Proceedings of the 2004 IEEE
International Conference on Robotics & Automation
New Orleans, LA • April 2004

V-GPS(SLAM): Vision-Based Inertial System for Mobile Robots

Darius Burschka and Gregory D. Hager
Computational Interaction and Robotics Laboratory
The Johns Hopkins University
Baltimore, MD 21218, USA
Email: {burschka|hager}@cs.jhu.edu

***Abstract*— In this paper we present a novel vision-based approach to Simultaneous Localization and Mapping (SLAM). We discuss it in the context of estimating the 6 DoF *pose* of a mobile robot from the perception of a monocular camera using a minimum set of three natural landmarks. In contrast to our previously presented V-GPS system, which navigates based on a set of known landmarks, the current approach allows to estimate the required information about the landmarks on-the-fly during the exploration of an unknown environment. The method is applicable to indoor and outdoor environments.**

The calculation is done from the image position of a set of natural landmarks that are tracked in a continuous video stream at frame-rate. An automatic hand-off process allows an update of the set to compensate for occlusions and decreasing reconstruction accuracies with the distance to an imaged landmark.

A generic sensor model allows a system configuration with a variety of physical sensors including: monocular perspective cameras, omni-directional cameras and laser range finders.

I. MOTIVATION

Localization is an essential task in a variety of applications that involve a camera moving in space. Examples of applications that rely on this kind of pose estimation spawn over a wide field from mobile navigation in unknown or partially known environments to estimation of camera movements in laparoscopic medical procedures where an endoscope camera is used for localization in the human body. All these tasks assume a capability of localization relative to a reference model of a given environment. Unfortunately, the model is often not known a-priori and needs to be estimated in parallel to the localization process. The problem of *simultaneous localization and mapping*, also known as SLAM, has attracted immense attraction especially in the mobile robotics literature. SLAM addresses the problem of building a map of an environment from a sequence of landmark measurements obtained from a moving system. Many popular SLAM implementations use laser range information as input to simplify the estimation process to a pure localization and registration since laser range finders estimate directly the 3D locations of the imaged points. We propose to extend this approach to a vision-based system where the information from a monocular camera is used as input.

The localization with monocular camera implicates additional challenge to the already difficult problem of the 3D model registration between consecutive camera frames. This challenge is the estimation of the 3D information itself from the two-dimensional projections in the two monocular camera images with uncertain motion between the two frames.

Since the motion of the robot is restricted to the plane of the floor, the registration of sensor images to each other is simplified to the estimation of the position within the plane of the floor and the orientation relative to a reference coordinate system. An information about the current orientation changes relative to a previous estimate can be calculated from the encoder information in the odometry on the robot [2] until a localization update becomes available again.

The situation changes drastically with the introduction of a different type of a robot designed to operate in off-road situations. Especially, the robot depicted in Fig. 1 introduces several new challenges emerging from the way the motion is generated. In addition, these robots are usually equipped with soft tires to reduce vibrations from rough road surfaces. These tires let them often bounce from the ground and tilt in an unpredictable way making any estimates based on odometry information useless.

Fig. 1. Example of an off-road capable robot requiring 6DoF pose estimation.

An inertial system (IMU) becomes necessary to deal with these sporadic changes induced by factors independent of the actual wheel rotations. An Inertial Measurement Unit (IMU) provides full six-degree-of-freedom (6DoF) motion sensing for applications such as navigation and control systems. Angular rate and acceleration are measured about three orthogonal axes. A drawback of this kind of systems is that they measure directly the velocities and accelerations. Therefore, they require an integration of these changes over time to obtain an absolute value of the position and orientation. We will refer to it in the following text as the *pose* of the mobile robot. This type of pose estimation is prone to errors due to drifts and offset errors in the measurements.

We address this problem by presenting a vision-based system for estimating the rapid changes in the pose by measuring directly the pose values instead of the motion parameters, thus, avoiding the sensitivity to integration errors of the conventional inertial systems. We propose a navigation system similar to a GPS system. It is based on a monocular camera.

This system operates on a set of landmarks in the world, similar to the satellites of the GPS system, whose positions are estimated in the *mapping step* as described in this paper. It is a generalization of our previously presented Vision-Based Control system [5], where an Image Jacobian Matrix was used to relate the deviation in the image to position errors in 3D space. An important extension to the previously presented system [4] is the ability to retrieve the necessary information about the tracked landmarks on-the-fly during an exploration mission in an unknown indoor or outdoor environment. In contrast to the previously presented approach the system does not require any odometry information.

Our system is motivated by the same idea as the system presented in [15], where a tracking approach for "2.5D space" was proposed. The system is supposed to compensate for the drawbacks of classical position-based visual servoing. In the approach presented in [15],eight landmarks are necessary to estimate the pose of an object in space. A reduction to four points is only possible in case that four co-planar points can be identified. The co-planarity constraint is a special case that is difficult to enforce in all situations. Additionally, a robust tracking of eight landmarks in the image is contradictory to our goal to build a compact system running on hardware with limited computational power that can usually be found on mobile systems. The smaller the number of landmarks that we need to track, the more processing power can be dedicated to other important tasks on the robot.

