Reconstructing the Esophagus Surface from Endoscopic Image Sequences

Introduction

The esophageal cancer is a highly lateral disease affecting a significant percentage of population in United States and in Western World [2],[8]. Recent technological developments provide powerful tools such as white light endoscopy, narrow-band endoscopic imaging (NBI) and auto-fluorescence imaging (AFI), which allow for advanced visualization of the malignant tissue. Current gold standard protocol for screening and surveillance of the esophageal cancer, known as Gastrointestinal (GI) endoscopy, consists of visual examination of the esophagus surface under endoscopic guidance and acquisition of biopsies from endoscopically visible lesions. The first time screening is followed by surveillance endoscopies at regular intervals in order to track the evolution of the malignant tissue. This protocol requires identification and retargeting of the examined regions in the surveillance endoscopy for potential biopsy acquisition. However, a method for accurate localization and retargeting of the examined regions under endoscopic guidance is currently not available. Creating a 3D visualization of the esophageal surface can serve the endoscopist as a roadmap and provide a significant support for retargeting during the procedure.

In this work, we explore the possibility of creating a 3D patient-specific visualization of the esophagus surface from endoscopic images by performing a 3D reconstruction of the tissue surface. Reconstruction of the esophagus surface from endoscopic videos involves several challenges: the esophagus tissue presents big deformations which can greatly modify the appearance of the tissue; moreover, several factors affect the quality of the images, e.g. liquid inside the esophagus generates specularities and fast motion of the camera leads to blurred images.

In this paper, we investigate the feasibility of using Monocular Simultaneous Localization and Mapping (MonoSLAM) [1] for 3D reconstruction of the esophagus surface in the presence of these challenges. We identify the necessary modifications for its in-vivo application and demonstrate that the MonoSLAM is a promising framework for 3D visualization of esophageal tissue.

3D surface reconstruction is a well-known research area in computer vision community. In the medical field, there has been an increasing interest in 3D surface reconstruction from endoscopic image sequences [4],[5], [6],[7],[12],[13],[9]. Stoyanov *et al.* [9] present a method for depth recovery from stereo laparoscopic images of deformable soft-tissue. Mourgues *et al.* [5] propose a correlation-based stereo method for surface reconstruction and organ modeling from stereo endoscopic images. Quartucci *et al.* [7] apply a shape-from-shading technique to estimate the surface shape by recovering the depth information from the surface illumination. Zhou *et al.* [13] reconstruct the 3D structure using a Circular Generalized Cylinder (CGC) model which decomposes the reconstruction into a series of 3D circles forming a tube that models the esophagus.

Recently, a number of techniques have been published which apply feature-based techniques for 3D reconstruction in endoscopy [4],[11],[12]. Wang *et al.* [11] track Scale Invariant Features (SIFT) [3] in the endoscopic sequence and use Adaptive Scale Kernel Consensus (ASKC), for robust motion estimation. In [12], Wu *et al.* also track SIFT features and use an iterative factorization method for structure estimation. Mountney *et al.* [4] present a technique for building a 3D map of the scene for minimally invasive endoscopic surgery while recovering the camera movement based on SLAM from a stereo endoscope.

In this paper we propose a MonosSLAM based approach for 3D surface reconstruction from monocular endoscopic images. First, SIFT features [3] are detected in the first frame. Then, tracking of these features in consecutive frames is performed by an intensity based method using normalized sum-of-squared differences similarity measure [1],[6]. The MonoSLAM algorithm is applied on the tracked features in order to create a 3D map of the observed esophagus surface and localize the endoscope (6 DoFs) within this map simultaneously.

The rest of the paper is organized as follows: the proposed approach and our results are presented in the Method and Experimental Results**Error! Reference source not found.** Sections respectively. The current state and potential improvements for future work are discussed in the discussion Section.

Method

The goal of this paper is to present a method to estimate the esophagus surface from a monocular endoscopic image sequence, in order to provide a 3D visualization tool to the endoscopist. The 3D surface estimation from a monocular image sequence can be performed within the MonoSLAM framework. This approach aims to localize the camera in the 3D surface map, while simultaneously estimating the map.

