Sensors **2014**, *14*, 6317-6337; doi:10.3390/s140406317

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

Monocular SLAM for Autonomous Robots with Enhanced Features Initialization

Edmundo Guerra [1], Rodrigo Munguia [2] and Antoni Grau [1,*]

[1] Automatic Control Department, Technical University of Catalonia UPC, 08028 Barcelona, Spain; E-Mail: edmundo.guerra@upc.edu

[2] Computer Science Department, CUCEI, Universidad de Guadalajara, 44430 Guadalajara, JAL, Mexico; E-Mail: rodrigo.munguia@upc.edu

***** Author to whom correspondence should be addressed; E-Mail: antoni.grau@upc.edu; Tel.: +34-93-401-6975; Fax: +34-93-401-7045.

Received: 6 January 2014; in revised form: 19 March 2014 / Accepted: 26 March 2014 / Published: 2 April 2014

Abstract: This work presents a variant approach to the monocular SLAM problem focused in exploiting the advantages of a human-robot interaction (HRI) framework. Based upon the delayed inverse-depth feature initialization SLAM (DI-D SLAM), a known monocular technique, several but crucial modifications are introduced taking advantage of data from a secondary monocular sensor, assuming that this second camera is worn by a human. The human explores an unknown environment with the robot, and when their fields of view coincide, the cameras are considered a pseudo-calibrated stereo rig to produce estimations for depth through parallax. These depth estimations are used to solve a related problem with DI-D monocular SLAM, namely, the requirement of a metric scale initialization through known artificial landmarks. The same process is used to improve the performance of the technique when introducing new landmarks into the map. The convenience of the approach taken to the stereo estimation, based on SURF features matching, is discussed. Experimental validation is provided through results from real data with results showing the improvements in terms of more features correctly initialized, with reduced uncertainty, thus reducing scale and orientation drift. Additional discussion in terms of how a real-time implementation could take advantage of this approach is provided.

Keywords: monocular SLAM; human-robot interaction; HRI; stereo matching; depth estimation

---

1. Introduction

Sensors are widely used in several scientific and technical fields like robotics enabling the perception of the environment and its elements surrounding the robotic systems. This has led to the development of several sensor-based problems within the field, such as simultaneous localization and mapping (SLAM). The SLAM problem states how a mobile robotic device can operate in an *a priori* unknown environment by means of only onboard sensors to simultaneously build a map of its surroundings and use it to track its position. Thus, the SLAM is one of the most important problems to solve in robotics heavily related with sensors and its applications.

Many approaches have been developed to deal with the SLAM problem, based on a wide selection of sensors and combinations of them. Generally speaking, exteroceptive sensors can be used to solve both mapping and localization, while proprioceptive sensors are only able to deal with localization. This difference makes exteroceptive sensors more useful in the SLAM context, and if the mapping is a requirement to deal with in any given approach, it will forcibly include this kind of sensors. In [1,2] many types of available sensors are discussed and the main drawbacks are commented. These characteristics and the continuous development of better yet cheaper camera devices produced a surge in camera-based SLAM works during the last decade.

The camera, as a sensor, provides huge amounts of data and information, which can be used to deal with several problems using techniques and approaches developed in computer vision. The extraction of features and data association problems can be treated with relative ease in a camera-based SLAM. The main issue with the utilization of cameras for SLAM is the depth estimation. Since each pixel on a camera sensor maps the view from a ray, its depth information cannot be obtained directly. The only way to obtain depth estimations with visual information is through triangulation, relying on at least two different images. This can be achieved in two ways: images from a single camera separated spatially and temporally, thus, producing the monocular approach; or taking images with different cameras simultaneously, producing the stereo vision approach. The latter approach relies generally in the epipolar geometry [3], and normally implies creating a system where the different cameras (as it is not necessarily restricted to two cameras) are of similar characteristics and calibrated. Besides, these systems follow a set of restrictions to optimize the data association exploiting epipolar geometry features. Those characteristics make the stereo SLAM an easier problem, with the weaknesses of requiring more computational power to deal with two or more video sequences at the same time, and the limit to the maximum depth that can be estimated imposed by the baseline of the rig.

On the other hand, monocular SLAM approaches constitute more complex problems normally, especially if the six DoF have to be recovered from an only-bearing sensor. Most of the approaches originated in robotics have been produced through filtering techniques, many of them derived from the Extended Kalman Filter (EKF). Many notable works have used other sensors besides the camera to help to estimate the parallax in order to find the depth value [4,5]. In [6,7], a good survey of approaches to SLAM is presented where several methods with different filtering techniques are discussed. Still, some of the most relevant works on SLAM are based on EKF, like the feasibility of real-time EKF monocular SLAM [8], the development of the inverse-depth (I-D) feature parametrization model [9], and more recently real-time relocalization [10,11].

The analogous problem in computer vision, known as structure from motion (SfM), has produced several solutions over time [12]. Although originally they were conceived as off-line solutions given their reliance on global non-linear optimization, these solutions eventually led to the creation of the keyframe methods. Though these methods are gaining popularity, they still rely on bundle adjustment [13], and so they are not the best option when the computational power might be a problem [14], and as they estimate the map only using some of the frames [15], a filtering approach able to reject data association errors [14] can produce maps with the similar levels of accuracy at a lesser cost.

A growing field in robotics research deals with the interaction of human and robotic devices, known as human-robot interaction (HRI) [16]. These trends also affect SLAM, as several recent works reveal, like exploration of large areas with a wide group of people and robots [17], or mapping a building explored by a human [18]. While this implies increased complexity in several aspects, like large maps management or data fusion from very different sensors, it also opens new strategies and approaches to several SLAM problems. In a collaborative context, monocular SLAM could be improved in terms of depth estimation of long range features, which both in the delayed [1] and undelayed approaches [19] tend to be difficult, as well as the initialization of a metric scale, which normally requires the introduction of a known scale information, *etc*.

In this work, authors present a collaborative framework for local scale SLAM based on previous works. Thus, the approach presented and discussed in [1,2,20,21], the delayed inverse-depth monocular SLAM (DI-D SLAM), is enhanced by enabling the utilization of data obtained through a different monocular sensor. This monocular camera is assumed to be worn by a collaborating human, which helps the robotic camera performing SLAM to explore a particular unknown environment. Section 2 briefly describes how the DI-D monocular SLAM works using HOHCT validation, detailing opportunities to improve it. Section 3 presents a brief discussion about stereo estimation, considering the characteristics found in our approach (non-constant, approximately known but variable calibration), and justifying the approach taken, before discussing which issues have been solved and the way to do so. The first issue solved is the metric scale initialization, and the second improved characteristic is the problem appeared when initializing features within the direction of movement. Section 4 describes the built experimental setup and some of the experiments, along with the assumptions done during data capture. This is followed by comments and discussion on the obtained results: achieved improvement, theoretical costs associated to those improvements, and the possibilities of a real-time implementation, describing possible architectures and optimizations.