Our pose estimation is based on an image-based approach that compares the 2D projections of an *internal 3D model* between images. The *internal 3D model* is estimated up to scale due to the limitations in the perception of a monocular camera system (see Section II-C.1.a). In [7] a recursive model-based object pose estimation is presented that is based on orthographic projection of points onto camera image. This approach is limited to configurations that can be projected onto a planar image. In our case, we propose a pose estimation method allowing robust pose verification from 3 tracked landmarks that can be placed anywhere around the sensor. Our approach operates in image coordinates of the camera using a novel representation for the 3D model that does not require any knowledge about the three-dimensional position in the world to register the reconstructions to each other.

The paper is structured as follows. In section II, we describe the details of our algorithm that allows us to track the pose of the system in frame-rate of the used sensor system. We present the evaluation of the system performance in section III and conclude in section IV with an overall evaluation of the presented approach and a discussion of our future goals.

II. Approach

In our approach, we track the image position of natural landmarks and we use this information to calculate the pose changes of a mobile system. The presented system does not depend on any specific tracking algorithm, but it can be used with a variety of tracking algorithms relying on color, pattern or depth information encoded in the image. The tracking result of a landmark is reported as a position of a point in the image that usually represents the center of the tracked region.

The problem that we are addressing in this section is how to use a pose estimation algorithm that is designed to be used with 3D vectors [10] in a system that provides only the direction of the vectors without any information about their lengths.

We begin with a presentation of a generic sensor model that is used in our algorithm in section II-A followed by a description of the algorithm in section II-B. This algorithm is capable of updating the *3D information* about a set of landmarks by tracking them in consecutive images. An important step is the initialization of the system (section II-C) that retrieves the necessary information from the camera images without any odometry information to provide the necessary state information for the processing described in section II-B. This information is updated by the system using image data. We conclude this section with a description of a way how to add new natural landmarks relying only on image data (section II-D) and how to remove landmarks from the system that are not reliable anymore (section II-E). This step is what classifies this approach into the SLAM field, although, in our case we solve it using a structure-from-motion approach based on a minimum of 3 corresponding points between two images from a monocular camera.

A. Ideal Generic Sensor Model

We discussed already in [6] the properties of an *ideal generic sensor* for navigation purposes. Images taken from different sensor systems, like conventional perspective cameras, omnidirectional systems and laser range finders can be transformed into a unified representation of this ideal sensor that samples the world in spherical coordinates.

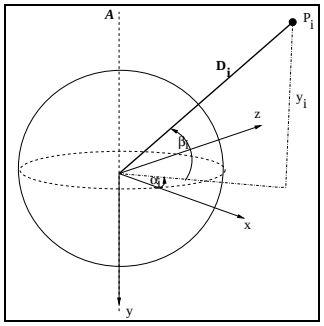

Fig. 2. Coordinate system of the ideal spherical sensor.

The position of any three-dimensional point P_i in space can be described as

$$P_i^*(\alpha_i, \beta_i, D_i) = \begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} = D_i \cdot \begin{pmatrix} \cos\beta_i \cos\alpha_i \\ \sin\beta_i \\ \cos\beta_i \sin\alpha_i \end{pmatrix} = D_i * \mathbf{n_i} \quad (1)$$

where (α_i, β_i) are the azimuth and elevation angles of the point P_i projected onto the spherical projection plane of the sensor and D_i is the distance to the point along the ray of

projection. Since the depth information D_i is not generally available from the most camera-based systems, the presented approach does not rely on this information, but this additional information simplifies the calculations significantly as we will show in Section II-C.1.

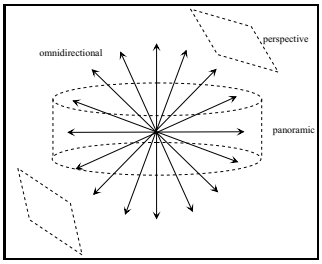

Fig. 3. The ideal sensor can be approximated with a variety of physical sensor configurations.

1) Approximation with a Laser Range System: A laser range system provides a complete set of the required information that includes the depth value D_i (1). The ideal sensor is actually identical to the sensor model of this sensor type. Unfortunately, most laser range finders scan the environment just in one elevation plane $(\beta_i = 0)$ limiting the angular changes in the vehicle pose just to rotations around the y-axis (Fig. 2).

