Chapter 5

Recent Developments in Monocular SLAM within the HRI Framework

Edmundo Guerra, Yolanda Bolea,
Rodrigo Munguia and Antoni Grau

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/63820

Abstract

This chapter describes an approach to improve the feature initialization process in the delayed inverse-depth feature initialization monocular Simultaneous Localisation and Mapping (SLAM), using data provided by a robot's camera plus an additional monocular sensor deployed in the headwear of the human component in a human-robot collaborative exploratory team. The robot and the human deploy a set of sensors that once combined provides the data required to localize the secondary camera worn by the human. The approach and its implementation are described along with experimental results demonstrating its performance. A discussion on the usual sensors within the robotics field, especially in SLAM, provides background to the advantages and capabilities of the system implemented in this research.

Keywords: mapping and localization, sensors, visual odometry, HRI, features initialization

1. Introduction

A great deal of the investigation done in the field of robotics is addressed to the Simultaneous Localisation and Mapping (SLAM) problem [1, 2]. The SLAM problem is generally described as that of a robot—or robotic device with exteroceptive sensor/s—which explores an unknown environment, performing two different tasks at the same time: It builds a map with the observations obtained through the exteroceptive sensor/s [3] and localizes itself into the map during the exploration, thus knowing the position and trajectory.

The works defining the origin of the field can be traced to Smith and Cheeseman [4], Smith et al. [5], and Durrant-Whyte [6], which established how to describe the relationships between landmarks while accounting for the geometric uncertainty through statistical methods. These eventually led to the breakthrough represented in Smith's work. In such a research, the problem was presented for the first time as a combined problem with a joint state composed of the robot pose and the landmark estimations. These landmarks were considered correlated due to the common estimation error on the robot pose. That work would lead to several works and studies, being [7] the first work to popularize the structure and acronym of SLAM as known today.

The problem related with SLAM techniques is considered of capital importance given that a solution to it is required to allow an autonomous robot to be deployed in an unknown environment and operate without human assistance. But there is a growing field of robotics research that deals with the interaction of human and robotic devices [8]. Thus, there are several applications of robotic mapping and navigation that include the human as an actor. The basic application would be the exploration of an environment by a human, but mapped through a robotic platform [9]. Other works deal with more complex applications, such as mapping the trajectory of a group of humans and robots during the exploration of an environment and coordinating them with the help of radio frequency identification (RFID) tags [10]. Another application gaining weight is the use of SLAM to allow assistance robots to learn environments, improving the usability of the device [11].

All these approaches solve some kind of SLAM problem variant where the human factor is present: to assist, to track, to navigate, etc. But none uses data captured by human senses. There are works that deal with the mapping of human-produced data into map generated by a robot, but these data are not used in the map estimation process, but 'tagged' to it. So currently, no approach uses the data from human into the solution to the SLAM problem. This is a waste of useful resources, given the power of the human sight, still superior in terms of image processing to the most advanced techniques which are increasingly adopting the strategies discovered by scientists, but designed and adopted by human evolution millennia ago.

So, in this chapter, we will discuss about the monocular SLAM problem in the context of human-robot interaction (HRI), with comments on available sensors and technologies, and different SLAM techniques. To conclude the chapter, a SLAM methodology where a human is part of a virtual sensor is described. His/her exploration of the environment will provide data to be fused with that of a conventional monocular sensor. These fused data will be used to solve several challenges in a given delayed monocular SLAM framework [12, 13], employing the human as part of a sensor in a robot–human collaborative entity, as was first described in authors' previous work [14].

2. Sensors in the SLAM problem

In robotic systems, all relations between the system and the physical environment are performed through transducers. Transducers are the devices responsible for converting one

http://dx.doi.org/10.5772/63820

kind of energy into another. There are basically 2 broad types of transducers: sensors and actuators. Actuators use energy from the robotic system to produce physical effects, such as forces and displacements, sound, and lightning. Sensors are the transducers responsible for sensing and measuring by way of the energy conversion they perform: turning the energy received into signals (usually of electrical nature), which can be coded into useful information.

The sensors used in SLAM, just like in any other fields of robotics, can be classified according to several criteria. From a theoretical point of view, one of the most meaningful classifications is that if the sensor is of proprioceptive or exteroceptive nature. Proprioceptive (i.e., '*sense of self*') sensors are generally responsible for measuring values internal to the robot system, like the position of a joint, the remaining battery charge, or a given internal temperature. On the other side, exteroceptive sensors measure different characteristics and aspects of the environment, normally with respect to the sensor itself.

The encoders are proprioceptive sensors, responsible for measuring the position or movement of a given joint. Although there are linear encoders, only the rotary encoders are frequently used in the SLAM problem [15]. These encoders can measure directly the position of the rotary axis, in terms of position if they are 'absolute encoders' or in terms of movement for the 'incremental encoders'. Their great accuracy when measuring rotation allows computing the exact distance traveled by a wheel, assuming that its radius is known. Still they present several problems related to the nature of how they measure: The derived odometers assume that all the movement against the wheel surface is transformed into rotation at a constant and exact rate, which is false in many circumstances. This makes them vulnerable to irregular and dirty surfaces. As a proprioceptive sensor, with no exterior feedback, the error of a pure odometry-based SLAM approach will grow unbound, suffering the drift due to dead reckoning.

