# EKF Monocular SLAM 3D Modeling, Measuring and Augmented Reality from Endoscope Image Sequences

Oscar G. Grasa[1], J. Civera[1], A. Güemes[2], V. Muñoz[3], and J.M.M. Montiel[1]

[1] I3A. Universidad de Zaragoza, Spain {oscgg,josemari}@unizar.es,
[2] Hospital Clínico Universitario "Lozano Blesa", Universidad de Zaragoza.
[3] Dpto. Informática e Ingeniería de Sistemas, Universidad de Malaga, Spain. *

**Abstract.** In recent years monocular SLAM has produced algorithms for robust real-time 3D scene modeling and camera motion estimation which have been validated experimentally using low cost hand-held cameras and standard laptops. Our contribution is to extend monocular SLAM methods to deal with images coming from a hand-held standard monocular endoscope. With the endoscope image sequence as the only input to the algorithm, a sparse abdominal cavity 3D model –a 3D map– and the endoscope motion are computed in real-time.
A second contribution is to exploit the recovered sparse 3D map and the endoscope motion to: 1) produce real-time photorealistic 3D models that ease cavity visualization; 2) measure distances in 3D between two points of the cavity; and 3) support augmented reality (AR) annotations. All this information can provide useful support for surgery and diagnose based on endoscope sequences. The results are validated with real hand-held endoscope sequences of the abdominal cavity.

## 1 Introduction

SLAM (Simultaneous Localization and Mapping) is a classical problem in mobile robotics: let be a mobile sensor following an unknown trajectory in an unknown environment, the goal is to estimate, simultaneously, both the environment structure –a map of 3D points– and the sensor location with respect to that map. Recently, SLAM research has focused on monocular cameras as the unique sensorial input, giving origin to monocular SLAM methods. 30 Hz real-time systems estimating full 3D camera motions and maps of 3D points using commodity cameras and computers have been reported in [1–4].

Our contribution is to extend, and validate with real monocular endoscope image sequences, one of the leading edge monocular SLAM algorithms [4] for its

* The authors would like to thank Dr. M. A. Bielsa form the Hospital Clínico Universitario "Lozano Blesa" for the "Abdominal wall image sequence". We would also like to thank Jonathan Richard Shewchuk for allowing free use of his Delaunay's Triangulation software (http://www.cs.cmu.edu/~quake/triangle.html). This work has been partially supported by Spanish FIT-360005-2007-9, DPI2006-13578 and DPI2009-07130, and EU funded RAWSEEDS Grant FP6-IST-045144.

use in medical applications. The primary result is a sparse 3D map composed of salient points –features– of the observed cavity. The main assumptions are that the cavity is rigid and that the endoscope undergoes a non-pure rotational motion. These conditions are fulfilled by a number of medical applications, such as laparoscopic ventral hernia repairs. The proposed work expands and details the abstract recently published by the same authors, [5].

Monocular SLAM methods recover the map up to an unknown scale factor, implying that only relative distances can be measured. However, in practice, a known size tool can provide the unknown scale factor and hence real distances can be recovered. Given the probabilistic nature of the SLAM map, the distance estimates are accompanied by an error estimate. We show how relative distances, along with the corresponding error estimates, are computed in live real-time while exploring a cavity.

Besides, the map is used as the backbone for real-time photorealistic modeling to ease the 3D cavity visualization. The textured 3D model allows the synthesis of a panorama that expands the limited FOV (field of view) of the endoscope. Finally, since the camera motion with respect to the 3D map is known accurately and in real-time, AR annotations can be supported live in medical sequences.

A number of techniques have been developed for cavity 3D reconstruction from endoscope sequences. [6,7] perform reconstructions using stereo endoscopes. In both works, the stereo endoscope always points to the same cavity area or moving organ to obtain the 3D structure. Stereo endoscopes have been successfully used in visual SLAM as reported in [8], where the performance of different image features when medical images are considered is analyzed. This is the closest work to ours. However, we are able to deal with monocular images.