2) Approximation with a Conventional Perspective Camera: A conventional perspective camera provides information about the projection angles (α_i, β_i) (1) that is encoded in the horizontal and vertical pixel-coordinates (u_i, ν_i) of a uni-focal camera (focal length f=1). The required angular values can be calculated using:

$$\begin{aligned} \alpha_i &= \arctan\, u_i \\ \beta_i &= \arctan \frac{\nu_i}{\sqrt{1+u_i^2}} \end{aligned} \quad \text{or} \quad \mathbf{n_i} = \frac{(u_i\ \nu_i\ 1)^T}{||(u_i\ \nu_i\ 1)^T||} \tag{2}$$

The image coordinates allow the recovery of the direction vector $\mathbf{n_i}$ in (1) leaving the length of the vector D_i unknown.

3) Approximation with an Omni-directional Camera: The field of view of a conventional perspective camera is usually limited to a cone with an opening angle of $50° - 80°$. This introduces a strong limitation on the distribution of the tracked landmarks, because the entire set needs to be visible in each image frame. This limitation can be loosened by using an omni-directional camera that can cover the entire hemisphere of the ideal sensor, extending the field of view significantly.

The required angular values can be calculated using

$$\begin{aligned} \alpha_i &= \operatorname{atan2}\left(\frac{\nu_i}{u_i}\right) \\ \beta_i &= \arcsin \frac{(1+b)\cdot\cos\gamma_i - 2\sqrt{b}}{2\sqrt{b}\cos\gamma_i - (1+b)} \end{aligned} \tag{3}$$

for a hyperbolic mirror with the shape $f(r) = \sqrt{\frac{b}{1-b}r^2 + b}$ as described in [4] in more detail. The value D_i from (1) remains unknown in this case as well.

B. Pose Estimation

The projection of a 3D-point P_i changes in the image of the ideal sensor due to motion of the mobile robot. The 6 DoF motion can be described with a rotation matrix $\tilde{\mathbf{R}}$ around the three-axes of the coordinate frame and a 3D translation vector $\mathbf{T}$ as

$$P_i^* = \tilde{\mathbf{R}} \cdot P_i + \mathbf{T} \tag{4}$$

According to (1), each point P_i can be expressed as a product of the direction vector $\mathbf{n_i}$ and the length of the vector D_i.

1) Calculation of the Pose Change from the Image Information: Let us assume for a moment that we have a guess for the values $P_{i\in\{1,\ldots,N\}} = D_i\mathbf{n_i}$ for the N tracked landmarks in the initial image.

We set the estimates for the points P_i^* in the current frame to

$$P_i^* = \lambda_i \cdot n_i^* = \tilde{R} \cdot P_i + \vec{T} \tag{5}$$

The predominant changes in the position are due to the strongly varying orientation angle of the camera, while the camera translation is small between the consecutive frames. Using this assumption, we set λ_i for a frame t in our system to

$$\hat{\lambda_i^t} = \begin{cases} D_i, & t = 0 \quad (\text{initial step}) \\ \lambda_i^{t-1}, & t > 0 \end{cases} \tag{6}$$

We have shown already in [4] that the matrix $\tilde{\mathbf{R}}$ and the vector $\mathbf{T}$ can be recovered by solving the following least-square problem for *N* landmarks

$$\min_{\tilde{\mathbf{R}},\mathbf{T}} \sum_{i=1}^{N} \|\tilde{\mathbf{R}}P_i + \mathbf{T} - P_i^*\|^2, \quad \text{subject to } \tilde{\mathbf{R}}^T\tilde{\mathbf{R}} = I. \tag{7}$$

Both are constant for all landmark transformations between two given images.

Such a constrained least squares problem [8] can be solved in closed form using quaternions [13], [16], or singular value decomposition (SVD) [12], [13], [16], [17].

The SVD solution proceeds as follows. Let P_i and P_i^* denote lists of corresponding vectors and define

$$\bar{P} = \frac{1}{n}\sum_{i=1}^{n} P_i, \quad \bar{P^*} = \frac{1}{n}\sum_{i=1}^{n} P_i^*, \tag{8}$$

that is, $\bar{P}$ and $\bar{P^*}$ are the centroids of $\{P_i\}$ and $\{P_i^*\}$, respectively. Define

$$P_i' = P_i - \bar{P}, \quad P_i'^* = P_i^* - \bar{P^*}, \tag{9}$$

and

$$\tilde{\mathbf{M}} = \sum_{i=1}^{n} P_i'^* P_i'^{\mathbf{T}}. \tag{10}$$

In other words, $\frac{1}{n}\tilde{\mathbf{M}}$ is the sample cross-covariance matrix between $\{P_i\}$ and $\{P_i^*\}$. Equation (10) allows to add confidence values as weights for the points as well, but for now we

assume these weights to be all equal to one. It can be shown that, if $\tilde{\mathbf{R}}^*$, $\mathbf{T}^*$ minimize (7), then they satisfy

$$\tilde{\mathbf{R}}^* = \text{argmax}_{\tilde{\mathbf{R}}} \text{tr}(\tilde{\mathbf{R}}^T \tilde{\mathbf{M}}) \quad (11)$$

$$\mathbf{T}^* = \bar{P}^* - \tilde{\mathbf{R}}^* \bar{P}. \quad (12)$$

Since we subtract the mean value $\bar{P}$ in (9), we remove any translation component leaving just a pure rotation.

