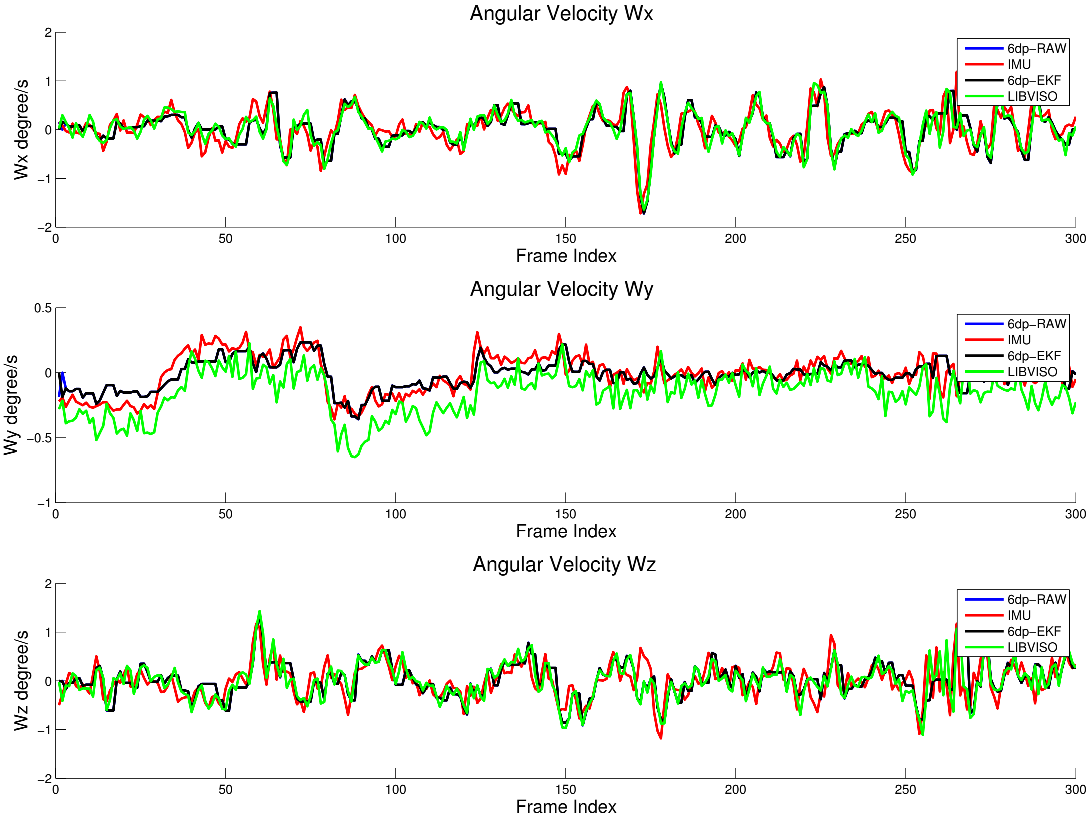

Fig. 11 Results for angular velocities estimation between IMU/GPS information (red), raw 6DP measurements (blue), filtered 6DP measurements 6DP-EKF (black), and 6D Visual Odometry Library LIBVISO (green). Even though all exhibit similar behaviors the filtered implementation 6DP-EKF is the one which is closer to the ”ground truth” IMU/GPS measurements (see also Table 1).

Table 1 Standard Mean Squared Error between IMU and Visual Odometry (LIBVISO and 6DP-EKF). The displayed results show a significant improvement of the 6DP-EKF method performance specially in the angular velocities estimation case.

	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-EKF	0.0884	0.0748	0.7789	0.0049	0.0021	0.0056	0.9421	0.0126

following form.

$$C_{x,y}(u,v) = \frac{P(x,y;u,v) - Q(x,y;u,v)}{\sqrt{F(x,y)}\sqrt{G(x,y;u,v)}} \tag{20}$$

where the numerator term can be calculated using the following equations:

$$P(x,y;u,v) = \sum_{x=x-N_x}^{x+N_x} \sum_{y=y-N_y}^{y+N_y} [f(x,y) \times g(x+u,y+v)]. \tag{21}$$

$$Q(x,y;u,v) = \frac{1}{(2N_x+1)(2N_y+1)} \left[\sum_{x=x-N_x}^{x+N_x} \sum_{y=y-N_y}^{y+N_y} f(x,y) \right] \times \left[\sum_{x=x-N_x+u}^{x+N_x+u} \sum_{y=y-N_y+v}^{y+N_y+v} g(x,y) \right]$$

(22)

On the other hand, although Q$(x,y;u,v)$ can be calculated using a sum-table approach, the term P$(x,y;u,v)$ involves cross correlation terms between both images and cannot be calculated recurring to a sum-table approach, since (u,v) are sliding window parameters.