Regarding the usage of monocular computer vision to determine 3D models from endoscope sequences, a significant paper is [9], where cavity 3D structure is computed to align it with a cavity CT scan model. More recently, in [10], classical two view RANSAC+bundle adjustment is applied to mannequin images to determine the 3D structure. This work presents a constraint-based factorization 3D modeling method from endoscope sequences to produce a dense 3D reconstruction in near real-time. Compared with these works, we demonstrate experimentally that our monocular SLAM algorithm features robust real-time performance when processing real endoscope images.

Monocular SLAM has proven a right tool for providing scene anchor points and camera motion estimation at frame rate in [2,11]. In this work we show that AR can be supported when monocular SLAM is adapted to medical images.

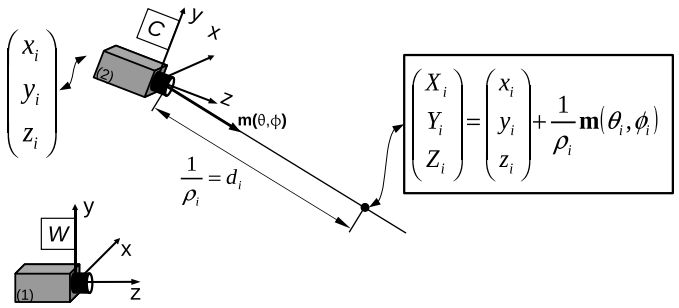
## 2 EKF Monocular SLAM

We focus on the EKF+ID+JCBB monocular SLAM approach. EKF (Extended Kalman Filter) monocular SLAM was initially proposed in [1]. ID (Inverse Depth) for map point coding, [4], improves measurement equation linearity and hence the overall estimation performance. JCBB (Joint Compatibility Branch

and Bound) [12] has proven essential for spurious rejection by enforcing scene rigidity. This combination was first used for monocular cameras in [13].

The estimated state is a Gaussian vector $\mathbf{x}$ that jointly codes the camera location with respect to the map, $\mathbf{x_v}$, and all map points coded in inverse depth, $\mathbf{y_i}$. The camera smooth motion prior is coded by means of a constant velocity motion model. Because of that, the camera location $\mathbf{x_v}$ includes: translation ($\mathbf{r^W}$), orientation defined by a quaternion ($\mathbf{q^{RW}}$), velocity ($\mathbf{v^W}$), and angular velocity ($\omega^{\mathbf{R}}$). The Gaussian estimate is defined by its mean, $\hat{\mathbf{x}} = \left(\hat{\mathbf{x}}_\mathbf{v}^\top \; \hat{\mathbf{y}}_\mathbf{1}^\top \; \hat{\mathbf{y}}_\mathbf{2}^\top \; \dots \right)^\top$, and covariance, $P$. Monocular sequence processing can recover 3D structure up to an unknown scale factor and up to an unknown 3D transformation with respect to an absolute reference. To remove the absolute transformation from the estimate, the first camera is defined as the absolute reference $W$ (see Fig. 1).

### 2.1 Inverse Depth




$$\begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} + \frac{1}{\rho_i}\mathbf{m}(\theta_i, \phi_i)$$

**Fig. 1.** Camera-(1) defines the world frame, $W$. A feature is observed for the first time by camera-(2), the feature location is defined with respect to the camera-(2) pose, $(x_i, y_i, z_i)^T$, using distance between camera-(2) and the feature, $d_i = 1/\rho_i$, and a directional vector, $\mathbf{m}(\theta, \phi)$, defined by its azimuth and elevation angles.