The standard SLAM algorithm is designed to estimate a rigid scene and mostly applied to indoor environments. To be able to build a consistent map, giving the location of the features in the 3D space, these features have to be robustly detected and tracked. In the case of endoscopic image sequences the environment has practical issues which make the tracking difficult. First, the scene is deformable. Second, the images can be blurred as a consequence of water or fast movement. In addition, the features in an indoor environment are representative

patches which are generally the high frequency variation of the image, like edges. In contrast, the endoscopic images of soft-tissue mainly consist of homogeneous regions and lack the discriminative features. This makes the task of feature tracking more difficult.

Our method starts with detecting the SIFT features [3], which are distinctive local patches of the image. These features are tracked in the consecutive frames using a template-matching algorithm with the normalized sum-of-squared difference (NSSD) similarity measure [1],[6]. The NSSD is computed as:

$$NSSD(d_i, d_j) = \frac{\sum_{i,j}^{n^2}\left(\frac{I(i,j)-\bar{I}}{\sigma_I} - \frac{J(i+d_i, j+d_j)-\bar{J}}{\sigma_J}\right)^2}{n^2} \quad (1)$$

where $I(i,j)$ and $J(i,j)$ denote the intensity values at image location (i,j) for the initial template patch (from the first frame) and for the patch in the current frame, respectively. $\bar{I}$ and $\bar{J}$ are the mean intensity values of the patches I and J . σ_I and σ_J represent the standard deviations of intensities of the patches I and J , and n denotes the width and height of the patch. The displacement $\vec{d} = (d_i, d_j)$ with the maximum NSSD measure is used to compute the new location of the tracked patch in the current frame. In the feature tracking task, the local image patches are warped using the transformation matrix obtained from the current estimation of the camera position in the scene in order to correct the projective deformation produced by the different viewpoint angles induced by camera motion [1]. To select the most reliable patches for the MonosSLAM algorithm, we perform a thresholding on the SIFT features based on their confidence value. We also guarantee that the detected and tracked feature patches do not overlap.

One standard way to solve the feature-based SLAM problem is to use the Extended Kalman Filter (EKF). In this formulation the map is composed of a series of features (also known as landmarks). The problem consists of estimating the position of these landmarks in a global coordinate system while localizing the camera within the map. In the EKF-SLAM, the position and orientation of the features and the camera are represented using a state vector, denoted as x_k . The camera motion and the updates of the feature locations are estimated at each time frame using a dynamic model:

$$x_k = f(x_{k-1}, u_k) + w_k \quad (1)$$

$$z_k = h(x_k, y) + v_k \quad (2)$$

where z_k denotes the measurement of the position of the feature at time k , f is a non-linear transformation of the camera location determined as a function of its previous location and u_k is the input of the model, in this case is the estimated acceleration from an impulse force applied in each time-step modeled by a zero mean Gaussian distribution, h is the transformation of the feature position in the image as a function of the camera and feature position in the map, and w_k is the motion disturbance modeled as a zero-mean Gaussian distribution with covariance Q_k . Similarly, v_k is the observation error modeled as w_k but with covariance R_k . Formally, the estimated state vector $\hat{x}$ is composed of the camera state vector $\hat{x}_k$ and the feature state vector $\hat{y}_i, i \in \{0...N\}$, where N is the number of features. Furthermore, we associate a covariance matrix P to the state vector $\hat{x}$ (see equation 3). P reflects the uncertainty of the current estimate of the camera and map, but also represents the relation between the camera position and feature map.

$$\hat{x} = \begin{pmatrix} \hat{x}_k \\ \hat{y}_k \end{pmatrix}, \qquad P = \begin{bmatrix} P_{xx} & P_{xy} \\ P_{yx} & P_{yy} \end{bmatrix}, \quad (3)$$

where $\hat{x}_k$ is the camera state vector which is composed of the 3D position and orientation and $\hat{y}_i$ is the state vector of feature i which consists of the 3D position and the estimated orientation of the patch.

$$\hat{x}_k = \begin{pmatrix} r^W & q^{WC} & v^W & \omega^C \end{pmatrix}^T, \quad (4)$$

$$\hat{y} = \begin{pmatrix} r^{WY} & d^Y \end{pmatrix}^T \quad (5)$$

The equation 4 shows the components of the used camera state vector $\hat{x}_k$ where r^W is the 3D position vector, q^{WC} is the orientation quaternion. v^W and ω^C are the velocity vector and the angular velocity vector, relative to a fixed world frame W and camera frame C , respectively. We assume that the camera has constant acceleration and angular acceleration between time $k-1$ and k . The equation 5 shows the components of the feature state vector $\hat{y}$ where r^{WY} is the 3D position vector and d^Y is a unitary vector which describes the orientation of the feature.