Range finders are exteroceptive sensors which measure distances between them and any point in the environment. They use a variety of active methods to measure distance, sending out sound, light, or radio waves and listening to the receiving waves. Generally, these are known as sonar, laser range finders (LRF), or radar systems. The devices destined to robotics applications generally perform scans, where a set of measurements is performed concurrently or over such a short time that they are considered all simultaneously. When scans are performed, each sub-measurement in a set is usually paired with bearing data, to note the relation between the different simultaneous measurements, generally performed in and arc.

Sonar systems use sound propagation through the medium to determine distances [16]. Active sonar creates a pulse of sound (a ping) and listens to its reflections (echoes). The time of the transmission of the pulse to its reception is measured and converted to distance by knowing the speed of sound being a time-of-flight measurement. Laser rangefinders (LRF) [17] can work on different principles, using time-of-flight measurements, interferometers, or the phase shift method. As the laser rays are generally more focused compared to the other types of waves, they tend to provide higher accuracy measurements. Radars [18] also employ electromagnetic waves, using time-of-flight measures, frequency modulation, and the phased array method

between others to produce the measurements. As they usually produce a repeated pulse at a given frequency (RPF), they present both a maximum and minimum range of operation.

These sensors can have great accuracy given enough time (the trade-off between data density and frequency is generally punishing), and as they capture the environment, they do not suffer from dead reckoning effects. On the other side, the data they provide are just a set of distance at given angles, so these data need to be interpreted and associated, requiring cloud matching methodology (like iterative closest point, ICP, and other similar and derived ones), which is computationally expensive. Besides, they have all their specific weaknesses: Sonar has limited usefulness outside of the water given how sound works on the air; LRF are vulnerable to ambient pollutants (dust, vapors) that may distort the lightning processes of the measurement; radar has very good range but tends to be lacking in accuracy compared to the other range-finders.

The Global Positioning System (GPS) [19] is a proprioceptive sensor based on synchronizing radio signal received from multiple satellites. With that information, it can compute the coordinates and height position of the sensor on any point of the world with up to 10 m margin. This 10 m margin grows rapidly if fewer satellites are visible (direct line of sight is required), making it useless on closed environments, urban canyons, etc. Besides, the weakness to satellite occlusion and wide error margin, the GPS presents other challenges, like a rather slow update rate for most of the commercial solutions.

The inertial measurement unit (IMU) is a proprioceptive sensor that combines several sensing components to produce estimations of the linear and angular velocities and the forces of the device. They have generally linear and angular accelerometers, and sometimes they include also gyroscopes and magnetometers, producing the sensory part of inertial navigation system (INS). The INS includes a computing system to estimate the pose and velocities without external references. The systems derived from the IMU have generally a good accuracy, but they are vulnerable to drift when used in dead reckoning strategies due to their own biases. The introduction on external reference can improve the accuracy, and thus, they are frequently combined with GPS. Introducing other external references leads to the development of the inertia-visual odometry field, which is closely related to the SLAM [20, 21]. Still, the accuracy gain is limited by the nature of the exteroceptive sensor added (which keeps its own weaknesses), and the IMU part of the system becomes unreliable in the presence of strong electromagnetic fields.

Vision-based sensors are exteroceptive sensors which measure the environment through the reflection of light on it, capturing a set of rays conformed as a matrix, thus producing images. The most common visual sensor is the camera, which captures images of the environment observed in a direction, similarly to the human eye. Still, there are many types of cameras, depending on the technology which they are based, which light spectrum they capture, how they convert measurement into information, etc. An standard camera can generally provide color or grayscale information as an output at 25 frames per second (fps) or more, being generally focused on the wavelength range visible by the human eye, and presenting that information in a way pleasant to the human eye. But specific cameras can work with different

http://dx.doi.org/10.5772/63820

frameworks as target, thus capturing other spectra not seen by human eye (IR, UV…), producing vastly higher fps rates, etc.

One of the main weaknesses of cameras within the context of the SLAM problem is that they produce only visual angular data: Each element of the matrix which composes an image shows the visual appearance information about a projected point where a ray (which theoretically can reach the infinite) finds an object. Thus, cameras alone cannot produce depth estimation in a given instant. This can be solved by more specific sensors, like time-of-flight cameras. These sensors generally have poorer resolutions, frame rate, dynamic range, and performance overall, while being several orders of magnitude more expensive, which made them barely used until few years ago.

There are other types of visual sensors that while still being cameras, they result more divergent from the standard monocular cameras. A good example would be the works on multiple camera stereo vision. Stereo cameras generally include two or more cameras and based on epipolar geometry can find the depth of the elements on the environment. Omnidirectional cameras expand the field of view, so that they can see almost all their surroundings at any given time, vision, presenting several challenges of their own in terms of image mapping and representation.

3. Classic only-bearing monocular SLAM approaches

There are many approaches to solve the SLAM problem, depending on the sensors available and the mathematical models and procedures used. From particle filters [22] to sums of Gaussian distributions, passing through the use of graph-based approaches [23] and RANSAC methods [24], the SLAM problem has been treated using many different mathematical techniques. Latest trends rely on bundle adjustment and other optimization methods [25, 26]. Still, one of the most commonly found approaches to the problem is using the extended Kalman filter (EKF) [27, 28], treating it as an incremental estimation problem.