Let $(\tilde{\mathbf{U}}, \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{V}})$ be a SVD of $\tilde{\mathbf{M}}$. Then the solution to (7) is

$$\tilde{\mathbf{R}}^* = \tilde{\mathbf{V}} \tilde{\mathbf{U}}^T \quad (13)$$

Note, that the optimal translation is entirely determined by the optimal rotation, and all information for finding the best rotation is contained in $\tilde{\mathbf{M}}$ as defined in (10). Hence, only the position of the 3D points relative to their centroids is relevant in the determination of the optimal rotation matrix.

These two values from (12) and (13) are used to calculate a new guess for P_i^* (5). These new better approximations are used to repeat the whole calculation. The iteration is terminated once the change in $|\Delta \tilde{\vec{T}}| < \epsilon_d$ is smalled than the required accuracy ϵ_d.

This approach requires a minimum of N=3 landmarks with increasing robustness to noise with additional landmarks. Our system tries to maintain a minimum of 4 landmarks to allow single landmarks to disappear due to noise or occlusions still maintaining an operational state of the system (see Sec. II-D).

C. Initialization of the System

There are different ways to initialize the system and the choice of one of the following three alternatives depends on the effort that can be spend to do the initialization. In unknown environments more complex initialization that does not require any preparation of the environment, like the one presented in Section II-C.1.a, may be preferable to simpler solutions requiring a reference pattern in the initial frame.

1) Calculation of the Point Representation - P_i: In Equation (1) the 3D point P_i is described by the values $(D_i, \mathbf{n_i})$.

A single camera image cannot provide any information about the depth value D_i. Additional information from a second image taken from a different location is required. What we are trying to solve here is a typical stereo problem in monocular stereo.

We solve the *correspondence problem* between the images through tracking of image regions using SSD tracking algorithms from our *XVision* image processing library [11] that allow to track gray-scale patterns (Fig 4) and, alternatively, color blob tracker to track unique color regions in the image of an omnidirectional camera.

The reconstruction problem can be solved in three ways that are described in the remaining part of this subsection. All these approaches are merely required for the initialization of the system. All new points P_i added later to the set of tracked landmarks are measured automatically using the current condition of the system as described in section II-D.

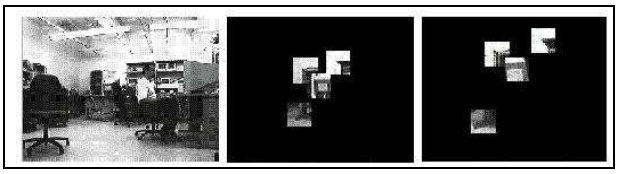

Fig. 4. Correspondences between consecutive image frames are maintained through tracking of image regions.

a) Point Estimation Based on Essential Matrix: A relation between the projections p_i, p_i^* in two camera images with known internal parameters can be expressed with the Essential Matrix $\tilde{\mathbf{E}}$ [8] as

$$p_i^* \tilde{\mathbf{E}} p_i = 0 \quad (14)$$

The Essential Matrix $\tilde{\mathbf{E}}$ consists of a product of two matrices

$$\tilde{\mathbf{E}} = \tilde{\mathbf{R}} \cdot \text{sk}(\mathbf{T}),$$

$$\text{with} \quad \text{sk}(\mathbf{T}) = \begin{pmatrix} 0 & -T_z & T_y \\ T_z & 0 & -T_x \\ -T_y & T_x & 0 \end{pmatrix} \quad (15)$$

Note that, given a correspondence, we can form a linear constraint on $\tilde{\mathbf{E}}$. It is only unique up to scale, therefore, we need 8 matches, then we can form a system of the form $\tilde{\mathbf{C}} \cdot e = 0$ where e is the vector of the 9 values in $\tilde{\mathbf{E}}$.

Using SVD, we can write $\tilde{\mathbf{C}} = \tilde{\mathbf{U}} \tilde{\mathbf{D}} \tilde{\mathbf{V}}^T$. The vector e is the column of $\tilde{\mathbf{V}}$ corresponding to the least singular value of $\tilde{\mathbf{C}}$, because of the rank deficiency of $\tilde{\mathbf{E}}$.