For the denominator calculation a global sum-table approach can be used:

$$F(x,y) = \sum_{x=x-N_x}^{x+N_x} \sum_{y=y-N_y}^{y+N_y} f^2(x,y) - \frac{1}{(2N_x+1)(2N_y+1)} \times \left[\sum_{x=x-N_x}^{x+N_x} \sum_{y=y-N_y}^{y+N_y} f(x,y) \right]^2 \tag{23}$$

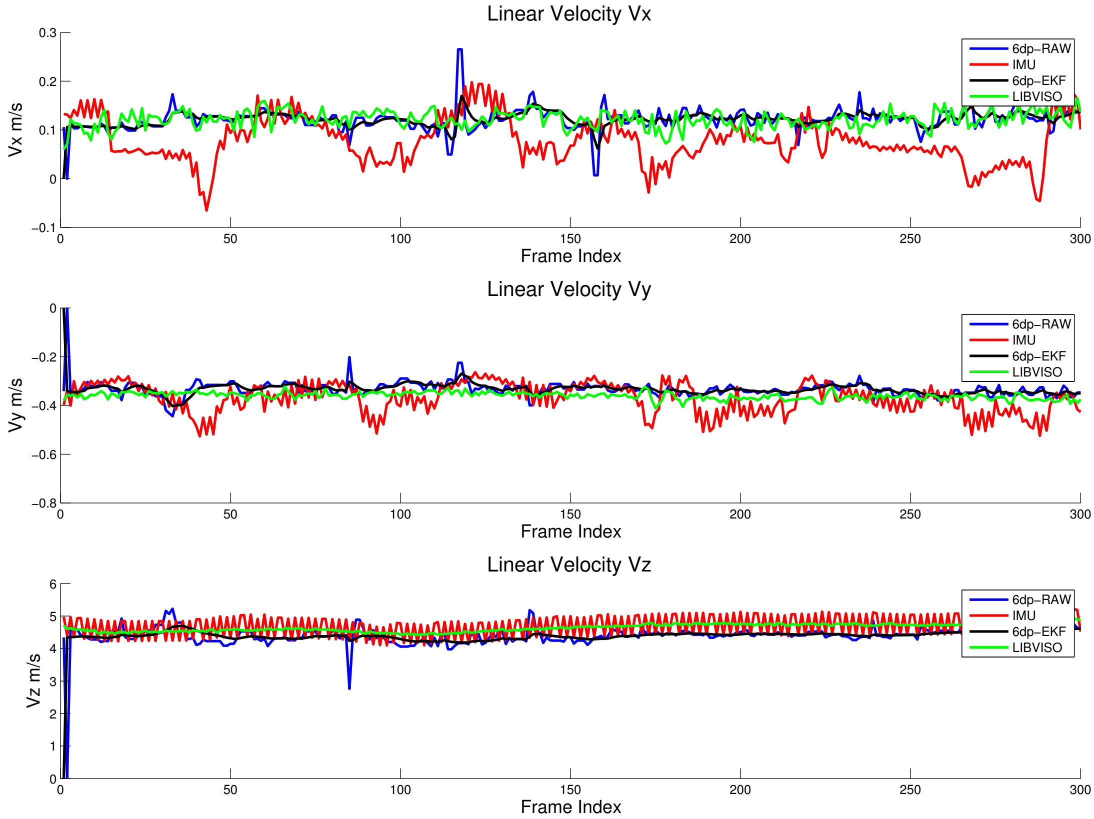

Fig. 12 Results for linear velocities estimation, where the LIBVISO implementation and 6DP-EKF display similar performance when compared to IMU/GPS performance.

$$G(x,y;u,v) = \sum_{x=x-N_x+u}^{x+N_x+u} \sum_{y=y-N_y+v}^{y+N_y+v} g^2(x,y) - \frac{1}{(2N_x+1)(2N_y+1)} \times \left[\sum_{x=x-N_x+u}^{x+N_x+u} \sum_{y=y-N_y+v}^{y+N_y+v} g(x,y) \right]^2 \tag{24}$$

where the four global sum schemes can be calculated as an integral window approach.