2. Monocular SLAM with DI-D Initialization

The procedure for local monocular DI-D SLAM with a pinhole camera can be summarised in the terms of Algorithm 1. Based on the inverse-depth parametrization, it tracks landmarks through frames until they are initialized. The augmented state vector, $\hat{\mathbf{x}}$ (Equation (1)), required by the EKF, is used to maintain the monocular camera position and the map. The first part of this column vector contains a vector $\hat{\mathbf{x}}_{\mathbf{v}}$ that represents a robotic camera device, describing its pose and movement speeds (Equation (2)). The position of the camera optical centre is represented by $\mathbf{r}^{WC}$, while its orientation with respect to the navigation frame is represented by a unit quaternion $\mathbf{q}^{WC}$. Linear and angular velocities are described by $\mathbf{v}^{W}$ and $\boldsymbol{\omega}^{W}$ respectively. The environment map estimation is represented by a set of N features $\hat{\mathbf{y}}_{\mathbf{i}}$,

with i = [1, *N*]. Each feature $\hat{\mathbf{y}}_\mathbf{i}$ is stored as a vector which models the estimated feature localization (Equation (3)) with respect to the world coordinates according to the inverse-depth model [9]. Coordinates x_i, y_i, z_i are the optical center of the camera when the feature was seen for the first time; θ_i, ϕ_i represent azimuth and elevation for the ray which traces the feature point; depth r_i to the feature is coded by its inverse: $\rho_i = 1/r_i$, see Figure 1.

$$\hat{\mathbf{x}} = \left[\hat{\mathbf{x}}_\mathbf{v}, \hat{\mathbf{y}}_\mathbf{1}, \ldots \hat{\mathbf{y}}_\mathbf{n}\right]^T \tag{1}$$

$$\hat{\mathbf{x}}_v = \left[\mathbf{r}^{WC} \quad \mathbf{q}^{WC} \quad \mathbf{v}^W \quad \boldsymbol{\omega}^W\right]^T \tag{2}$$

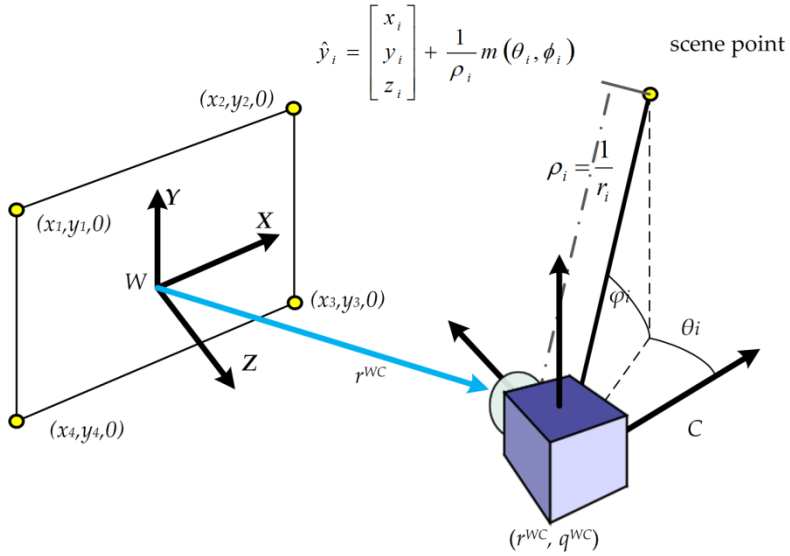
$$\hat{\mathbf{y}}_\mathbf{i} = \left[x_i \quad y_i \quad z_i \quad \theta_i \quad \phi_i \quad \rho_i\right]^T \tag{3}$$

Algorithm 1 Monocular SLAM with DI-D feature initialization (vid, refpose).

vid sequential image video
ref_{pose} initial reference points

begin
k: = 0
($\hat{x}_0$, P_0): = *Initialize* (*vid.FirstFrame*(), ref_{pose})
while true **do**
img: = *vid.NextFrame*()
($\hat{x}_{k+1}$, P_{k+1}): = *StatePrediction* ($\hat{x}_k$, P_k)
(h_k, ∇H_k): = *MeasurementPrediction* ($\hat{x}_{k+1}$, P_{k+1})
(z_k, S_k): = *Matching&Validation* (h_i, ∇H_i,P_{k+1},img)
($\hat{x}_k$, P_k): = *Update* ($\hat{x}_{k+1}$, P_{k+1}, h_k, z_k, S_k)
($\hat{x}_k$, P_k): = *AddFeatures*($\hat{x}_k$, P_k, img)
k: = k + 1
end while
end

Figure 1. Inverse-depth problem with PnP scale initialization, adapted from reference [2].



A metric scale initialization is required to estimate parallax in the DI-D monocular SLAM approach. Thus a set of accurate features are required initially, so that other new features can be initialized afterwards. This initialization process, based on the PnP problem [22], allows to set an initial value for the extrinsic camera parameters, and to map four points (see Figure 1). The problem of this approach is that those four points had to be coplanar, with known spatial relationships, and be able to be introduced manually on the algorithm or through artificially calibrated features guaranteeing to be seen at the start of the video sequence. Other common approach is producing unknown scale maps, introducing the scale later on. Both approaches are undesired for exploration tasks. Once the state is initialized, the EKF procedure is started. The camera is assumed to follow an unconstrained constant-acceleration camera motion prediction model [8,23], with Gaussian noise processes used to produce impulses with linear and angular speeds to move the camera, as seen on Equation (4):

$$f_v = \begin{bmatrix} \mathbf{r}_{k+1}^{WC} \\ \mathbf{q}_{k+1}^{WC} \\ \mathbf{v}_{k+1}^{W} \\ \boldsymbol{\omega}_{k+1}^{W} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_k^{WC} + \left(\mathbf{v}_k^W + \boldsymbol{\upsilon}_k^W\right)\Delta t \\ \mathbf{q}_{k+1}^{WC} \times \mathbf{q}\left(\left(\boldsymbol{\omega}_k^W + \boldsymbol{\zeta}_k^W\right)\Delta t\right) \\ \mathbf{v}_k^W + \boldsymbol{\upsilon}_k^W \\ \boldsymbol{\omega}_k^W + \boldsymbol{\zeta}_k^W \end{bmatrix} \tag{4}$$

$$P_{k+1} = \nabla F_x P_k \nabla F_x{}^T + \nabla F_u Q \nabla F_u{}^T \tag{5}$$

The features are assumed to remain static, which is the hypothesis generally used for mapping. As mapping dynamic features would damper the map, the validation of this used data association will remove them. The uncertainty of this prediction model is propagated through the covariance matrix (Equation (5)), where ∇F_x and ∇F_u are the Jacobian of the prediction model and process noise, respectively.

$$\mathbf{h}^c = \begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix} = R^{CW}\left(\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + \frac{1}{\rho_i}\mathbf{m}\left(\theta_i, \phi_i\right) - r^{WC}\right) \tag{6}$$

The observation model of a point $\hat{\mathbf{y}}_i$ computes a ray expressed in the camera frame as (Equation (6)), where $\mathbf{h}^C$ is observed by the camera through its projection in the image. R^{CW} is the transformation matrix from the global reference frame to the camera reference frame. The Jacobian of the measurement model is computed at the same time, as it will be needed through the data matching phase. The matching process is performed through active search with a patch cross-correlation technique between images. The search is limited to elliptical search regions derived from the innovation matrix (Appendix, Equation (a4)). The set of pairing obtained are validated with a batch gating approach based on the "highest order hypothesis compatibility test" (HOHCT) [2,24], which makes an ordered search to validate the data association pairing using the Mahalanobis distance [25], outperforming the Joint Compatibility Branch & Bound (JCBB). The state and covariance update equations follow classical EKF formulation (see Appendix Equations (a1)–(a4)).

To add features to the state, the DI-D feature initialization is used, [1,26,27]. Based on stochastic triangulation, a hypothesis of initial depth for a feature using a delay is defined. To achieve this initialization a database of candidate feature points is created, and these landmarks are tracked until

they achieve enough parallax to produce depth estimation, otherwise they are rejected. This parallax is estimated thanks to the initial metric scale initialization, which allows estimation of the camera trajectory. Still, this delayed approach can probe a weakness when the camera trajectory is not smooth enough to allow continuous tracking of features. Keyframe methods can implement techniques [28] to deal with these problems at the cost of higher computational requirements or even the introduction of additional sensors [29].