The general monocular EKF-SLAM procedure is based on detecting points of interest which can be detected and distinguished from other robustly, introduce them into the map representation which is being built inside the filter, and track them through the sequence of frames, estimating both their pose and the camera odometry. For each landmark, a patch of the image around it, describing its 'appearance', is stored and will be used to identify it, and the landmark itself is generally modeled through unified inverse depth parametrization [29], although other model exists [30].

The estimation process is based on probabilistic filtering, where an initial prediction step makes a prediction of the movement of the robot and so of the position of the camera. Data from any sensors can be used; although in pure monocular SLAM methodologies, due to lack of data, predicted motion is assumed to be described by Gaussian distributions [31]. Thus, a constant velocity movement model is used, with random impulses of angular and linear accelerations modeled as white noise. The prediction of the map is much simpler: As the

landmarks or points of interest in the map are assumed to be part of the environment, the hypothesis used is that they will remain static and so their position does not change.

After the prediction step, a conventional EKF-SLAM would produce an actual measurement from the sensors and compare them with the predicted measurements obtained through the direct observation model. This step requires solving the data association problem, which consists in matching the predicted measurements with the actual measurements from the sensors. Given the computational cost of extracting all the possible points of interest at each frame and matching them with those predicted points, and the issues produced by the uncertainty of the prediction given that it is based on random movements, an active search strategy is used to deal with the problems [32]. Under this strategy, the features in the map are predicted into the pixel space using the direct observation model, and for each pixel, a search is performed looking for the most similar point (according to the stored patch), using zero-normalized cross-correlation (ZNCC). Ideally, each feature predicted will be matched to a new pixel that is the same feature in the latest frame. This process can fail due to visual artifacts, the geometry of the environment, the presence of dynamic objects, and several other causes. So these pairs of points (namely, the feature predicted and the match found in the latest frame) are checked through a data association validation methodology, in our case the HOHCT [13].

Once the predicted landmarks and its associated pairing are found in image space, in pixel coordinates, the innovation of the Kalman filter or residual can be computed following the usual EKF methodology.

Although the iterative estimation of the map as an EKF is pretty straightforward, there is still a critical process which defines many characteristics of any give monocular SLAM approach: the feature initialization process. When using conventional point detector and descriptors, it is frequent that several dozens or even hundreds of points will appear in an image, and most of them will be ignored for the SLAM process—based on spatial distribution, position on image, etc.—but still, no depth information is available in an instant way. Thus, two main strategies exist to deal with this issue: Undelayed approaches try to 'guess' the value to initialize the depth, normally relying on heuristics, while delayed approaches track a feature over a time, until they have a good estimation of its depth and only then proceed to initialize it.

These two types of strategies define many characteristics of the SLAM procedures. As undelayed approaches try to use point features as landmarks just after have been seen, the points are quickly introduced into the filter, accepting many outliers that have to be validated later or rejected at the data association validation step [28, 31]. On the other side, delayed approaches track and estimate the points before using them, so the used landmarks are generally more stable and reliable with delayed initialization [33].

The delayed inverse-depth (DI-D) monocular SLAM is a delayed feature initialization technique [12, 13]. The delay between a landmark being observed for the first time and being initialized allows estimating the parallax achieved through the estimated odometry. This in turn enables obtaining depth estimations for the landmarks through triangulation.

http://dx.doi.org/10.5772/63820

4. Introducing the human component into monocular SLAM

The DI-D procedure, although it was shown to be a strong monocular EKF-SLAM methodology, still presents several features that reduces its usability and scalability, mainly the need for an initialization process using synthetic or known a priori landmarks. These known landmarks would help initially to produce the odometry estimation and thus are critical to solve the scale problem of the map.

Shifting the monocular SLAM problem from an isolated sensor fusion point of view to a component into a bigger human-robot collaborative effort allows considering new options. Given the features of current exploratory robots, it is worth noting that an exploratory team composed of robots and humans will outperform any robotic device. If the desired tasks increase in complexity (emergency situations, those required management and decision under high uncertainty), the advantage of a human-robot collaborative team increases dramatically. Assuming that the human wears a headwear device with several sensors, the SLAM capabilities of the robot can be improved (**Figure 1**). Thus, the camera deployed in the helmet will be used to obtain 'instantaneous parallax', thus achieving complete measurement when the human is looking at the same direction as the robot, in a stereo-like situation, as it was initially proposed and described in authors previous work [14].

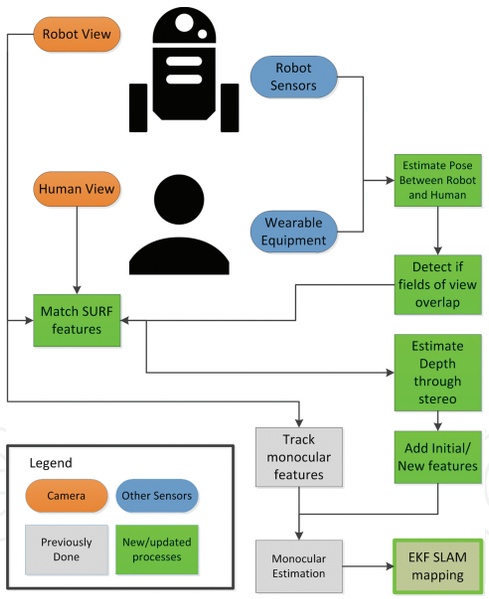

Figure 1. DI-D monocular EKF-SLAM components within human-robot collaborative.