Since the constraint (15) defines the correspondence up to a scale, we remove the absolute value of $\mathbf{T}$ from $\tilde{E}$ using

$$\text{tr}(\tilde{\mathbf{E}}^T \tilde{\mathbf{E}}) = 2||\mathbf{T}||^2 \quad (16)$$

We normalize the matrix $\tilde{\mathbf{E}}$ with $\sqrt{\text{tr}(\tilde{\mathbf{E}}^T \tilde{\mathbf{E}})/2}$. This allows us to calculate the rotation matrix $\tilde{\mathbf{R}}$. We can solve for normalized $\mathbf{T}'$ from $\tilde{\mathbf{E}}'$ by calculating the position of the epipole [8] representing directly the translation vector.
We define

$$w_i = \mathbf{E}'_\mathbf{i} \times \mathbf{T}' \ \wedge \mathbf{R_i} = w_i + w_j x w_k \quad (17)$$

The three vectors $\mathbf{R_i}$ estimated from the permutations of 1,2,3 for the values of (i, j, k) construct the rotation matrix R.

This way we are able to calculate a modified form of (4) as

$$\frac{D_i^*}{m} \mathbf{n_i^*} = \tilde{\mathbf{R}} \cdot \frac{D_i}{m} \mathbf{n_i} + \frac{\mathbf{T}}{m} \quad (18)$$

with an unknown scaling factor *m*. From (9) we can see that it is sufficient to estimate the rotation matrix $\tilde{\mathbf{R}}$ with our approach, because the translation is removed from the matrix $\tilde{\mathbf{M}}$ through the differences to the mean values.

This method requires a tracking of 8 or more landmarks in the initial state to estimate the values for D_i from the homographies between the camera images. Once the D_i values are estimated, a set of N best landmarks according to [6] is chosen for further processing. The reader may ask, why we do not use this approach as a solution, but tracking of 8 or

more points is a far more complex problem than tracking of 3 points as required by our approach.

The initial vehicle pose is the reference for further processing. The orientation in the initial frame defines the orientation of the coordinate frame for following processing.

b) Point Estimation based on a Constrained Reference Configuration: As explained already before, the point estimation is only required in the initial state of the system. In most cases, a simple feature pattern on the floor with N=3 landmarks can be used to initialize the system without violating the requirement of being able to approach unknown areas. If the system was in a vertical position (no angle to the vertical Y-axis) in the initial frame then we can set the initial values for D_i to

$$D_i = \frac{1}{|n_{iy}|}. \tag{19}$$

This sets the Y components of all points P_i to 1. It corresponds to a vertical alignment of the camera pointing at the floor looking at a set of points on the floor. We know from (18) and (9) that in this way we will still obtain the correct rotation matrix $\tilde{\mathbf{R}}$, which is the only value important for an inertial system.

We can extend the system to a true *vision-based odometry* if we know the mounting height H of the camera system. In this case we modify (19) to

$$D_i = \frac{1}{|n_{iy}|} H. \tag{20}$$

In this case, we are able to estimate the true initial values for D_i, so that the computed translation vector $\mathbf{T}$ represents the metric translation in the world (m=1).

c) Point Estimation based on a Known Reference Configuration: Outdoors, it may be difficult to find a horizontal reference plane that could be used for calibration. It is obvious that following (20) we can use a known reference point configuration with known D_i values as an initial estimate.

D. Hand-off Procedure to add new Landmarks

The initial state is used as a reference orientation for consecutive measurements. Natural occlusions in the environment and decreasing spatial resolution with the distance to the sensor limit the usability of a tracked landmark to a local area segment. This requires an addition of new landmarks to the set if any of the landmarks disappear.

Since the system is supposed to add new information on-the-fly to the set while operating in an unknown environment, we developed a strategy to extract the necessary landmark information using the current state of the system. As mentioned already in Section II-B.1, the system requires a minimum of 3 landmarks to maintain the state information from the visual perception of the camera. Depending on the complexity of the scene, additional landmarks are tracked to maintain a minimum number of visible landmarks at 3 at all times.

In simple lab scenarios with small occlusion probabilities a system may try to keep the set of tracked landmarks at N=4. In Section II-C.1, we used three reference markers to bootstrap the system to the initial step. That means that the system will try to add an additional landmark to the set. The choice of the landmark candidates is done based on the considerations described in [6], [9]. The new landmarks are tracked in at least two consecutive images while the mobile robot is moving. In highly cluttered scenes the system may decide to move backwards after losing a landmark to ensure a visibility of the remaining landmarks during this process.

We have seen in Equation (18) that a landmark can be reconstructed only up to a scale without external reference about its size or odometry information. We require in our approach that the newly added landmark is scaled in the same way as the already existing set of landmarks that are currently tracked. The scale can be the actual metric distance to the landmark or any other scale depending on the initialization.