3. Feature Based Stereo Estimation of Depth

3.1. Stereo Discussion

Some of the weak points discussed before can be solved within a cooperative context exploiting data from another monocular camera. Assuming that the other camera device (the "free camera", or C_f) with known pose is near to the robotic camera performing SLAM ("SLAM camera", or C_s), joining observations from both cameras allow performing stereo-like estimation when their fields of view overlap. This way, a new non-constant stereo inverse-depth feature initialization approach will be used to address the issues.

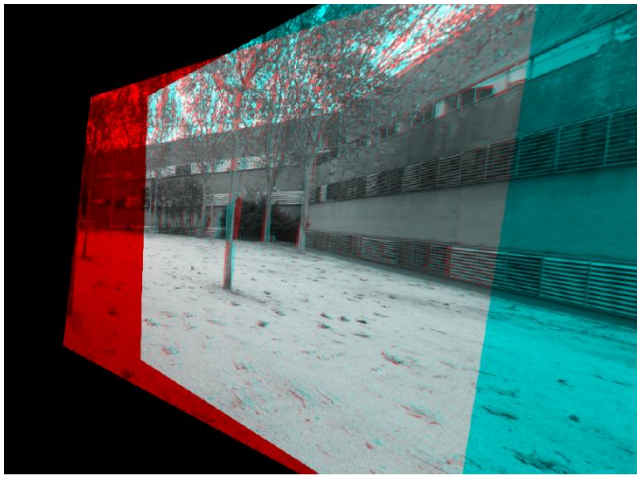
Classical stereo approaches [30–32] rely on epipolar geometry to create a calibrated camera rig with multiple constraints. These constraints typically include that both cameras' projection planes lie in the same plane in world coordinates, thus allowing optimization of the correspondence problem, as the match on an image of another's image pixel will lie in the corresponding epipolar line, and rectification can turn them into straight-lines parallel to the horizontal axis. Several works have dealt with rectification of stereo images for unrestricted pose cameras both calibrated [31] and uncalibrated [33,34].

In [31], Fusiello detailed the first method to rectify stereo pairs with any given pairs of calibrated cameras. The method is based on rotating the cameras until they have one of their axis aligned to the baseline, and forcing them to have their projective planes contained within the same plane to achieve horizontal epipolar lines. Other works have proposed similar approaches to rectification of stereo pairs assuming calibrated, uncalibrated, or even multiple view [35–37] stereo rigs. These approaches need to warp both images according to the rectification found (see Figure 2 *versus* Figures 3 and 4), and in some cases producing great variations in terms of orientation and scale (Figure 4), thus rendering them less attractive in terms of our approach.

Figure 2. Pair of images captured at experimental environment.

Figure 3. Pair of images rectified according to [31].

Figure 4. Stereo pair rectified and matched.



At any case, dealing with stereo features without rectified images is not a big problem in the proposed approach. As next subsections will describe further, the process of stereo features search and matching will be done sparsely, only to introduce new features: during the initialization, or when the filter needs new features. For both cases only a part of the image will be explored, and when adding new features in a system already initialized, additional data from the monocular phase can be used to simplify the process.

3.2. Initialization of Features with Non-Constant Stereo Estimation

The requirement of metric scale initialization of the DI-D method can be easily solved under the cooperative assumption. Authors' previous approach required the presence of a set of known, easily identifiable features to estimate them initially through the PnP problem. Then, assuming that at the start of the exploration a cooperating, free moving camera is near, the data from this camera can produce the features needed through stereo estimation. This process is shown in the diagram in Figure 5, where, after the poses of the SLAM camera C_s and the free camera C_f are known, the maximum distance from a camera where a point with a given minimum parallax (pl_{min}) could lie is found. This distance is employed to build a model of the field of view of each camera, as a pair of pyramids, with each apex in the position of a camera, and the base centered along the view axis. Then it can be

guaranteed that any point with parallax—between cameras—equal or greater than pl_{min} will lie in the space intersected by the two fields of view modelled as pyramids, as seen in Figure 6. So the intersection between the different polygons composing the pyramids is computed as a set of segments (two point tuples), as described by Algorithm 2. Once all the segments are known, they are projected into the 2D projective space of each camera, and a search region is adjusted around them, determining the regions of interest where the stereo correspondence may be useful and significant.

Figure 5. New metric scale initialization process.

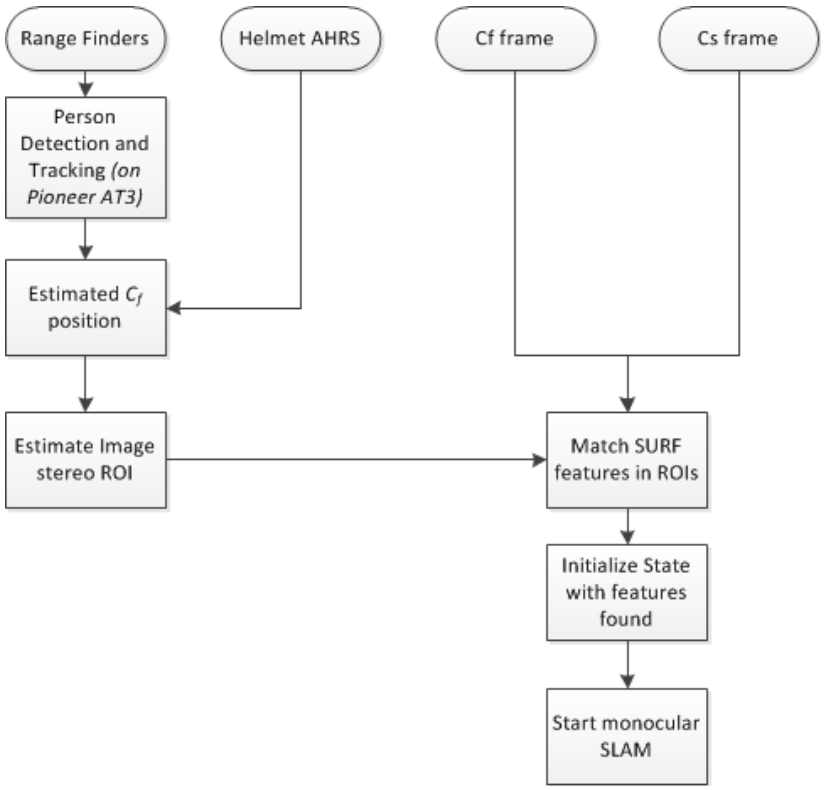
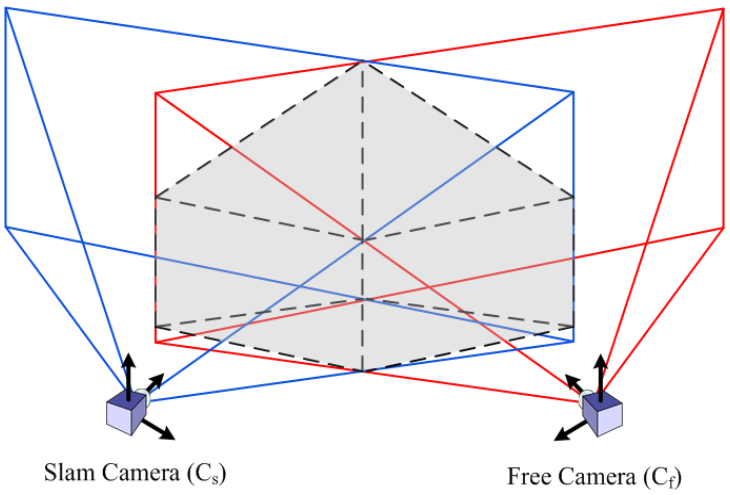


Figure 6. Polyhedron found intersecting fields of view until a depth where minimum parallax pl_{min} could be found.