To achieve this, in addition to the new camera on the human (C_h) which will perform the depth estimation with the robot camera (C_r), a combination of sensors and mechanisms able to estimate the pose between the cameras should be deployed. From a software point of view, the modules required to treat these data and estimate the pose, and those new which will deal with the new depth estimation process, must be implemented. From the EKF-SLAM methodology, the initialization of features is a local process that introduces the features into the EKF

using the inverse-depth parametrization, which remains the same, but will require also using a new inverse observation model that treats the whole system to estimate features as a single complex sensory process.

4.1. Multiple monocular vision sensor array: pseudo-stereo vision

The weak points discussed before can be solved within a cooperative context exploiting data from another monocular camera. Assuming that the C_h with known pose is near to the robotic camera performing SLAM (C_r), 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 [34, 35] 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; this allows optimizing 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 [35] and uncalibrated [36].

Figure 2. Image pair sample captured at one experimental sequence.

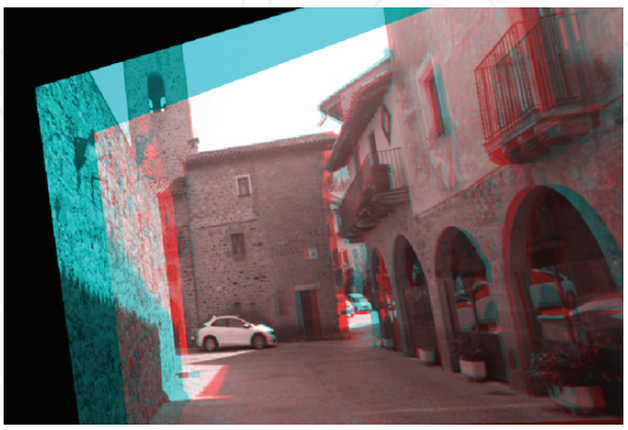

Figure 3. Rectification of images left and right at **Figure 2**. Scale distortions are produced due to the multiple reprojection operations.

Fusiello et al. [35] 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 rectifying stereo pairs assuming calibrated, uncalibrated, or even multiple view [37, 38] stereo rigs. These approaches need to warp both images according to the rectification found (see **Figure 2** left and right and **Figure 3**) and, in some cases, producing great variations in terms of orientation and scale (**Figure 3**), thus rendering them less attractive in terms of our approach.

At any case, dealing with stereo features without rectified images is not a big problem in the proposed approach. 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.

4.2. Scaled feature initialization with collaborative sensing

The requirement of metric scale initialization of the DI-D method can be avoided under the assumption of a cooperative framework. Classical DI-D required the presence of a set of known, easily identifiable features to estimate them initially through the PnP problem and initiate the EKF with scale. 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 pseudo-stereo estimation. This process is shown in **Figure 4**, where, after the pose between the robot camera and the human camera is 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 optical center of a pinhole 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 modeled as pyramids, as seen in **Figure 5**. So the intersection between the different polygons composing the pyramids is computed as a set of segments (two point tuples), as described by **Algorithm 1**. 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.

In the interest regions found, SURF-based feature descriptors [39] are matched to produce new stereo features to initialize in the EKF state vector when needed. SURF is chosen over SIFT and FAST [39] due to the more convenient trade-off offered in terms of matching accuracy and efficiency, and could be replaced by any other feature descriptor. Each pair of matched points between cameras allows estimating the world coordinates of the landmark feature seen through triangulation, back tracing the points on the images from the robot camera and the human camera. Then, the landmarks found and fully measured (with real depth estimation) are introduced in the monocular EKF according to the unified inverse depth parametrization.

To take advantage from the computational effort made during the non-overlapping frames, the landmarks that were being tracked to be initialized prior to the pseudo-stereo measurement are given priority to be introduced; these landmarks are robust because they were tracked for several frames previously.

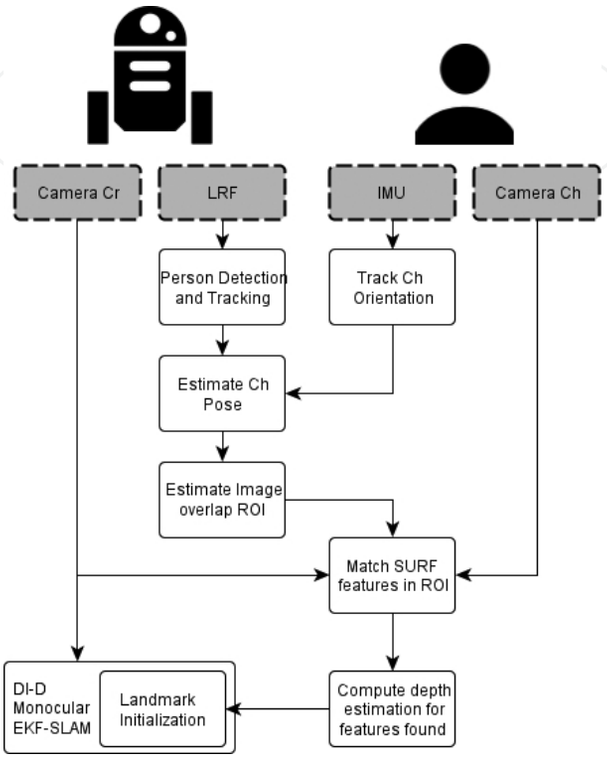

Figure 4. Block diagram of the final implementation, with sensors on gray boxes and software processes on clear blocks. The updated landmark initialization is one of the cornerstone processes in any feature-based EKF-SLAM technique.