We decided to use the constancy of the difference vector between two landmarks to scale the new landmark to the right size. If point $P_1 = D_1\mathbf{n_1}$ is one point from the currently tracked set of landmarks and $P_x = D_x\mathbf{n_x}$ is the newly added landmark then the following equation (21) should hold for any two images, where the two points are visible.

$$P_1^* - P_2^* = D_1^* \cdot \mathbf{n_1^*} - D_x^* \cdot \mathbf{n_x^*} = \tilde{\mathbf{R}} \cdot D_1 \cdot \mathbf{n_1} + \mathbf{T}' - (\tilde{\mathbf{R}} \cdot D_x \cdot \mathbf{n_x} + \mathbf{T}') \tag{21}$$

with (D_x, D_x^*) being the unknown variables, we can write (21) as

$$\begin{pmatrix} \tilde{\mathbf{R}}\mathbf{n_x} & -\mathbf{n_x^*} \end{pmatrix} \begin{pmatrix} D_x \\ D_x^* \end{pmatrix} = \tilde{\mathbf{R}} D_1 \mathbf{n_1} - D_1^* \mathbf{n_1^*} \tag{22}$$

This measurement is used to estimate the pose of the system in the subsequent period of time. We achieve a better noise and error suppression by adding data from an additional image using at least three image frames for the estimate. The pose change from image $1 \rightarrow 2$ is annotated here as $(\tilde{\mathbf{R_1}}, \mathbf{T_1})$ and the pose change between images $2 \rightarrow 3$ are annotated as $(\tilde{\mathbf{R_2}}, \mathbf{T_2})$

$$\begin{pmatrix} \tilde{\mathbf{R_1}}\mathbf{n_x} & -\mathbf{n_x^*} & 0 \\ \tilde{\mathbf{R_2}}\tilde{\mathbf{R_1}}\mathbf{n_x} & 0 & \mathbf{n_x^{**}} \end{pmatrix} \begin{pmatrix} D_x \\ D_x^* \\ D_x^{**} \end{pmatrix} = \\ = \begin{pmatrix} \tilde{\mathbf{R_1}} D_1 \mathbf{n_1} - D_1^* \mathbf{n_1^*} \\ \tilde{\mathbf{R_2}}\tilde{\mathbf{R_1}} D_1 \mathbf{n_1} - D_1^{**} \mathbf{n_1^{**}} \end{pmatrix} \tag{23}$$

This over-determined system of linear equations in (23) can be solved using the least square approach with the pseudo-inverse matrix (24)

$$\tilde{\mathbf{A}}^{-1} = \left(\tilde{\mathbf{A}}^{\mathbf{T}} \tilde{\mathbf{A}} \right)^{-1} \tilde{\mathbf{A}}^{\mathbf{T}} \tag{24}$$

E. Removal of Unreliable Landmarks

An error in the estimation due to poor detection or motion in the scene (moving landmark) can be detected by the system

during operation by evaluating the smallest singular value of the matrix

$$\tilde{\mathbf{H}} = \begin{pmatrix} \tilde{\mathbf{R}}\mathbf{n_1} & -\tilde{\mathbf{R}}\mathbf{n_2} & 0 & -\mathbf{n_1^*} & \mathbf{n_2^*} & 0 \\ \tilde{\mathbf{R}}\mathbf{n_1} & 0 & -\tilde{\mathbf{R}}\mathbf{n_3} & -\mathbf{n_1^*} & 0 & \mathbf{n_3^*} \end{pmatrix} \quad (25)$$

If the system is still maintaining the correct information about the landmarks then the smallest singular value from H in (25) should be close to zero. This matrix is derived from (22) by solving it for the unknown vector lengths D_i.

$$\tilde{\mathbf{H}} \cdot (D_1\ D_2\ D_3\ D_1^*\ D_2^*\ D_3^*)^T = 0 \quad (26)$$

Let $(\tilde{\mathbf{U}}, \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{V}})$ be the SVD of $\tilde{\mathbf{H}}$ then the solution for the unknown values D_i is the column of $\tilde{\mathbf{V}}$ corresponding to the smallest singular value. This solution gives us just the ratios between the vector lengths in the tracking set.

This check can only be performed while the mobile robot is moving. If the position of the robot does not change then the vectors $(\mathbf{n_i^*}, \tilde{\mathbf{R}}\mathbf{n_i})$ are identical and the rank deficiency of $\tilde{\mathbf{H}}$ is greater than 1. A translation $||T|| > 0$ between the two images is necessary to modify both vectors $(\mathbf{n_i^*}, \mathbf{n_i})$ in a way that cannot be corrected by a pure rotation $\tilde{\mathbf{R}}$.

Periodic checks of different triplets in the set of tracked landmarks allow also to identify landmarks that cannot be tracked with a sufficient accuracy any more and need to be replaced by a new selection (see Section II-D).

III. Results

A. Convergence of the Pose Estimation

A main problem with the IMU systems is that they report changes between consecutive measurements and not the absolute values of the pose. The absolute pose is estimated by integrating the measurements over time. Small errors and offsets in the measurements result in an accumulative error in the final absolute estimates.

The presented system can calculate both values: the incremental change between two frames and the absolute difference to a reference frame. Depending on the requirements in the system, both modes are of interest. An important question here is the accuracy of the system for varying distances from the original configuration.