References

1. Alismail, H., Browning, B., Dias, M.B.: 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)
2. Bonin-Font, F., Ortiz, A., Oliver, G.: Visual navigation for mobile robots: A survey. J. Intell. Robotics Syst. **53**, 263–296 (2008)
3. Comport, A., Malis, E., Rives, P.: Real-time quadrifocal visual odometry. The International Journal of Robotics Research **29**(2-3), 245–266 (2010)
4. Domke, J., Aloimonos, Y.: A Probabilistic Notion of Correspondence and the Epipolar Constraint. In: Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06), pp. 41–48. IEEE (2006)
5. Fischler M.A., B.C.: Random sample consensus a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
6. Goodall, C.: Procrustes Methods in the Statistical Analysis of Shape. Journal of the Royal Statistical Society. Series B (Methodological) **53**(2), 285–339 (1991)
7. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of The Fourth Alvey Vision Conference, pp. 147–151 (1988)
8. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, second edn. Cambridge University Press, ISBN: 0521540518 (2004)
9. Howard, A.: Real-time stereo visual odometry for autonomous ground vehicles. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2008, pp. 3946–3952. Ieee (2008)
10. Huang, J., Zhu, T., Pan, X., Qin, L., Peng, X., Xiong, C., Fang, J.: A high-efficiency digital image correlation method based on a fast recursive scheme. Measurement Science and Technology **21**(3) (2011)
11. Kai, N., Dellaert, F.: Stereo tracking and three-point/one-point algorithms - a robust approach. In: Visual Odometry, In Intl. Conf. on Image Processing (ICIP, pp. 2777–2780 (2006)
12. Kazik, T., Kneip, L., Nikolic, J., Pollefeys, M., Siegwart, R.: Real-time 6d stereo visual odometry with non-overlapping fields of

view. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 1529 –1536 (2012)

13. Kitt B., G.A., Lategahn, H.: Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. In: IEEE Intelligent Vehicles Symposium (IV), pp. 486–492. IEEE (2010)
14. Kneip, L., Chli, M., Siegwart, R.: Robust real-time visual odometry with a single camera and an imu. In: Proc. of the British Machine Vision Conference (BMVC) (2011)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
16. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. pp. 674–679 (1981)
17. Maimone, M., Matthies, L., Cheng, Y.: Visual Odometry on the Mars Exploration Rovers. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 903–910. IEEE (2005)
18. Maimone, M., Matthies, L., Cheng, Y.: Two years of visual odometry on the mars exploration rovers: Field reports. J. Field Robot. **24**(3) (2007)
19. Milella, A., Siegwart, R.: Stereo-based ego-motion estimation using pixel tracking and iterative closest point. In: in IEEE International Conference on Computer Vision Systems, p. 21 (2006)
20. Moreno, F., Blanco, J., González, J.: 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, pp. 932–942. Springer-Verlag (2007)
21. Nelder, J.A., Mead, R.: A simplex method for function minimization. The Computer Journal **7**(4), 308–313 (1965). DOI 10.1093/comjnl/7.4.308
22. Ni, K., Dellaert, F., Kaess, M.: Flow separation for fast and robust stereo odometry. In: IEEE International Conference on Robotics and Automation, ICRA 2009, vol. 1, pp. 3539–3544 (2009)
23. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE Trans. Pattern Anal. Mach. Intell. **26**, 756–777 (2004)
24. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry for ground vehicle applications. Journal of Field Robotics **23**(1), 3–20 (2006)
25. Obdrzalek, S., Matas, J.: A voting strategy for visual ego-motion from stereo. In: 2010 IEEE Intelligent Vehicles Symposium, pp. 382–387
26. Olson, C., Matthies, L., Schoppers, M., Maimone, M.: Rover navigation using stereo ego-motion. Robotics and Autonomous Systems **43**, 215–229 (2003)
27. Rehder, J., Gupta, K., Nuske, S.T., Singh, S.: Global pose estimation with limited gps and long range visual odometry. In: IEEE Conference on Robotics and Automation (2012)
28. Rusinkiewicz, S., Levoy, M.: Efficient variants of the ICP algorithm. In: Third International Conference on 3D Digital Imaging and Modeling (3DIM) (2001)
29. Scaramuzza D., F.F.: Visual odometry tutorial, part i. Robotics Automation Magazine, IEEE **18**(4), 80 –92 (2011)
30. Silva, H., Bernardino, A., Silva, E.: Combining sparse and dense methods for 6d visual odometry. 13th IEEE International Conference on Autonomous Robot Systems and Competitions, Lisbon Portugal (April 2013)
31. Zhang, Z., Deriche, R., Faugeras, O., Luong, Q.T.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. Artificial Intelligence Special Volume on Computer Vision **78**(2), 87 – 119 (1995)