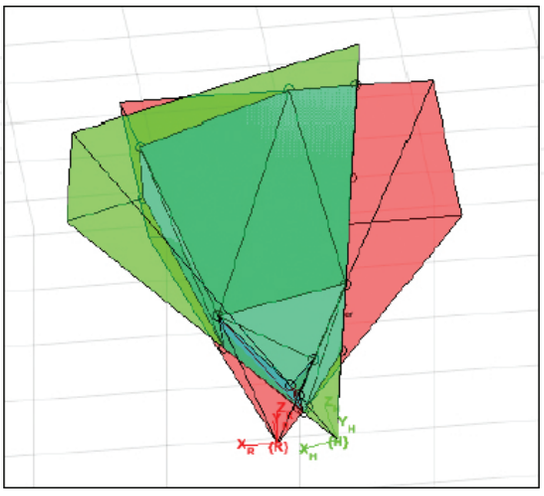

Figure 5. Graphical representation of the intersection of both cameras' fields of view.

ALGORITHM 1:

(ri_r, ri_h):= Find Stereo ROI (cam_r, cam_h, pl_{min})

begin

ri_r:= Ø; ri_h:= Ø

distance:= *FindDistance* (cam_r.pose, cam_h.pose)

PyramidDepth:= *FindMaxDepth* (distance, pl_{min})

Py1:= *ModelFoV*(cam_r,PyramidDepth)

Py2:= *ModelFoV*(cam_h,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

if ¬(intersection = Ø) **then**

  ri_r := *Envelope* (*ProjectTo2D* (cam_r.pose, intersection.points))

  ri_h := *Envelope* (*ProjectTo2D* (cam_h.pose, intersection.points))

end if

end

5. Experimentation and results

The approach described in this work was fully implemented and tested with real data. The DI-D SLAM with pseudo-stereo feature initialization was programmed in MATLAB® to test and evaluate it. Several sequences were captured in semi-structured environments using a robotic platform and wearable headgear.

5.1. Experimental system implementation

The sequences were reduced to a resolution of 720 × 480 pixels and grayscale color, shortening the computational effort for the image processing step. 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_h and C_r, 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 C_h. During the sequences, the camera C_r whose sequence would be used for the SLAM process was deployed looking forward, towards the advance direction. This kind of

movements produces singularities in estimation, as the visual axis of the cameras is aligned with the movement, producing 'tunnel blindness', where the elements near the centre of the captured images produce negligible parallax, and thus, only variations in scale are perceptible in short intervals.

The robot functions were performed by a robotic platform based on the Pioneer 3 AT (see **Figure 6**). The platform runs a ROS distribution over an Ubuntu 14.04 OS. The platform is equipped with a pair of laser range finders Leuze RS4-4 and a Logitech C920 webcam, able to work up to 30 frames per second (fps) at a resolution of 1080p. The sensors worn by the human are deployed on a helmet, including another Logitech webcam camera and an Xsens AHRS. All the data have been captured and synchronized through ROS in the robotic platform hardware. The ROS middleware provides the necessary tools to record and time-stamp the data from the sensors connected to the platform.

Figure 6. The robotic platform (based on the Pioneer 3 AT) used to capture the data sequences to test the described approach.

To estimate the pose of C_h, orientation data from the IMU are combined with the approximate pose of the human, estimated with the range finders [17]. 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.

The pose of the camera C_h with respect to the C_r is not assumed to be perfectly known. Instead, it is considered that a 'noisy' observation of the pose of C_h with respect to C_r 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_h was 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_h, for instance, using different sensors.

5.2. Experiments and results

The introduction of an auxiliary monocular sensor which can provide non-constant stereo information was proven useful. One of the weaknesses discussed earlier of the DI-D was the need to set an initial metric scale through synthetic feature, 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 it is not limited to the artificial landmark, the scale propagates in a smoother way with reduced drift on the local scale.

5.2.1. Visual odometry accuracy and scaling

Figure 7 shows results for two sample trajectories, with and without the utilization of the proposed non-constant stereo DI-D feature initialization approach, in blue and orange lines, respectively. The trajectory on **Figure 7** (left) was captured in an inner courtyard, with several seats and trees. This trajectory ran for 19 m, with two 90° turns, on a semi-structured environment with plenty of objects within view that could be mapped. On the other side, the trajectory shown in **Figure 7** (right) was capture as a straight 10 m movement in a blind alley with painted walls with homogenously textured surfaces, reducing the chances to obtain robust features. These two sequences contained plenty the most disadvantageous characteristics for monocular SLAM: singular movements where parallax cannot be observed for the central region of the camera, quick changes in orientation and turning velocities, surfaces/environments showing low count of robust visual features, natural lightning and shadows, etc.