In Figures 5 and 6, we show the number of iterations required to estimate the motion parameters with an accuracy below 1cm. The proof of global convergence is mathematically derived in [10]. The number of iterations to find the best transformation explaining the changes between the reference and the current positions of the projections varies depending on the initial differences between the reference model and the current pose.

We observe in Figure 6 that a rotation of the system occasionally helps to find the "best fit" between the reference model and the current position. The number of iterations increases significantly for the positions close to the selected landmarks.

In the above examples we did not propagate the changes in the λ_i lengths between the steps. We show in Fig. 7 the improvement in the convergence using (6).

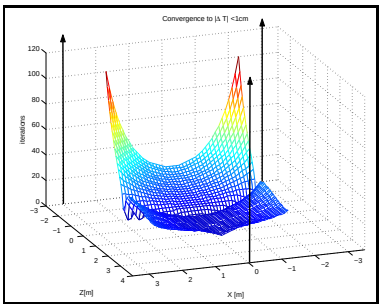

Fig. 5. Convergence to residual error $|\Delta T| < 1cm$ under $(0°, 0°, 0°)$ rotation (pure translation), arrows show the positions of the markers on the fbor.

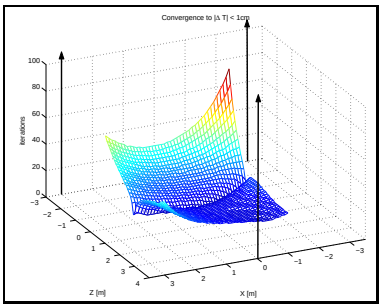

Fig. 6. Convergence to residual error $|\Delta T| < 1cm$ under $(20°, 30°, 50°)$ rotation in the same scenario as in Fig. 5.

From our analytical evaluation of optimal landmark distribution [6] we follow that the best placement of landmarks is around the robot. The currently used strategy is to maintain a set of landmarks, where the polygon of lines connecting the two closest neighbors contains the mobile robot inside of the polygon. New landmarks are selected accordingly.

B. Sensitivity to Calibration Errors

In case of a standard perspective camera with square pixels $(p_x = p_y)$ pointing down at the floor we can write the following equations for (α_i, β_i)

$$\begin{aligned} \alpha_i &= \operatorname{atan2}(\frac{\frac{u_x p_x}{f}}{\frac{\nu_y p_y}{f}}) = \operatorname{atan2}(\frac{u_x}{\nu_y}) \\ \beta_i &= \arctan\sqrt{(\frac{u_x p_x}{f})^2 + (\frac{\nu_y p_x}{f})^2} \end{aligned} \quad (27)$$

with (u_x, ν_y) being the pixel coordinates of the tracked points.

An error in the estimate of the focal length results in a non-linear distortion of the β_i-estimate.

Due to the way we estimate the motion parameters $\tilde{\mathbf{R}}, \mathbf{T}$, the rotation is estimated with a high accuracy as our evaluation in Fig. 8 shows.

The transformation to ideal sensor representation (Section II-A) requires the knowledge of internal parameters. Errors in the estimates have an influence on the accuracy of the angular representations as shown in Figure 8 for an error in the estimate of the focal length to f'=1.05*f. The internal calibration is important to ensure the correct transformation from the physical sensor to the ideal sensor.

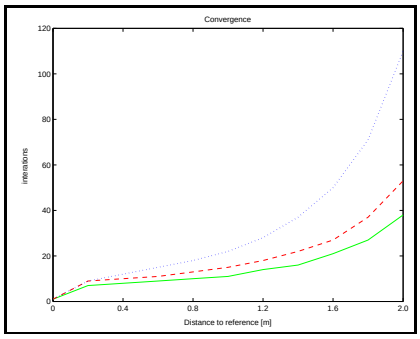

Fig. 7. Convergence along the diagonal path in Fig. 5 from (0,0) to (2,2): (dotted) using the estimated lengths from reference point (data from Fig. 5);(dashed) using previous estimate moving 20cm between images; (solid) using previous estimate moving 10cm between images.

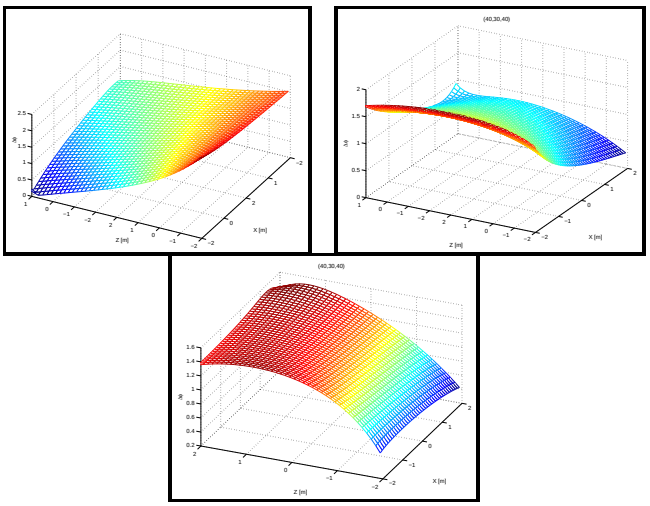

Fig. 8. Rotational error $(\Delta\varphi_x, \Delta\varphi_y, \Delta\varphi_z)$ of the estimate for f'=1.05*f after constant 50 iterations for λ_i for pose rotation $(40°, 30°, 40°)$.

