Intraoperative on-the-fly Organ-Mosaicking for Laparoscopic Surgery

S. Bodenstedt, D. Reichard, S. Suwelack, M. Wagner, H. Kemgott, B. Müller-Stich, R. Dillmann, S. Speidel

Institute for Anthropomatics and Robotics, Department of Computer Science, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany;
Department of General, Abdominal and Transplantation Surgery, University of Heidelberg, Heidelberg, Germany

ABSTRACT

The goal of computer-assisted surgery is to provide the surgeon with guidance during an intervention using augmented reality (AR). To display preoperative data correctly, soft tissue deformations that occur during surgery have to be taken into consideration. Optical laparoscopic sensors, such as stereo endoscopes, can produce a 3D reconstruction of single stereo frames for registration. Due to the small field of view and the homogeneous structure of tissue, reconstructing just a single frame in general will not provide enough detail to register and update preoperative data due to ambiguities. In this paper, we propose and evaluate a system that combines multiple smaller reconstructions from different view points to segment and reconstruct a large model of an organ. By using GPU-based methods we achieve near real-time performance. We evaluated the system on an ex-vivo porcine liver (4.21 mm ± 0.63) and on two synthetic silicone livers (3.64 mm ± 0.31 and 1.89 mm ± 0.19) using three different methods for estimating the camera pose (no tracking, optical tracking and a combination).

Keywords: Endoscopic image processing, Stitching, SLAM, Visualization, Surgical vision, Quantitive endoscopy

1. INTRODUCTION

The amount of minimally-invasive surgery performed yearly is rapidly increasing. This is largely due to the numerous benefits these types of intervention have on the patient side: Shorter stay in hospital, less trauma, minimal scaring and lower chance of post-surgical complications. There are several drawbacks for the surgeon though, his hand-eye coordination is limited, he has no haptic feedback, no direct line of sight and a limited field of view.

Computer-assisted surgery tries to alleviate some of these drawbacks for the surgeon by providing him with information relevant to the intervention, e.g. as the position of a liver tumor or a preoperatively planned resection line. To correctly display such information, soft tissue deformations have to be taken into account. For registering preoperative data, the organ surface can be observed with optical laparoscopic sensors, such as the ones reviewed in, that provide a 3D-reconstruction of a single video frame. These methods in general only reconstruct a small field of view and due to the homogeneous structure of tissue, a single frame in general will not provide enough detail to rule out ambiguities during registration.

In this paper, we present a system that takes 3D reconstructions generated online by a stereo endoscope from multiple viewpoints, while segmenting out parts of an organ in each view. The reconstructions and the segmentations are then combined into one organ model. To compute a 3D point cloud from a stereo image pair, the algorithm outlined in is used. The segmentation of the organ of interest is done on the color image by labeling each pixel as part of an organ of interested or the background using a random forest. The resulting point clouds and their respective labels are then integrated into a voxel-volume using a Kinect-Fusion based algorithm.

Given enough viewpoints, the voxel-volume will contain a combined model more suited for registration than the model generated from just one single frame.

The procedure of localizing the camera in the world and simultaneously mapping it, know as simultaneous localization and mapping (SLAM) in literature, is a well-know problem in computer-assisted laparoscopic surgery.
surgery. There are many solutions that try to address it. Recently, the KinectFusion method providing dense reconstructions of the entire scene in real-time using a Microsoft Kinect was introduced by Newcombe et. al. in for usage in non-medical fields. In an extension of is used to reconstruct the surgical situs using multiple views taken by a 160 x 120 px time-of-flight camera.

The novelty of the approach presented in this work is that we use a modality already available in surgical workflow, a stereo endoscope with a 720 x 576 px resolution, to reconstruct an entire scene from multiple viewpoints online, while simultaneously segmenting one or more organs of interest. In the following, we will present a more detailed description of our reconstruction workflow, followed by an evaluation, using two synthetic livers and one ex-vivo porcine liver. Three methods for determining the camera pose are also evaluated: Optical tracking, no tracking and a combination of the two methods.

2. METHODS

Our system for reconstructing the scene consists of multiple steps (figure 1). We first reconstruct a 3D point cloud from stereo image frames. At the same time, the organs of interest are segmented in the video image. The reconstruction is then combined with the segmentation results and integrated into a truncated signed distance (TSD) volume. From this volume, after every time step, a mosaicked model of all the reconstructions combined with the segmentation can be retrieved. Using a TSD volume allows us to incorporate information from different viewpoints to create a larger model than from a single view, while simultaneously reducing noise in the model.