ID point coding [4] improves the measurement linearity –and hence the EKF performance– at low parallax, which takes place when feature depth is much bigger than the camera translation. At initialization, even features close to the camera produce low parallax. As a result, ID improves performance even for maps composed of close features only. An ID feature is a 6 parameters vector:

$$\mathbf{y_i} = \left(x_i \; y_i \; z_i \; \theta_i \; \phi_i \; \rho_i\right)^T \tag{1}$$

The projection ray of a map point when it is first observed is coded as: $x_i$, $y_i$, $z_i$ (camera location when the point was observed for the first time), and $\theta_i$ and $\phi_i$ (azimuth and elevation angles), which define the ray unit vector $\mathbf{m}(\theta_\mathbf{i}, \phi_\mathbf{i})$. Point

depth is coded by its inverse $\rho_i = 1/d_i$, so a point location $\mathbf{x_i}$ is (see Fig.1):

$$\mathbf{x_i} = \begin{pmatrix} X_i \\ Y_i \\ Z_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} + \frac{1}{\rho_i}\mathbf{m}(\theta_i, \phi_i), \quad \mathbf{m}(\theta_i, \phi_i) = \begin{pmatrix} \cos\phi_i \sin\theta_i \\ -\sin\phi_i \\ \cos\phi_i \cos\theta_i \end{pmatrix} \tag{2}$$

### 2.2 Data Association and Feature Detection

The quality of the SLAM reconstruction strongly depends on the data association accuracy. Active matching in monocular SLAM combines an accurate geometrical prediction with an image correlation score. EKF innovation defines for every map feature a predicted location and an elliptical uncertainty region. The match is searched for inside this area by correlation with a texture patch, which in our case is 11x11 pixels in size, stored when the map point is first observed and warped according to the point-camera relative location. As all innovations are highly correlated through the camera location uncertainty, JCBB is applied, checking if all matches are jointly compatible. If not, a Branch and Bound algorithm is applied to detect inconsistent matches before the EKF update.

Monocular SLAM in robotics uses a correlation score based on luminance, neglecting color information. However, internal organ images have high red and low blue content, so we use the green band, which contains a rich contrast to fire a point detector and a nice texture to produce distinctive patches for recognition.

To initialize a map point, a Harris saliency detector is applied. In endoscope scenes the light source is fixed to the camera producing reflections that erroneously fire the detector. To remove these reflections we assume they produce high gray level pixels, and if any pixel in a patch around the feature is over a threshold, the point is rejected. We use a theshold of 140 over 255.

## 3 SLAM Geometrical Map Exploitation

Monocular SLAM produces a sparse 3D map and the camera motion along with the corresponding covariance. We propose using the SLAM map as the geometrical backbone to support useful information for medical applications.

### 3.1 Photorealistic reconstruction

A mesh of triangular elastic textured tiles is built on top of the SLAM map. It is a generalization for 3D scenes of the mosaic method proposed in [14]. The tiles are defined by a standard 2D Delaunay's triangulation over a projection of the 3D map on the XY plane of the absolute reference $W$.3D triangle texture is gathered from the images that observe the complete corresponding 3D triangle. Fig. 2 sketches the photorealistic modeling process.

Since triangulation is a live process –map points, and consequently triangles, are continuously created, erased and their estimates changed–, maintenance operations are performed to deal with new and deleted triangles as the SLAM estimation evolves, and to take textures for the triangles from the images.

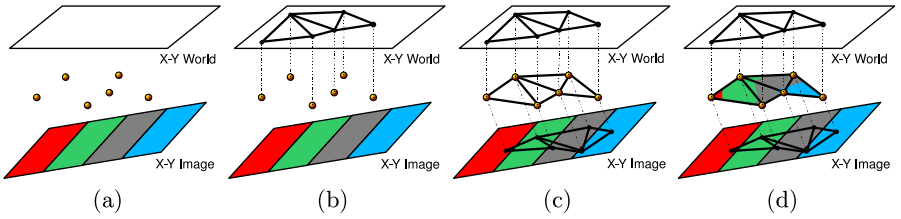



**Fig. 2.** *a*: Features in 3D, the current image and the world X-Y plane. *b*: Features projected onto a X-Y plane and triangulation on this plane. *c*: Triangulation image backprojection to obtain textures. *d*: Final photorealistic reconstruction.