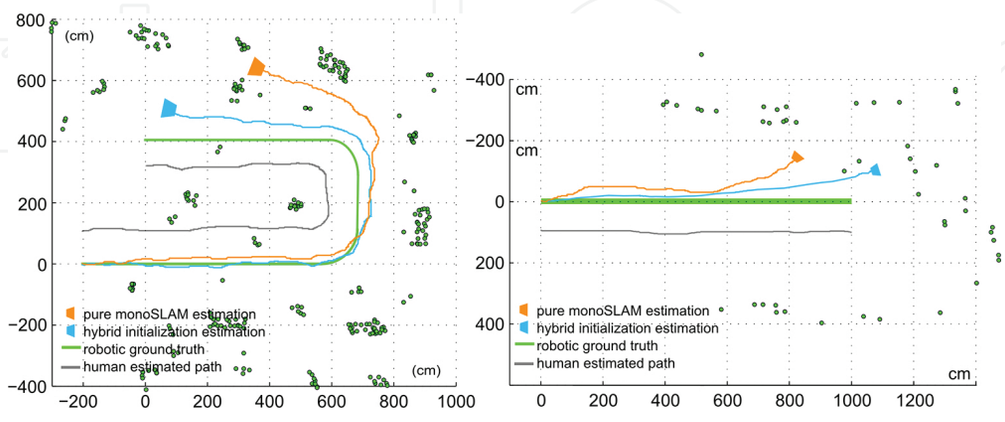

Figure 7. Trajectories estimated with classical DI-D monocular SLAM (orange plots) and with the new non-constant stereo DI-D approach for feature initialization (blue plots). Green line denotes robot ground truth; gray line denotes C_h ground truth.

The introduction of the pseudo-stereo initialization of features enables initialization of features with actual depth estimation instantly, without relying on heuristic or having a delay where data are being processed but not used in the estimation. Each of these situations has strong deterrents; for example, the heuristics used for depth initialization can vary between sequences, or even different SLAM 'runs' of the same video sequence, accounting for the uncertainty in the prediction model and the feature selection process. When a feature is initialized with a delayed method after it has been seen, computational power is spent on estimating a landmark that likely will not be used and never introduced into the EKF map.

At the end, the proposed approach made the system more resilient, especially to quick view changes, such as turning, and long singular movements—front advance. These movements can be seen in **Figure 7**, left and right, respectively. During close turns, delayed monocular SLAM approaches have very little time to initialize features because the environment changes quickly and the features are observed for short periods. This produces a decrease in the number of initialized features that decreases the odometry estimation accuracy. At the end of the run, the uncertainty becomes so big that errors cannot be corrected, the EKF loses convergences, and the estimation process results become useless. **Figure 7** (left) shows how the two turns can greatly degrade the orientation estimation for a classic delayed SLAM method, while the proposed approach can track the turns much more closely, with less than half the error. On the other side, **Figure 7** (right) illustrates the issues of singular movements: The odometry scale is very hard to estimate for pure monocular methods, because features present reduced parallax. Not only the length of the trajectory is affected by this phenomenon, but the accuracy of the orientation estimation also becomes compromised due to the inability of the EKF to reduce uncertainty quickly enough.

5.2.2. Computational costs analysis

The apparent increase in 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 computational costs as low as possible.

In the considered sequence set, there were a total of 9527 frames for C_r. Although C_h had a paired frame for each one on the C_r sequences, the overlap only was found in 3380 of the pairs. This means increasing the visual processing and feature capturing costs on a 35.47% of the frames. Increasing the computational cost of the most demanding step in a third of iterations may look daunting, but there are few considerations. The technique rarely implies processing an additional full frame: The region where the overlap is interesting is predicted and modeled as a ROI into the C_h image, limiting the area to explore. Besides, the cost increase is bounded by the number of frames where it is applied, so, if there are enough features visible in the map, there is no need to execute the pseudo-stereo depth estimation.

Moreover, it is worth noting that newly proposed approach made less effort per feature to initialize it, as it can 'instantly' estimate landmarks on the process of being initially measured through parallax accumulation. This trades off with the fact that the pseudo-stereo initialization can initialize with more frequency weak features which the delayed initialization would not been able to handle, and must be rejected during the data association validation step.

http://dx.doi.org/10.5772/63820

Table 1 shows the features initialized by each approach and the tracking effort required until the initialization of the features is done. Note how the non-constant stereo DI-D feature initialization approach uses about 8% more features, but the effort used to initialize them is much lower, as seen by the number of frames which the feature is tracked prior to being introduced into the map. 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_h processing and stereo-based initialization.

	Classic EKF monocular SLAM	**Pseudo-stereo feature initialization**
Features initialized (total)	1871	2032
Features per sequence (avg.)	267.28	290.26
Average frames to initialize a feature	19.63	8.53

Table 1. Statistics of initialized features and frames from first observation until initialization for DI-D SLAM and pseudo-stereo initialization.

Furthermore, in real-time applications employing this technique, the C_r sensor could be upgraded to an 'intelligent' sensor, with processing capabilities, using off-the-shelf technologies—low-cost microcomputers, FPGA, etc. This approach would integrate image processing in the C_h sensor, allowing parallel processing of features, and sending only extracted features, reducing required bandwidth and transmission time. This processing step could be done while the robotic camera C_r 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 novel approach to monocular SLAM has been described, where the capabilities of additional hardware introduced in a human-robot collaborative context are exploited to deal with some of the hardest problems it presents. Results in quickly changing views and singular movements, the bane of most of the EKF-SLAM approaches, are greatly improved, proving the proposed approach.