Algorithm 2 (ris, rif): = Find Stereo ROI (cams, camf, plmin).

cam_s, cam_f	SLAM camera, free camera (pose, intrinsic matrix)
pl_{min}	minimum parallax considered meaningful
ri_s, ri_f	stereo ROI for images from cam_s and cam_f respectively

```
begin
  distance: = FindDistance (cam_s.pose, cam_f.pose)
  PyramidDepth: = FindMaxDepth (distance, pl_min)
  Py1: = ModelFoV(cam_s,PyramidDepth)
  Py2: = ModelFoV(cam_f,PyramidDepth)
  intersection = Ø
  for each polygon_i in Py1
    for each polygon_j in Py2
      segment: = Intersect(polygon_i, polygon_j)
      intersection.add(segment)
    end for
  end for
  ri_s: = Ø; ri_f: = Ø
  if (intersection = Ø) then
    ri_s: = Envelope(ProjectTo2D(cam_s.pose, intersection.points))
    ri_f: = Envelope(ProjectTo2D(cam_f.pose, intersection.points))
  end if
end
```

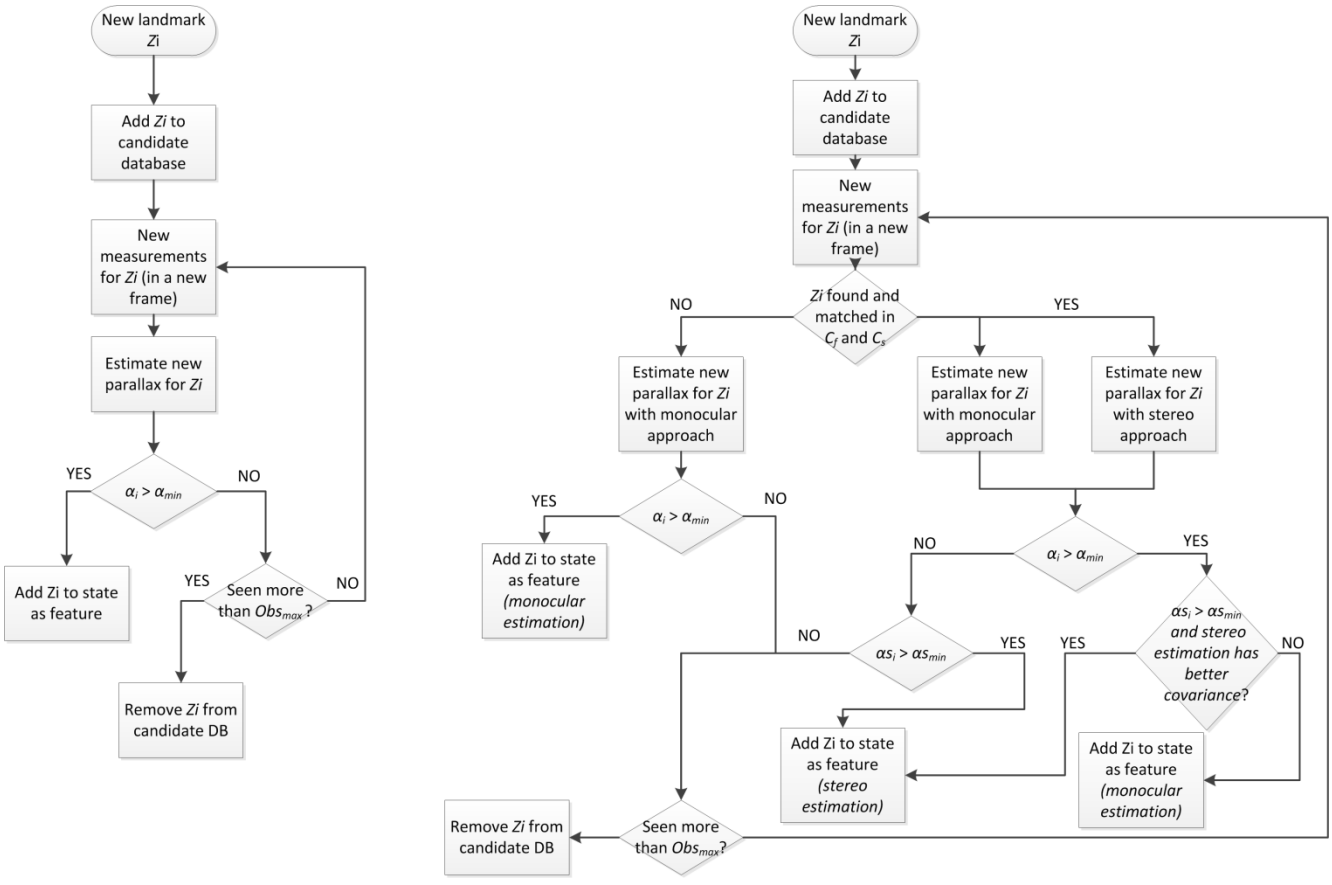
In the interest regions found, SURF-based feature descriptors [38] are matched to produce up to ten new stereo features to initialize the EKF state vector, see the end of the process at Figure 5. SURF is chosen over SIFT and FAST due to the more convenient trade-off offered in terms of matching accuracy and efficiency [39]. Each pair of matched features allows estimating the world coordinates of the feature point seen with simple triangulation, backtracking the point on the images from the SLAM camera and the free camera. Then, the set of landmarks found and estimated are introduced in the monocular EKF according to the inverse depth model.

3.3. Introducing New Features under Stereo

As described in [1,26], the DI-D initialization adds new landmarks into the map when a feature achieves parallax enough. This process may be easily disrupted if the features cannot be tracked long enough due to motion blur, illumination problems, trajectory irregularities, *etc.*, probably disrupting the filtering performance and convergence. The introduction of data association validation, though generally improves convergence of the filter [2,24], may also deprive the filter of features, as it may remove them faster that they are introduced. These problems can be mitigated under the assumption of a temporary stereo correspondence between cameras C_s and C_f, introducing the features much earlier with accurate depth estimation, using a non-constant stereo I-D feature initialization approach.

Figure 7 shows the schemes for the single camera feature initialization process (Figure 7, left), and the new initialization process (Figure 7, right), with the help of the free camera C_f. The approach shown on the left follows a classic strategy of storing and tracking candidate landmarks, detected through the Harris salience operator, to see if enough parallax is reached ($\alpha_i > \alpha_{min}$) within a given number of frames, and then to initialize them with an estimated depth value (with respect to the camera). On the proposed approach (Figure 7, right), if at prediction step stereo correspondence is found, the process to introduce new features will have different chances to introduce candidate landmarks as features. Those candidates that have enough parallax will be given preference to be inserted as full features ($\alpha_i > \alpha_{min}$). The depth estimation for these features will have the most available accuracy, obtained through parallax or through stereo estimation (available if $\alpha s_i > \alpha s_{min}$). Secondly, candidate landmarks which have not achieved enough parallax, but they are within the overlap region, are considered to be initialized as full features if they comply with depth estimation obtained through stereo matching. The last resource to initialize features would be finding new landmarks, and introduce them using only stereo matching, just like in the state initialization. These last two cases will also have to comply with restrictions (available if $\alpha s_i > \alpha s_{min}$) to guarantee a minimal accuracy when estimating the feature depth.

Figure 7. Feature initialization process according to the single monocular camera approach (**left**) and the proposed method with non-constant stereo I-D feature initialization (**right**).