IV. Conclusions and Future Work

We have presented a navigation system capable of operation under significant pose variations between the reference and the actual position. Especially, its capability to operate robustly under large rotational errors distinguish it from pure Image Jacobian approaches. Due to the linearizations in the Jacobian matrix, approaches based on Image Jacobian matrices have a limited area of convergence. Our approach avoids this problem by switching to a recursive model-adaptation method operating in the 2D image space. This method fits a scaled 3D model into the camera projections taken from two different positions estimating the six dimensional rotational and translational parameters during this process.

Our main contribution is a reduction of the required set of landmarks to three tracked points in the scene allowing a significant reduction in the requirement on the computational power on the mobile system. The presented system can be used with a variety of physical sensor configurations, like standard perspective cameras, omni-directional cameras and laser range finders. An image-based method for landmark initialization during hand-off process based solely on the image information is presented.

The correspondence problem between the saved reference representing the reference pose and the current image is solved by tracking the landmarks in consecutive images over time. The system is currently successfully used for camera localization in endoscopic procedures in human skull showing its generality. The resulting 3D model is used to register the reconstructed 3D features from the endoscope to an anatomic CT-scan of the patient.

Our future work will focus on development of better landmark selection strategies that will allow a more robust selection during the hand-off process and extensive tests with different physical sensor configurations.

Acknowledgments

This work was supported by the DARPA MARS project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

[1] S. Baker and S.K. Nayar. A Theory of Catadioptric Image Formation. *Proc. of IEEE International Conference on Computer Vision*, Bombay, 1998.

[2] L. Ojeda, H. Chung, and J. Borenstein. Precision-calibration of Fiber-optics Gyroscopes for Mobile Robot Navigation. *Proc. of IEEE International Conference on Robotics and Automation*, pages 2064–2069, San Francisco, April 2000.

[3] A.M. Bruckstein and T.J. Richardson. On Hyperbolic Mirrors for Omnidirectionality. *ATT Bell Laboratories Lab Notes*, February 1996.

[4] Darius Burschka and Gregory D. Hager. V-GPS – Image-Based Control for 3D Guidance Systems. In *Proc. of IROS*, pages 1789–1795, October 2003.

[5] D. Burschka and G. Hager. Vision-based control of mobile robots. In *Proc. International Conference on Robotics and Automation*, pages 1707–1713, 2001.

[6] Darius Burschka, Jeremy Geiman, and Gregory D. Hager. Optimal Landmark Confi guration for Vision-Based Control of Mobile Robots. In *Proc. of International Conference on Robotics and Automation (ICRA)*, 2003. To appear.

[7] Daniel F. DeMenthon and Larry S. Davis. Model-Based Object Pose in 25 Lines of Code. *International Journal of Computer Vision*, 15:123–141, June 1995.

[8] O. Faugeras, *Three-Dimensional Computer Vision*, The MIT Press, 1993.

[9] G. Hager, D. Kriegman, E. Yeh, and C. Rasmussen. Image-based prediction of landmark features for mobile robot navigation. In *IEEE Conf. on Robotics and Automation*, pages 1040–1046, 1997.

[10] G. Hager, C-P. Lu, and E. Mjolsness. Object pose from video images. *PAMI*, 22(6):610–622, 2000.

[11] G. Hager and K. Toyama. The XVision System: A General-Purpose Substrate for Portable Real-Time Vision Applications. *Computer Vision and Image Understanding*, 69(1):23–37, 1995.

[12] B. K. P. Horn, H. M. Hilden, and S. Negahdaripour, "Closed-form solution of absolute orientation using orthonomal matrices," *J. Opt. Soc. Amer.*, vol. A-5, pp. 1127–1135, 198.

[13] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternion," *J. Opt. Soc. Amer.*, vol. A-4, pp. 629–642, 1987.

[14] B. K. P. Horn. *Robot Vision*. MIT Press, 1986.

[15] Ezio Malis, Franois Chaumette, and Sylvie Boudet. 2D 1/2 visual servoing. *IEEE Transactions on Robotics and Automation*, 15(2):238–250, April 1999.

[16] M. W. Walker, L. Shao, and R. A. Volz, "Estimating 3-D location parameters using dual number quaternions," *CVGIP: Image Understanding*, vol. 54, no. 3, pp. 358–367, 1991.

[17] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fi tting of two 3-D point sets," *IEEE Trans. Pat. Anal. Machine Intell.*, vol. 9, pp. 698–700, 1987.