### 3.2 Distance Measurement

Distance measurements are needed for some surgical procedures, e.g. ventral hernia repairs where the hole size is used to define the prothesic mesh dimensions.

Given the probabilistic 3D map, up to a scale factor, and an element of known size, for example a tool, real distances between points of the map, along with measurement error estimates, are available.

Considering two reference points, $(r_1, r_2)$, at a known distance, which define the scale factor, $s$, the distance between two map points $(i, j)$ is:

$$d(i,j) = s\frac{d_m(i,j)}{d_m(r_1,r_2)} \tag{3}$$

where $d_m(i,j)$ and $d_m(r_1,r_2)$ are the Euclidean distances between points $(i,j)$ and reference points $(r_1,r_2)$, respectively, measured in the SLAM map.

As the distance is a function of the SLAM state vector, $\mathbf{x}$, the covariance of the distance estimation can be propagated linearly from the SLAM covariance by means of the corresponding Jacobian matrix, $\mathbf{J}$:

$$\mathbf{J} = \frac{\partial d(i,j)}{\partial \mathbf{x}}, \mathbf{x} = \left(\mathbf{x_v}^\top \, \mathbf{y_1}^\top \dots \mathbf{y_{r1}}^\top \dots \mathbf{y_{r2}}^\top \dots \mathbf{y_i}^\top \dots \mathbf{y_j}^\top \dots\right)^\top \tag{4}$$

where all features are coded in ID. Since $d(i,j)$ only depends on $i, j, r_1$ and $r_2$, $\mathbf{J}$ is sparse, and reduced Jacobian ($\mathbf{J}_r$) and covariance ($\mathbf{P_r}$) matrices are used instead of the full matrices to compute the measurement error estimate $\sigma_d^2$:

$$\sigma_d^2 = \mathbf{J_r P_r J_r}^\top, \quad \mathbf{P_r} = \begin{pmatrix} P_{y_{r1}y_{r1}} & P_{y_{r1}y_{r2}} & P_{y_{r1}y_i} & P_{y_{r1}y_j} \\ P_{y_{r2}y_{r1}} & P_{y_{r2}y_{r2}} & P_{y_{r2}y_i} & P_{y_{r2}y_j} \\ P_{y_iy_{r1}} & P_{y_iy_{r2}} & P_{y_iy_i} & P_{y_iy_j} \\ P_{y_jy_{r1}} & P_{y_jy_{r2}} & P_{y_jy_i} & P_{y_jy_j} \end{pmatrix} \tag{5}$$

### 3.3 Augmented Reality

AR annotations in endoscope images need accurate real-time estimates of the live camera motion with respect to the observed scene. Monocular SLAM based only on images gathered by a camera has proven capable of providing camera

motion in real-time at 30 Hz, [2,11], for rigid scenes. AR is useful in laparoscopic surgery because it enables to visualize notations and fuse other modal images, such as 3D models of CT or MR, with endoscope images live during surgery.

Our contribution is to show that EKF monocular SLAM can be successfully applied to support AR annotations using as only input image sequences corresponding to a real hand-held endoscope observing the abdominal cavity.

## 4 Results

Experimental validation is performed over real images ($360 \times 288$) at 25 Hz gathered by a hand-held monocular endoscope observing the abdominal cavity, the image sequence being the only data input to the algorithm. Real-time performance at frame rate is achieved in all the experiments. Endoscope intrinsic parameters have been calibrated using a standard planar pattern calibration method, based on Zhang's initial solution [15], followed by Bundle Adjustment. A two parameter radial distortion model has been applied.