4. Experiments

The approach proposed in this work has been implemented in MATLAB® to test and evaluate it. A set of five sequences corresponding to the same trajectory has been captured in a semi-structured environment. The sequences have been reduced to a resolution of 720 × 480 pixels and greyscale color, shortening the computational effort for image processing. Each sequence corresponds to a collaborative exploration of the environment at low speed, including a human and a robotic platform, each one equipped with the monocular sensors assumed earlier, C_f and C_s, respectively. The data collected include monocular sequences, odometry from the robot, estimation of the human pose with respect to the robot, and the orientation of the camera for the five sequences.

Figure 8. Robotic platform Pioneer AT3 with webcam and laser range finders. Helmet with camera and AHRS.

The robot role is performed by a robotic platform based on the Pioneer 3 AT (Figure 8). The platform runs a ROS Fuerte distribution over an Ubuntu 12.04 OS. The platform is equipped with a pair of Leuzer RS4-4 laser range finders and a Logitech C920 webcam, able to work up to 30 frames per second (fps) at a resolution of 1,080 pixels. The sensors worn by the human are deployed on a helmet, including a GoPro Hero camera and an Xsens AHRS (Figure 9). All the data, except the GoPro camera video, have been captured and synchronized through ROS in the platform hardware. The ROS middleware provides the necessary tools to record and timestamp the data from the sensors connected to the platform. The C_f image sequence was offline synchronized with the C_s image sequence by marking frames with a laser pointer observed in both C_f and C_s sequences.

Figure 9. View of the helmet with the camera and AHRS unit placement detail.

To estimate the pose of C_f, orientation data from the AHRS are combined with the approximate pose of the human, estimated with the range finders [40,41]. The final position of the camera is computed geometrically as a translation from the estimated position of the atlas and axis vertebrae (which allow most of the freedom of movement of the head). These vertebrae are considered to be at a vertical axis over the person position estimated with the range finders, with height modeled individually for each person. In this work it is assumed that the environment is a flat terrain, easing the estimation process.

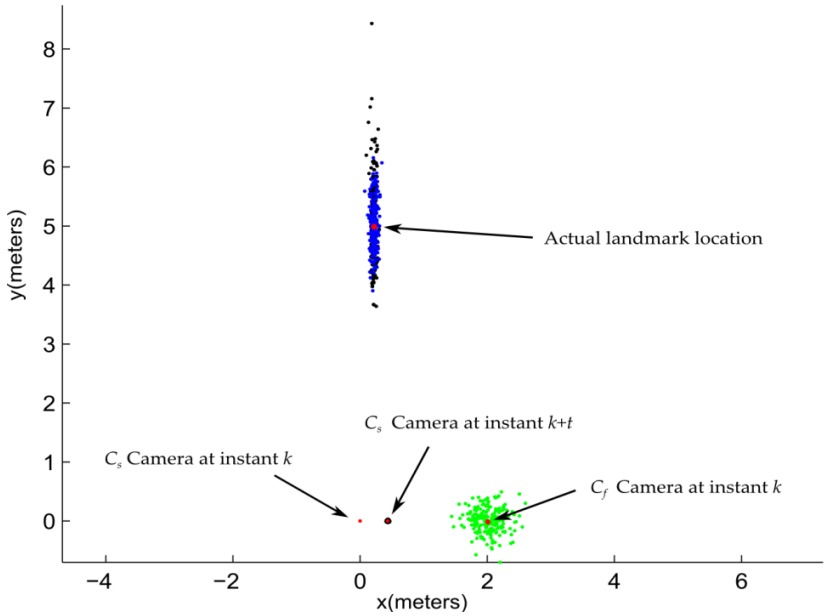
It is worth noting that for the proposed method, the pose of the camera worn by the human respect to the SLAM camera is not assumed to be perfectly known. Instead, it is considered that when needed, a "noisy" observation of the pose of C_f respect C_s is available by means of the methodology described above. The inherent error to the observation process is modeled assuming that the observation is corrupted by Gaussian noise. The value of the parameters used to model the inaccuracies for computing the pose of C_f were obtained statistically by comparing actual and estimated values. It is also important to note that an alternate method could be used for computing the relative pose of C_f, for instance using different sensors. However, even with the use of a more reliable methodology the errors would not be completely eliminated.

On the bright side, the errors introduced by the pseudo-calibrated stereo approach are not as critical as one could be expecting. Figure 10 shows the results obtained from simulating the initialization of a single feature using: (i) the ID-delayed monocular method; and (ii) the pseudo-calibrated stereo rig approach. In the simulation, the C_s is located at $[x, y] = [0, 0]$ at instant k. C_f is located at $[x, y] = [2, 0]$ at instant k. Thus, it is assumed that the base-line between C_s and C_f is equal to 2 meters. A landmark is located at $[x, y] = [0.21, 5]$. For comparison purposes it is assumed that C_s was moved (at some instant $k + t$) to its right to $[x, y] = [0.42, 0]$ in order to generate a parallax equal to 5 degrees. This amount is a typical value used as a threshold in the ID-Delayed method for initializing new features.

In the simulation, the drift associated with the estimated displacement of C_s is modeled by adding Gaussian noise with standard deviation $\sigma = 1$ cm to the actual location of C_s at instant $k + t$. The angular measurements provided by C_s are modeled by adding to its actual values Gaussian noise with $\sigma = 0.5$ degrees. In order to model the inaccuracies associated with the pseudo-calibrated stereo

rig approach, the estimated location of C_f was modeled by adding a Gaussian noise with $\sigma = 20$ cm to its actual location. The errors introduced by the AHRS device have been taken into account by considering that the angular measurements provided by C_f are corrupted by Gaussian noise with $\sigma = 1.5$ degrees.

Figure 10. Initialization of a single landmark using: (i) the ID-delayed monocular method; and (ii) the pseudo-calibrated stereo rig approach. The actual position of the cameras and the landmark are indicated by red dots. Estimated locations of the landmark obtained by the ID-delayed method are indicated by the cloud of black points. Estimated locations of the landmark obtained with the pseudo-calibrated stereo rig are indicated by the cloud of blue points. Every green point indicates an observation of the C_f location.



Using the above conditions, the location of the landmark has been estimated by triangulation with the location of C_s (at instant $k + t$) and with C_f. In both cases the location of C_s (at instant k) was used as common pivot. The experiment was carried out 200 times.

For this experimental setup, even considering that the location of C_f is poorly estimated, the likelihood region obtained with the pseudo-calibrated stereo rig is always smaller than the likelihood region obtained with the ID-delayed approach. This is because landmark depth estimation is heavily dependent on parallax. In this case, for the pseudo-calibrated stereo rig, the parallax is about 22 degrees.

5. Results and Discussion

5.1. Experimental Results

The introduction of an auxiliary monocular sensor which can provide non-constant stereo information has proven itself useful. One of the weaknesses discussed on earlier works [2] was the need to manually introduce an initial metric scale, which has been removed. This grants more autonomy to the system, exploiting the implicit human-robot interaction without enforcing utilization of artificial landmarks. Besides, as the metric scale initialization can introduce more features into the

initial state because is not limited to the artificial landmark, the scale propagates in a smoother way with reduced drift on the local scale.

Figure 11 shows results for one of the actual trajectories, with and without the utilization of the proposed non-constant stereo I-D feature initialization approach, right and left maps respectively. The introduction of stereo initialization of features allows introducing features with good depth estimation in a shorter time, thus making the system more resilient to quick view changes, such as turning. This can be seen on Figure 11 right, where the orientation drift is visibly minor. On the left trajectory estimation, the accumulated drift forces estimations so distant from actual observations within the data validation algorithm that most of the features are rejected. These rejections, combined with the drift itself, disrupt the estimation. On the other side, the trajectory estimated with non-constant stereo I-D feature initialization minimizes the drift and orientation deviation, thus keeping an accurate estimation even after the U-turn. Results obtained are consistent through all the 5 runs considered in the experimental setup, as shown on Figure 12 and Table 1.