The SLAM algorithm is implemented through two recursive steps which are composed of a prediction (time-update) and a correction (measurement-update). The time-update is computed using equations 6 and 7, and represents the update of the camera position in the feature map at time $k-1$.

$$\hat{x}_{k|k-1} = f(\hat{x}_{k-1|k-1}, u_k) \quad (6)$$

$$P_{xx,k|k-1} = \nabla f \cdot P_{xx,k-1|k-1} \cdot \nabla f^T + Q_k, \quad (7)$$

where ∇f is the Jacobian of f evaluated at $\hat{x}_{k-1|k-1}$.

The observation-update is computed using equation 8 and 9, and represents the update of the camera position and feature positions using the observed features in the image at time k .

$$\begin{bmatrix} \hat{x}_{k|k} \\ \hat{y}_k \end{bmatrix} = \begin{bmatrix} \hat{x}_{k|k-1} \\ \hat{y}_{k-1} \end{bmatrix} + W_k \left[z_k - h(\hat{x}_{k|k-1}, \hat{y}_{k-1}) \right] \quad (8)$$

$$P_{k|k} = P_{k|k-1} - W_k S_k W_k^T \quad (9)$$

$$S_k = \nabla h \cdot P_{k|k-1} \cdot \nabla h^T + R_k$$

$$W_k = P_{k|k-1} \cdot \nabla h^T \cdot S_k^{-1}$$

where ∇h is the Jacobian of h evaluated at $\hat{x}_{k|k-1}$ and $\hat{y}_{k-1}$. S_k is the innovation (or residual) covariance matrix and W_k is the optimal Kalman gain matrix.

The used observation model h in this approach is the pinhole projection and radial distortion of the camera in

order to incorporate the high distortion of the endoscope camera to the model. Refer to [1] for further details.

Experimental Results

The patient data for the experiments is acquired during an endoscopic examination using Olympus Exera II endoscope with NBI image enhancement. NBI provides an enhancement of the contrast in the mucosal pattern and submucosal vasculature [10]. Thus, it enables more structured visualization of the tissue allowing for more robust feature detection and tracking.

In our experiments, first, the internal camera and the distortion parameters of the endoscope are computed using the method presented in [1]. The radial distortions present in the NBI image sequence are then corrected using the estimated distortion parameters. Our proposed approach based on MonoSLAM is applied on the NBI patient data consisting of 31 consecutive frames.

To initialize the SLAM algorithm is necessary to provide some known features which are selected manually in order to define the real scale of the scene. We use two known features and estimate their position in the map by projecting the image point using a distance of 20 cm. from the camera to the scene.

An example frame from the NBI sequence is illustrated in Figure 1a. The example image with the tracked features is demonstrated in Figure 1b. The red boxes show the tracked features and the blue boxes the estimated features in the image. Figure 1c and 1d demonstrate the reconstructed surface map from the frontal and site view, respectively.

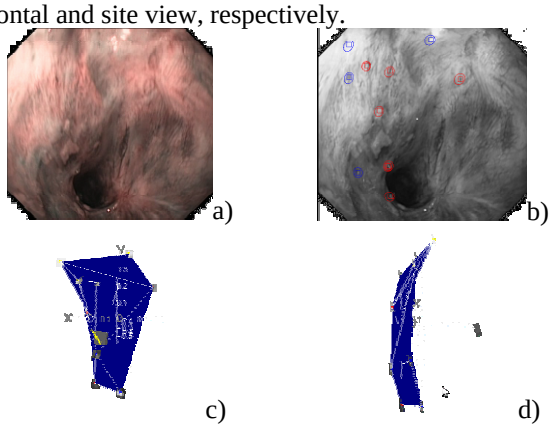

Fig. 1: Surface reconstruction results. An example image of a) the NBI patient data, b) with tracked features. Reconstructed surface from c) frontal and d) site view.