In Figs. 3 and 4 the textured triangular mesh model is shown. Despite the sparse map being composed of a reduced number of points, the photorealistic model provides easy understanding of the 3D cavity structure. Fig. 4 shows another example of a 3D live photorealistic model estimate corresponding to the abdominal wall, during a hernia repair surgery. The video "reconstruction.avi" shows the live model estimation and a detailed visualization of the 3D model from different points of view to assess the quality of the model.

Distances, together with their errors, have been calculated relative to a scale. The conversion to real distances is immediate if the real distance between two points in the reconstruction is known, as the scale can then be resolved (see Fig. 3). Fig. 3(b) shows the estimation history both for the distance and the error. Initially, error uncertainty is big, but as the camera translates, point location error decreases and consequently the distance error decreases as well. As the uncertainty is computed in live real-time, visual feedback gives the surgeon information on how to move the camera in order to reduce the distance error (see video "measurement.avi").

Since the 3D map and the camera location with respect to the map are available in live real-time, it is possible to anchor AR annotations to map points. Fig. 4 shows an AR cylinder, both in 3D and superimposed on the endoscope live image. As the virtual insertions are fixed to the map, they can be observed at their real location even when they are out of the camera FOV. The video "augmentedReality.avi" shows the corresponding movie.

## 5 Discussion and Future Work

We have shown how cavity exploration with a hand-held monocular endoscope can be casted as a monocular SLAM problem. A sparse map of 3D features and the camera motion are computed in live real-time at 25 Hz using the endoscope

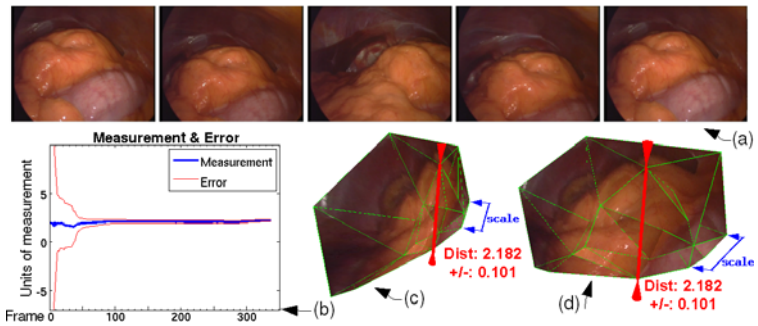



**Fig. 3.** Hand-held endoscope 341 frames sequence. *(a)* Several frames. *(c,d)* Photorealistic 3D model and distance estimate with the $2\sigma$, 95%, error interval. *(b)* Historical distance and error estimates. Notice the error reduction as the camera moves and gathers information from different points of view providing higher parallax. See also "measurement.avi" in the additional material.

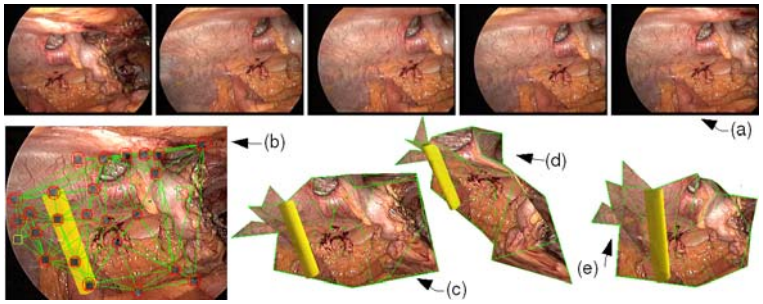



**Fig. 4.** Hand-held endoscope abdominal wall sequence. *(a)* Several frames. *(c,d,e)* Photorealistic 3D model and an augmented reality cylindrical insertion. *(b)* AR cylinder backprojected in live endoscope video. See also "reconstruction.avi" and "augmentedReality.avi" in the additional material.

image sequence as only input. The proposed algorithm is the state-of-the-art EKF+ID+JCBB monocular SLAM approach adapted to medical images.