Figure 11. Trajectory estimated with classical DI-D monocular SLAM (**left**) and with the new non-constant stereo I-D approach. Green line denotes robot ground truth, orange line denotes C_f ground truth, and the estimated C_s trajectory is shown in blue. Red features (only in left plot) have been artificially calibrated and introduced to have an initial scale estimation in the DI-D approach.

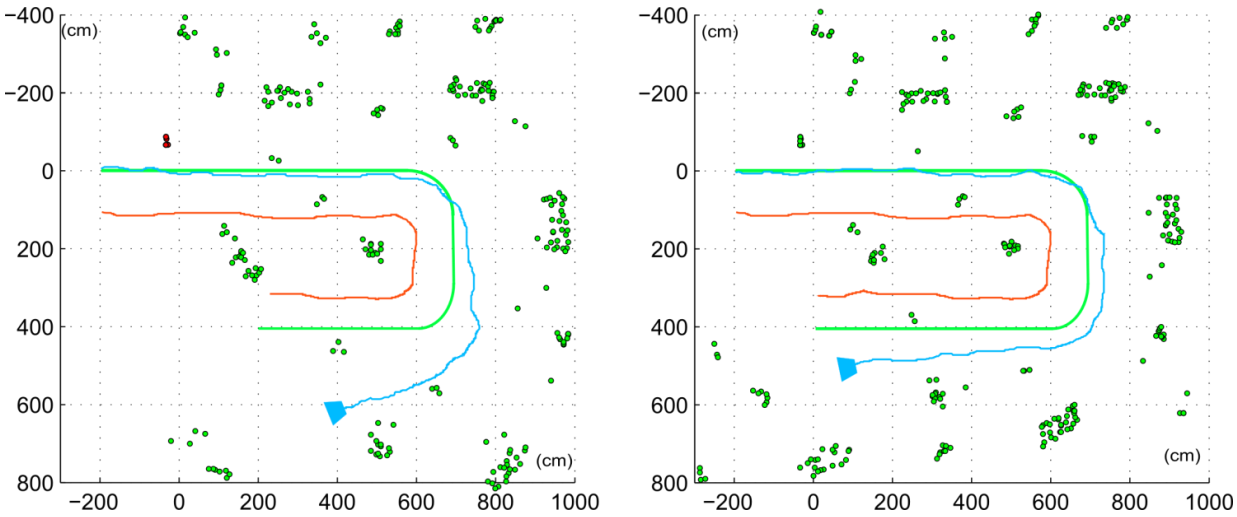


Figure 12. Estimations obtained for the rest of captured sequences, performing the trajectory several times and processed with the proposed approach.

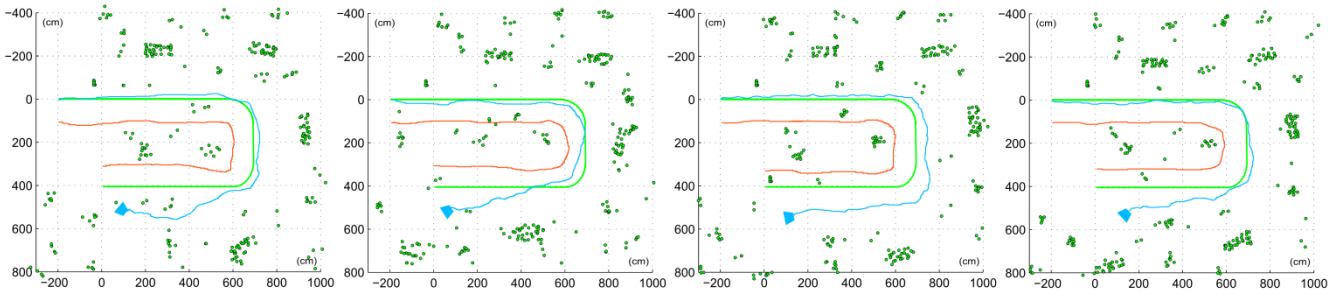
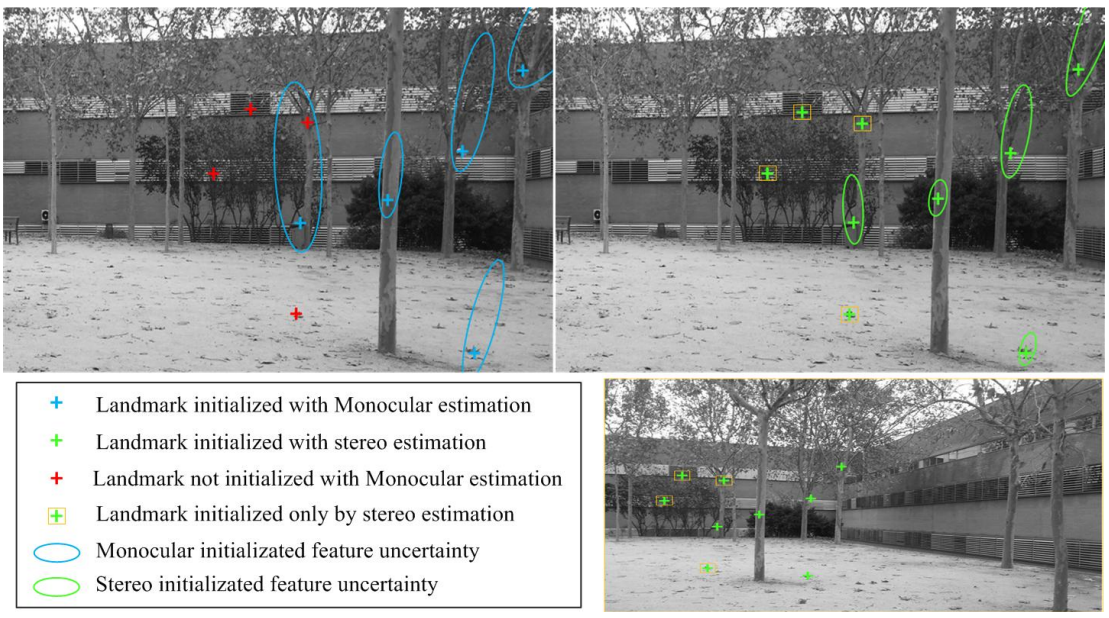


Table 1. Final pose estimation errors at the end of the trajectory.

Experiment # and Figure	Classic DI-D Initialization				Proposed Initialization			
	|x| (m)	|y| (m)	d (m)	Angle (°)	|x| (m)	|y| (m)	d (m)	Angle (°)
1 (Figure 11)	2.02	2.14	2.93	26	0.93	1.02	1.39	13.6
2 (Figure 12a)	1.53	2.92	3.29	−51	0.89	0.71	1.14	−29.4
3 (Figure 12b)	1.30	1.89	2.30	43	0.62	0.93	1.12	37.6
4 (Figure 12c)	3.41	3.17	4.64	77	1.07	1.38	1.75	14.2
5 (Figure 12d)	2.15	1.78	2.79	48	1.33	1.25	1.83	31.2

The introduction of the initialization of features with stereo estimation allows using more features that normally are rejected in the DI-D approach after being unable to achieve enough parallax. There are two cases were features are unable to achieve enough parallax: the first one is that they are distant features, and the camera would have to travel great distances in certain ways to see parallax; the second one affects features which lie in projective rays near the visual axis when camera moves forward. Several works [9,42] have dealt with distant features initializing them with heuristic values, as in an undelayed approach, as they tend to be useful to estimate orientation [26]. In the presented approach, the number of distant features used increases with respect to the DI-D approach, but those lying near the visual axis are the most benefit, as they will always present more parallax between C_f and C_s. Figure 13 shows an example of a set of features initialized thanks to the stereo estimation. Note how several distant features could not be initialized under the DI-D SLAM, but were actually introduced under the non-constant stereo I-D approach. Besides, as it was detailed on Figure 7 (left) and commented above, the non-constant stereo I-D approach initializes features with the most accurate estimation.