A set of experiments on semi-structured scenarios, where a human wearing a custom robotic headwear explores the unknown environments with a robotic platform companion, were captured to validate the approach. The system proposed profits from the sensors carried out by the human to enhance the estimation process performed through monocular SLAM. As such, data from the human-carried sensors are fused during the measurement of the points of interest, or landmarks. To optimize the process and avoid unnecessary image processing, the usefulness of the images from the camera on the human is predicted with a geometrical model

which estimates if the human was looking at the same places that the robot, and limits the search regions in the different images.

During the tests using real data, the MATLAB implementation of the approach proved itself to be more reliable and robust than the other feature initialization approaches. Besides, the main weakness of the DI-D approach, the need of a calibration process, was removed, thus producing a locally reliable technique able to benefit from more general map extension and loop closing techniques. While the model to estimate the pose between cameras has a given uncertainty very difficult to reduce (accumulated through the kinematic chain of the model), the measurement uncertainty is still lower than that of the purely monocular measurements, even with the parallax-based (in the delayed DI-D case) approach.

To conclude, the system proves the validity of a novel paradigm in human-robot collaboration, where the human can become part of the sensory system of the robot, lending its capacities in very significant ways with low-effort actions like wearing a device. This paradigm can open up the possibility of improving the capabilities of robotics systems (where a human is present) at a faster pace than what purely technical development would allow.

Acknowledgements

This research has been funded by Science Spanish ministry project reference DPI2013-42458-P.

Author details

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

*Address all correspondence to: antoni.grau@upc.edu

1 Automatic Control Dept., Techncial Univ of Catalonia, Spain

2 Computer Science Dept., University of Guadalajara, Guadalajara, Mexico

References

[1] Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: part I. IEEE Robot Autom Mag. 2006;13(2):99–110.

[2] Bailey T, Durrant-Whyte H. Simultaneous localization and mapping (SLAM): part II. IEEE Robot Autom Mag. 2006;13(3):108–17.

[3] Thrun S, et al. Robotic mapping: a survey. Explor Artif Intell New Millenn. 2002;1:1–35.

[4] Smith RC, Cheeseman P. On the representation and estimation of spatial uncertainty. Int J Robot Res. 1986;5(4):56–68.

[5] Smith R, Self M, Cheeseman P. Estimating uncertain spatial relationships in robotics. In: IEEE International Conference on Robotics and Automation Proceedings. 1987. pp. 850–850.

[6] Durrant-Whyte HF. Uncertain geometry in robotics. IEEE J Robot Autom. 1988;4(1):23–31.

[7] Durrant-Whyte H, Rye D, Nebot E. Localization of Autonomous Guided Vehicles. In: Giralt G, Dr.-Ing GHP, editors. Robotics Research [Internet]. Springer London; 1996 [cited 2013 Jun 4]. pp. 613–25. Available from: http://link.springer.com/chapter/10.1007/978-1-4471-1021-7_69

[8] De Santis A, Siciliano B, De Luca A, Bicchi A. An atlas of physical human–robot interaction. Mech Mach Theory. 2008 Mar;43(3):253–70.

[9] Fallon MF, Johannsson H, Brookshire J, Teller S, Leonard JJ. Sensor fusion for flexible human-portable building-scale mapping. In: Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on [Internet]. 2012 [cited 2013 Jun 13]. pp. 4405–4412. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6385882

[10] Kleiner A, Prediger J, Nebel B. RFID Technology-based Exploration and SLAM for Search and Rescue. In: IEEE; 2006 [cited 2013 Jun 10]. p. 4054–9. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4059043

[11] Cheein FAA, Lopez N, Soria CM, di Sciascio FA, Pereira FL, Carelli R. SLAM algorithm applied to robotics assistance for navigation in unknown environments. J Neuroeng Rehabil. 2010;7(1):10.

[12] Munguía R, Grau A. Monocular SLAM for visual odometry: a full approach to the delayed inverse-depth feature initialization method. Math Probl Eng. 2012;2012:1–26.

[13] Guerra E, Munguia R, Bolea Y, Grau A. A highest order hypothesis compatibility test for monocular SLAM. Int J Adv Robot Syst. 2013, 10:311.

[14] Guerra E, Munguia R, Grau A. Monocular SLAM for autonomous robots with enhanced features initialization. Sensors. 2014;14(4):6317–37.