It has also been shown how, building on the sparse SLAM map, a photorealistic model of the cavity can be computed in real time. The sparse 3D map provides support for AR annotations and 3D distance measurement. Our main contribution is to show this capabilities on real images gathered by a monocular endoscope observing the abdominal cavity.

Once we have tested the feasibility of the basic technique several venues of future work are open. As near future work, experiments to validate accuracy with respect to a ground truth in medical imagery would be valuable. Additionally, a cross-fertilization between engineers and physicians is needed to identify medical procedures that can benefit from the current state of the monocular SLAM technology; up to now we have focused on the abdominal cavity.

Current algorithms assume: 1) scene rigidity, 2) smooth endoscope motion, and 3) low motion clutter. These assumptions do not hold in general medical

scenes: non rigidity is almost prevalent, sudden motions are frequent, and tools cause a significant motion clutter. Thus, future work is being directed to cope with these issues. Relocation algorithms such as [16, 17] can provide robustness with respect to non smooth motion or motion clutter. Further research is also needed in models that code the scene as a deformable one.

## References


1. Davison, A.J.: Real-time simultaneous localisation and mapping with a single camera. In: International Conference on Computer Vision. (2003)
2. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: Symposium on Mixed and Augmented Reality (ISMAR). (2007)
3. Eade, E., Drummond, T.: Unified loop closing and recovery for real time monocular SLAM. In: British Machine Vision Conference. (2008)
4. Civera, J., Davison, A., Montiel, J.M.M.: Inverse depth parametrization for monocular SLAM. IEEE Trans. on Robotics and Automation **24**(5) (2008) 932–945
5. García, O., Civera, J., Güemes, A., Muñoz, V., J.M.M., M.: Real-time 3d modeling from endoscope image sequences. In: Workshop on Advanced Sensing and Sensor Integration in Medical Robotics (ICRA2009)
6. Stoyanov, D., Darzi, A., Yang, G.Z.: A practical approach towards accurate dense 3d depth recovery for robotic laparoscopic surgery. Computer Aided Surgery (2005) 199–208
7. Mourgues, F., Devernay, F., Coste-Manière, É.: 3D reconstruction of the operating field for image overlay in 3D endoscopic surgery. In: IEEE/ACM Symp. Augmented. Reality. (2001) 191–192
8. Mountney, P., Stoyanov, D., Davison, A., Yang, G.Z.: Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention. (2006) 347–354
9. Burschka, D., Li, M., Taylor, R., D., H.G., Ishii, M.: Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery. Medical Image Analysis **9**(5) (2005) 413–426
10. Wu, C., Sun, Y., Chang, C.: Three-dimensional modeling from endoscopic video using geometric constraints via feature poisitioning. IEEE Trans. on Biomedical engineering **54**(7) (2007)
11. Davison, A., Reid, I., Molton, N., Stasse, O.: Monoslam: Real-time single camera slam. IEEE Trans. on Pattern Analysis and Machine Intelligence **29**(6) (2007)
12. Neira, J., Tardos, J.D.: Data association in stochastic mapping using the joint compatibility test. IEEE Trans. on Robotics and Automation (2001) 890–897
13. Clemente, L.A., Davison, A.J., Reid, I.D., Neira, J., Tardós, J.D.: Mapping large loops with a single hand-held camera. In: Robotics Science and Systems. (2007)
14. Civera, J., Davison, A.J., Magallón, J.A., Montiel, J.M.M.: Drift-free real-time sequential mosaicing. Int. Journal of Computer Vision **51**(2) (2009) 128–137
15. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. on Pattern Analysis and Machine Intelligence **22**(11) (2000) 1330–1334
16. Williams, B., Klein, G., Reid, I.: Real-time SLAM relocalisation. In: Proc. International Conference on Computer Vision. (2007)
17. Cummins, M., Newman, P.: FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. The International Journal of Robotics Research **27**(6) (2008) 647–665