Figure 13. Landmarks initialized through stereo estimation. Left image: DI-D approach. Right top image: non-constant stereo I-D feature initialization. Right bottom image: image pair from C_f camera.



5.2. Costs and Analysis

The apparent increase of the computational effort that would suppose the utilization of the presented approach could be hard to justify within the field of filtering based SLAM, which generally try to keep reduced computational costs. But the cost increase is limited and could be further reduced. For our C_s sequence set, made of a total of 7,120 frames in five sequences, only 38.22% (2,721 frames) presented a field of view overlap with the C_f camera. While this fact supposed an overhead of processing almost 40% more images, the exploration area was reduced with the search of the stereo ROI. With respect to this, it is interesting seeing how the newly proposed approach made less effort per feature to be initialized, compensating the bigger number of features used.

Table 2 shows the features used on each approach and the tracking effort required until the initialization of the features is done. Note how the non-constant stereo I-D feature initialization approach uses about 4% more features, but the time required to initialize them is smaller. This is because many features that are being tracked are instantly initialized through stereo once they lay in the overlapped field of view. This is advantageous because it allows to introduce features known to be strong (enough to be tracked) directly without more tracking effort, compensating the effort used for the C_f processing and stereo based initialization.

Table 2. Statistics of features used and frames being tracked until initialization.

	DI-D Monocular SLAM	**Proposed Initialization**
Features initialized (total)	1487	1549
Features initialized (in average)	297.4	309.8
Average frames tracking a feature	24.6	10.4

Furthermore, in real-time applications employing this technique, the C_f sensor could be upgraded to an "intelligent" sensor, with real-time processing capabilities. This approach would integrate image processing in the C_f sensor, allowing parallel processing of SURF features, and sending only extracted features, minimizing communications delay. This processing step could be done while the robotic camera C_s makes the general EKF-SLAM process, and thus it would be possible to have the SURF landmarks information after the EKF update, in time for the possible inclusion of new features.

6. Conclusions

A new approach for feature initialization in SLAM is presented, the non-constant stereo I-D feature initialization. This is based on the DI-D approach [1,2,20,21], and heavily focused towards human-robot interaction framework, under the form of a collaborative exploration of the environment. The human collaborates through a second monocular sensor with total freedom of movement and approximately known pose. As this monocular sensors moves freely, sometimes its field of view will be concurrent with the field of view of the camera doing monocular SLAM, producing non-constant stereo pairing. As the relative pose between the cameras and the calibration matrices of each one of them are known, the fundamental matrix of a stereo system can be found. Even though this allows using stereo rectification based on epipolar geometry, it proved inconvenient for our approach, and SURF-based feature matching is used when needed.

The introduction of the non-constant stereo has allowed improvements on the performance of two specific aspects in the local scale framework. Firstly, the introduction of an initial metric scale through synthetic features has been removed, substituted by the initialization of a set of features based on stereo estimation. This depth estimation has proven to have a slighted advantage: as the number of features introduced initially is not limited to four coplanar points, the use of a greater number of features placed at more varied distances makes the metric scale propagation smoother. Secondly, the introduction of later landmarks through stereo enables utilization of far distance features with real depth estimation, instead of the heuristically assigned value used in other works, and the initialization of frontal landmarks when the camera C_s moves forward. These changes have produced a locally strong and robust SLAM approach, thus enabling its future utilization on larger scale SLAM, as commented on [14]. This would further reduce the drift of the estimated trajectory, thanks to the covariance reduction of loop closure.

Once the viability of the proposed approach has been demonstrated, further research should focus on maximizing the advantages obtained from the HRI, while studying with more depth the impact to a system in terms of costs. In terms of exploiting the HRI, the stereo data, when available, should be used widely, evaluating the impact of its introduction during the measurement and update step of the EKF SLAM. This would probably require a full overhaul of the prediction and observation models currently used, but should improve significantly the accuracy of the approach. In line with this overhaul, the use of non-constant stereo enables to reinitialize a metric scale when the field of view overlaps, permitting the introduction of submapping techniques and other techniques to improve map management and achieving larger trajectories, including loop closing.

The proposed technique could be also expanded, with some modifications, to deal with more C_f agents, *i.e.*, a group of different humans could explore an environment accompanied by a robot mapping their surroundings with data from the sensors deployed on the humans. While this approach would require much more computational power and a more insightful architecture, it would be of great interest due its resemblance to hypothetical real cases, where not a human alone but a team, would explore new zones with robotic assistance. Besides, this approach could be evaluated and compared with robot network approaches, similar to [17].

Acknowledgments

This research has been funded with the Spanish Science ministry project DPI2010-17112.

Author Contributions

The work presented in this paper corresponds to a collaborative development by all the authors. Antoni Grau and Rodrigo Munguia defined the research line. Edmundo Guerra and Antoni Grau designed and implemented the feature extraction based on stereo vision and performed the experiments, and Edmundo Guerra and Rodrigo Munguia designed and implemented the localization algorithms.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix: Nomenclature and Equations

Nomenclature

Nomenclature used during the paper and annexes:

Scalar

x_i, y_i, z_i	camera coordinates
θ_i, ϕ_i	camera azimuth and elevation
r_i	real depth to feature
ρ_i	inverse depth to feature
α_{min}	min. landmark parallax angle (monocular case)
αs_{min}	min. landmark parallax angle (stereo case)
pl_{min}	min. parallax to crop FoV intersection

Vector

$\hat{\mathbf{x}}$	augmented state vector
$\hat{\mathbf{x}}_v$	robot camera state vector
$\mathbf{r}^{WC}$	camera optical center position quaternion
$\mathbf{q}^{WC}$	robot camera orientation quaternion
$\mathbf{v}^{W}$	robot camera linear speed vector
$\boldsymbol{\omega}^{W}$	robot camera angular speed vector
$\hat{y}_i$	*i-th* feature vector
$\boldsymbol{f}_v$	camera motion prediction model
$\boldsymbol{\upsilon}^{W}$	linear speed pulse assumed by model
$\boldsymbol{\zeta}^{W}$	angular speed pulse assumed by model
$\mathbf{h}^{c}$	measurement prediction model
$\mathbf{g}$	Kalman innovation vector
$\mathbf{z}$	features measurement vector
$\mathbf{h}$	predicted features vector

Matrix

Q	process noise matrix
∇F_u	process noise Jacobian matrix
R^{CW}	world to camera transformation matrix
∇H_k	measurement model Jacobian matrix
S_k	innovation covariance matrix
R_{uv}	measurement noise matrix
W_k	Kalman gain matrix
P_k	augmented state covariance matrix
∇F_x	stated prediction model Jacobian matrix

Equations

The following equations describe the update step of the Extended Kalman Filter used, mentioned in Section 2 (full details in [1]):