[15] Johannsson H, Kaess M, Fallon M, Leonard JJ. Temporally scalable visual SLAM using a reduced pose graph. In: Robotics and Automation (ICRA), 2013 IEEE International Conference on [Internet]. IEEE; 2013 [cited 2016 Feb 10]. pp. 54–61. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6630556

[16] Diosi A, Taylor G, Kleeman L. Interactive SLAM using laser and advanced sonar. In: Robotics and Automation, ICRA 2005 Proceedings of the IEEE International Conference

on [Internet]. 2005 [cited 2013 Jun 10]. pp. 1103–1108. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1570263

[17] Sanfeliu A, Andrade-Cetto J. Ubiquitous networking robotics in urban settings. In: Workshop on Network Robot Systems Toward Intelligent Robotic Systems Integrated with Environments Proceedings of [Internet]. 2006 [cited 2013 Dec 26]. pp. 10–13. doi: 10.1.1.75.2850&rep=rep1&type=pdf

[18] Checchin P, Gérossier F, Blanc C, Chapuis R, Trassoudaine L. Radar Scan Matching SLAM Using the Fourier-Mellin Transform. In: Howard A, Iagnemma K, Kelly A, editors. Field and Service Robotics [Internet]. Springer Berlin Heidelberg; 2010 [cited 2013 Jun 10]. pp. 151–61. (Springer Tracts in Advanced Robotics). doi: 10.1007/978-3-642-13408-1_14

[19] Joerger M, Pervan B. Autonomous ground vehicle navigation using integrated GPS and laser-scanner measurements. In: San Diego: Position, Location, and Navigation Symposium [Internet]. 2006 [cited 2013 Jun 10]. Available from: http://mmae.iit.edu/~gps/publications/open/mathieu%20plans%2006.pdf

[20] Li M, Mourikis AI. High-precision, consistent EKF-based visual–inertial odometry. Int J Robot Res. 2013;32(6):690–711.

[21] Lupton T, Sukkarieh S. Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions. IEEE Trans Robot. 2012;28(1):61–76.

[22] Montemerlo M, Thrun S, Koller D, Wegbreit B. FastSLAM: a factored solution to the simultaneous localization and mapping problem. In: Proceedings of the National conference on Artificial Intelligence [Internet]. 2002 [cited 2013 Jun 4]. pp. 593–598. Available from: http://www.aaai.org/Papers/AAAI/2002/AAAI02-089.pdf

[23] Folkesson J, Christensen H. Graphical SLAM—a self-correcting map. In: 2004 IEEE International Conference on Robotics and Automation, 2004 Proceedings ICRA'04. 2004. pp. 383–390, Vol.1.

[24] Civera J, Grasa OG, Davison AJ, Montiel JMM. 1-Point RANSAC for extended Kalman filtering: application to real-time structure from motion and visual odometry. J Field Robot. 2010;27(5):609–631.

[25] Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. In: Mixed and Augmented Reality, ISMAR 2007 6th IEEE and ACM International Symposium on [Internet]. 2007 [cited 2013 Dec 26]. pp. 225–234. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4538852

[26] Newcombe RA, Lovegrove SJ, Davison AJ. DTAM: Dense tracking and mapping in real-time. In: 2011 IEEE International Conference on Computer Vision (ICCV). 2011. pp. 2320–7.

[27] Abrate F, Bona B, Indri M. Experimental EKF-based SLAM for mini-rovers with IR sensors only. 3rd European Conference on Mobile Robots (ECMR 2007), Friburg, Germany, 2007.

[28] Grasa OG, Civera J, Montiel JMM. EKF monocular SLAM with relocalization for laparoscopic sequences. In: IEEE International Conference on Robotics and Automation (ICRA), [Internet]. 2011 [cited 2013 Jun 6]. pp. 4816–4821. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5980059

[29] Civera J, Davison AJ, Montiel JMM. Unified inverse depth parametrization for monocular slam. In: Proceedings of Robotics: Science and Systems. 2006.

[30] Sola J, Vidal-Calleja T, Civera J, Montiel JMM. Impact of landmark parametrization on monocular EKF-SLAM with points and lines. Int J Comput Vis. 2012;97(3):339–368.

[31] Davison AJ. Real-time simultaneous localisation and mapping with a single camera. In: IEEE International Conference on Computer Vision. 2003. p. 1403–1410.

[32] Davison AJ, Murray DW. Mobile robot localisation using active vision. Proc 5th European Conference on Computer Vision, Freiburg, Germany, 1998; Vol.2: 809–825.

[33] Munguia R, Grau A. Camera localization and mapping using delayed feature initialization and inverse depth parametrization. In: IEEE Conference on Emerging Technologies and Factory Automation, 2007 ETFA. 2007. pp. 981–8.

[34] Loop C, Zhang Z. Computing rectifying homographies for stereo vision. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 1999. p. 131, Vol. 1.

[35] Fusiello A, Trucco E, Verri A. A compact algorithm for rectification of stereo pairs. Mach Vis Appl. 2000;12(1):16–22.

[36] Fusiello A, Irsara L. Quasi-euclidean uncalibrated epipolar rectification. In: ICPR 2008 19th International Conference on Pattern Recognition, [Internet]. 2008 [cited 2013 Dec 26]. pp. 1–4. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4761561

[37] Kang SB, Webb JA, Zitnick CL, Kanade T. A multibaseline stereo system with active illumination and real-time image acquisition. In: Proceedings of the Fifth International Conference on Computer Vision, [Internet]. 1995 [cited 2013 Nov 18]. pp. 88–93. Available from: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=466802

[38] Fanto PL. Automatic Positioning and Design of a Variable Baseline Stereo Boom [Internet]. Virginia Polytechnic Institute and State University; 2012 [cited 2013 Nov 14]. Available from: http://scholar.lib.vt.edu/theses/available/etd-07252012-081926/

[39] Juan L, Gwun O. A comparison of sift, pca-sift and surf. Int J Image Process IJIP. 2009;3(4):143–152.

IntechOpen