Discussion

In this work, we investigated the feasibility of MonoSLAM for 3D surface reconstruction from monocular NBI sequences. Our results demonstrate that MonoSLAM is a promising tool for surface reconstruction in NBI endoscopic images sequences. However, our study shows that there are some challenges for further improvement, which need to be addressed before MonoSLAM can be applied on endoscopic images *in-vivo* and *in-situ*. The tissue deformation present in the esophagus can greatly modify the appearance of the tissue causing a failure in the feature tracking. Moreover, due to the specular reflections and blurring caused by the esophageal surface or fast camera motion the image quality can be very poor. Addressing these challenges requires a more robust approach for the detecting and the tracking of the landmark features.

Literature

[1] Davison A. and Molton N.: Monoslam: Real-time single camera slam. IEEE Trans. Pattern Anal. Mach. Intell., 29(6):1052–1067, 2007. Member-Reid, Ian D. and Member-Stasse, Olivier.

[2] Fitzgerald R.C.: Review article: Barrett's oesophagus and associated adenocarcinoma UK perspective. Aliment Pharmacol Ther 20 (Suppl. 8): 45-49 (2004)

[3] Lowe D.: Distinctive image features from scale-invariant keypoints, Int. J. of Computer Vision, 60:91–110, 2004.

[4] Mountney P., Stoyanov D., Davison A.J., and Yang G.Y.: Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. MICCAI, 2006.

[5] Mourgues F., Devernay F., Coste-Manière E: 3D reconstruction of the operating field for image overlay in 3d-endoscopic surgery. In ISAR '01: Proceedings of the IEEE and ACM InternationalSymposium on Augmented Reality (ISAR'01), page 191, Washington, DC, USA, 2001. IEEE Computer Society.

[6] Pangercic D., Bogdan R. and Beetz M.: 3D-Based Monocular SLAM for Mobile Agents Navigating in Indoor Environments. In Proceedings of the 13th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Hamburg, Germany, September 15-18, 2008.

[7] Quartucci Forster C.H. and Tozzi C.L: Towards 3d reconstruction of endoscope images using shape from shading. Computer Graphics and Image Processing, Brazilian Symposium, 2000.

[8] Sharma P., Bansal A., Mathur S., Wani S., Cherian R., McGregor D., Higbee A., Hall S. and Weston A.: The utility of a novel narrow band imaging endoscopy system in patients with Barrett's esophagus, Gastrointestinal Endoscopy, 64, 167-175, 2006.

[9] Stoyanov D, Darzi, A., Yang, G-.Z. Dense 3d Depth Recovery for Soft Tissue Deformation During Robotically Assisted Laparoscopic Surgery. MICCAI, 41-48, 2004.

[10] Usui S., Satodate H., Fukami N., Kudo S., Yoshida T. and Inoue H.: Narrow-band imaging system with magnifying endoscopy for superficial esophageal lesions. Gastrointestinal Endoscopy, 59(2):288-295, 2004.

[11] Wang H.Z., Mirota D., Ishii M., and Hager G.D.: Robust motion estimation and structure recovery from endoscopic image sequences with an adaptive scale kernel consensus estimator. pages 1-7, 2008.

[12] Wu C., Sun Y., Chen Y., and Chang C.: Endoscopic feature tracking and scale-invariant estimation of soft-tissue structures. IEICE - Trans. Inf. Syst., E91-D(2):351-360, 2008.

[13] Zhou J., Das A., Li F., and Li B.: Circular generalized cylinder fitting for 3d reconstruction in endoscopic imaging based on mrf. 1-8, 2008.

Affiliation of the first Author

Victor Castañeda Zeman
Chair of Computer Aided Medical Procedures (CAMP) - Technische Universität München
Boltzmannstr. 3, Garching bei München, 85748, Germany
castaned@in.tum.de