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k+1} + W\mathbf{g} \tag{a1}$$

$$P_k = P_{k+1} - WS_iW^T \tag{a2}$$

$$W = P_{k+1}\nabla H_k^T S_k^{-1} \tag{a3}$$

$$S_k = \nabla H_k P_{k+1} \nabla H_k{}^T + R_{uv} \tag{a4}$$

References

1. Munguia, R.; Grau, A. Monocular SLAM for visual odometry: A full approach to the delayed inverse-depth feature initialization method. *Math. Probl. Eng.* **2012**, *2012*, ID676385.
2. Guerra, E.; Munguia, R.; Bolea, Y.; Grau, A. New validation algorithm for data association in SLAM. *ISA Trans.* **2013**, *52*, 662–671.
3. Cipolla, R. The visual motion of curves and surfaces. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **1998**, *356*, 1103–1121.
4. Piniés, P.; Lupton, T.; Sukkarieh, S.; Tardós, J.D. Inertial aiding of inverse depth SLAM using a monocular camera. In Proceedings of the IEEE International Conference on Robotics and Automation, Rome, Italy, 10–14 April 2007; pp. 2797–2802.
5. Strelow, D.; Singh, S. Optimal motion estimation from image and inertial measurements. In Proceedings of Workshop on Integration of Vision and Inertial Sensors (InerVis 2003), Coimbra, Portugal, June 2003.
6. Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* **2006**, *13*, 99–110.
7. Bailey, T.; Durrant-Whyte, H. Simultaneous localization and mapping (SLAM): Part II. *IEEE Robot. Autom. Mag.* **2006**, *13*, 108–117.
8. Davison, A.J. Real-time simultaneous localisation and mapping with a single camera. In Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; pp. 1403–1410.
9. Civera, J.; Davison, A.J.; Montiel, J.M.M. Unified inverse depth parametrization for monocular slam. In Proceedings of Robotics: Science and Systems Conference, Philadephia, PA, USA, August 2006.
10. Williams, B.; Klein, G.; Reid, I. Real-Time SLAM Relocalisation. In Proceedings of the IEEE 11th International Conference on Computer Vision, Rio do Janeiro, Brazil, 14–21 October 2007; pp. 21–28.
11. Grasa, O.G.; Civera, J.; Montiel, J.M.M. EKF monocular SLAM with relocalization for laparoscopic sequences. In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 4816–4821.
12. Besl, P.J.; McKay, N.D. A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1992**, *14*, 239–256.

13. Konolige, K.; Agrawal, M. FrameSLAM: From Bundle Adjustment to Real-Time Visual Mapping. *IEEE Trans. Robot.* **2008**, *24*, 1066–1077.
14. Strasdat, H.; Montiel, J.M.M.; Davison, A.J. Real-time monocular SLAM: Why filter? In Proceedings of the IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 3–7 May 2010; pp. 2657–2664.
15. Klein, G.; Murray, D. Parallel tracking and mapping for small AR workspaces. In Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan, 13–16 November 2007; pp. 225–234.
16. De Santis, A.; Siciliano, B.; de Luca, A.; Bicchi, A. An atlas of physical human-robot interaction. *Mech. Mach. Theory* **2008**, *43*, 253–270.
17. Kleiner, A.; Dornhege, C.; Dali, S. Mapping disaster areas jointly: RFID-Coordinated SLAM by Humans and Robots. In Proceedings of the IEEE International Workshop on Safety, Security and Rescue Robotics, Roma, Italy, 27–29 September 2007; pp. 1–6.
18. Fallon, M.F.; Johannsson, H.; Brookshire, J.; Teller, S.; Leonard, J.J. Sensor fusion for flexible human-portable building-scale mapping. In Proceedings of the IEEE/SRJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7–12 October 2012; pp. 4405–4412.
19. Sola, J.; Monin, A.; Devy, M.; Lemaire, T. Undelayed initialization in bearing only SLAM. In Proceedings of the IEEE/SRJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 2499–2504.
20. Munguia, R.; Grau, A. Monocular SLAM for Visual Odometry. In Proceedings of the IEEE International Symposium on Intelligent Signal Processing, Alcala de Henares, Spain, 3–5 October 2007; pp. 1–6.
21. Munguia, R.; Grau, A. Concurrent Initialization for Bearing-Only SLAM. *Sensors* **2010**, *10*, 1511–1534.
22. Williams, B.; Cummins, M.; Neira, J.; Newman, P.; Reid, I.; Tardós, J. A comparison of loop closing techniques in monocular SLAM. *Robot. Auton. Syst.* **2009**, *57*, 1188–1197.
23. Chiuso, A.; Favaro, P.; Jin, H.; Soatto, S. 3-D motion and structure from 2-d motion causally integrated over time: Implementation. In Proceeding of the European Conference on Computer Vision, Dublin, Ireland, 26 June–1 July 2000; pp. 734–750.
24. Guerra, E.; Munguia, R.; Bolea, Y.; Grau, A. Validation of Data Association for Monocular SLAM. *Math. Probl. Eng.* **2013**, *2013*, ID 671376.
25. Neira, J.; Tardós, J.D. Data association in stochastic mapping using the joint compatibility test. *IEEE Trans. Robot. Autom.* **2001**, *17*, 890–897.
26. Munguia, R.; Grau, A. Closing Loops with a Virtual Sensor Based on Monocular SLAM. *IEEE Trans. Instrum. Meas.* **2009**, *58*, 2377–2384.
27. Munguia, R.; Grau, A. Delayed inverse-depth feature initialization for sound-based SLAM. In Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation, Hamburg, Germany, 15–18 September 2008; pp. 817–824.
28. Klein, G.; Murray, D. Improving the agility of keyframe-based SLAM. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 802–815.

29. Weiss, S.; Achtelik, M.W.; Lynen, S.; Chli, M.; Siegwart, R. Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments. In Proceedings of the IEEE International Conference on Robotics and Automation, Vilamoura, Portugal, 14–18 May 2012; pp. 957–964.
30. Loop, C.; Zhang, Z. Computing rectifying homographies for stereo vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, USA, 23–25 June 1999; pp. 120–131.
31. Fusiello, A.; Trucco, E.; Verri, A. A compact algorithm for rectification of stereo pairs. *Mach. Vis. Appl.* **2000**, *12*, 16–22.
32. Howard, A. Real-time stereo visual odometry for autonomous ground vehicles. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, 22–26 September 2008; pp. 3946–3952.
33. Fusiello, A.; Irsara, L. Quasi-euclidean uncalibrated epipolar rectification. In Proceedings of the 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
34. Kumar, S.; Micheloni, C.; Piciarelli, C.; Foresti, G.L. Stereo rectification of uncalibrated and heterogeneous images. *Pattern Recognit. Lett.* **2010**, *31*, 1445–1452.
35. Kang, S.B.; Webb, J.A.; Zitnick, C.L.; Kanade, T. A multibaseline stereo system with active illumination and real-time image acquisition. In Proceedings of the 5th International Conference on Computer Vision, Cambridge, MA, USA, 20–23 June 1995; pp. 88–93.
36. Gallup, D.; Frahm, J.-M.; Mordohai, P.; Pollefeys, M. Variable baseline/resolution stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 125–138.
37. Fanto, P.L. Automatic Positioning and Design of a Variable Baseline Stereo Boom. Master's Thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2012.
38. Bay, H.; Tuytelaars, T.; van Gool, L. SURF: Speeded up robust features. In Proceeding of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 404–417.
39. Juan, L.; Gwun, O. A comparison of SIFT, PCA-SIFT and SURF. *Int. J. Image Proc.* **2009**, *3*, 143–152.
40. Sanfeliu, A.; Andrade-Cetto, J.; Barbosa, M.; Bowden, R.; Capitán, J.; Corominas, A.; Gilbert, A.; Illingworth, J.; Merino, L.; Mirats, J.M.; *et al*. Decentralized Sensor Fusion for Ubiquitous Networking Robotics in Urban Areas. *Sensors* **2010**, *10*, 2274–2314.
41. Garrell, A.; Sanfeliu, A. Cooperative social robots to accompany groups of people. *Int. J. Robot. Res.* **2012**, *31*, 1675–1701.
42. Clemente, L.A.; Davison, A.J.; Reid, I.; Neira, J.; Tardós, J.D. Mapping large loops with a single hand-held camera. In Proceedings of Robotics: Science and Systems Conference, Atlanta, GA, USA, June 2007.

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).