![Figure 1. A system overview](image)

2.1 Reconstruction and Segmentation

The stereo endoscope provides left and right camera images, which are first preprocessed to remove distortion and to rectify the image pair. Using correspondence analysis we first calculate a disparity map between the two images and then triangulate those matches, resulting in a dense 3D point cloud \( R_i \) in camera coordinates for each time step. The preprocessing and the correspondence analysis were both implemented on the GPU.

Every pixel in the scene is simultaneously classified using a random forest into foreground, i.e. organs of interest and background. As features, the hue and saturation channels from the HSV color space and the color-opponent dimensions \( a \) and \( b \) from the LAB color space were used. The classifier thus provides for each time step a mapping \( C_i(r) \to \{1 \ldots n\}, r \in R_i \) from 3D point to a class-label.
The random forest was trained on multiple previously labeled images. We trained a forest consisting of 50 trees with a maximum depth of 10. To allow real-time processing, the classification portion of the random forest was ported to the GPU.

### 2.2 Integration into TSD-Volume

Assuming the pose $P_i$ of the camera in each time step is known, the point clouds $R_i$ can be transformed into the world coordinate system $R_W^i = P_i(R_i)$. At every time step, $R_W^i$ is integrated into a TSD volume $S_i(p) \rightarrow \{F_i(p), K_i(p, j), W_i(p)\}$, where $p$ is a voxel in the volume. The truncated signed distance value $F_i(p)$ and the weight $W_i(p)$ are computed as suggested in. We included $K_i(p, j)$ in the volume to account for class membership of $p$:

\[
K_i(p, j) = \frac{W_{i-1}(p)K_{i-1}(p, j) + W^W_{i-1}(p)K^W_{i-1}(p, j)}{W_{i-1}^W_{i-1} + W^W_{i-1}}
\]

(1)

\[
K^W_{i-1}(p, j) = \begin{cases} 
1 & \text{if } C_i(R^W_{i-1}(p)) = j \\
0 & \text{else} 
\end{cases}
\]

(2)

where $R^W_{i}(p)$ represents the point in $R^W_{i}$ that lies in $p$.

The class membership $C_i(p)$ at the current time step can then be computed as:

\[
C_i(p) = \underset{j \in \{1..n\}}{\text{argmax}} K_i(p, j)
\]

(3)

This way of smoothing class membership over time allows our system to cope with potential misclassifications.

### 2.3 Camera Pose

To integrate the point cloud $R_i$ into the TSD volume, the pose $P_i$ of the camera at time step $i$ has to be known. In this paper, we consider three methods for estimating $P_i$:

- We make the same assumption as in\textsuperscript{4} that the pose of the camera changes only slightly between frames. By registering $R_i$ with a ray cast of the TSD volume using the iterative closest point algorithm,\textsuperscript{7} we estimate $P_i$. (ICP)
- We use an NDI Polaris optical tracking system to track both camera and the patient. (Polaris)
- We combine the two methods by using the tracking information as a seed for the ICP. (Mixed)

### 2.4 Experimental Setup

Our experimental setup (figure 2) consists of a calibrated PAL stereo endoscope (from Richard Wolf, with a resolution of $720 \times 576$ pixels and a $30^\circ$ optic), an NDI Polaris optical tracking system used to track the endoscope and the patient, and a patient phantom with silicone organs. A workstation PC with an Intel Core i7-2700K CPU, a GeForce GTX 650Ti GPU and 16 Gbyte of RAM was used.

For evaluation purposes recordings from three livers, one ex-vivo porcine liver and two synthetic silicone livers, taken with the stereo endoscope, were used (figure 2.4). For the porcine liver and one of the silicone livers, we used laser scans as ground truth. For the second silicone liver, the ground truth was a CT scan.
3. RESULTS

We performed three experiments to evaluate our system using one porcine liver (Porcine) and two silicone livers (Silicone 1 and Silicone 2). For each liver, a ground truth was also computed. In each experiment, we passed the stereo endoscope over the liver and used the captured images to reconstruct and segment the liver simultaneously. For each experiment three mosaicked models, each with a different method for tracking the camera pose, were constructed. We then computed for each model the average distance of each point to the ground truth. The mosaicked porcine liver was registered to the ground truth using ICP. For the sake of comparison, we also computed the average distance of the unprocessed single frame point clouds $R_{W}$ to the ground truth. The camera pose used for transforming each point cloud into the world coordinate system was given by the Polaris. The results are given in table 1. The data clearly shows that our method, when using the optical tracking system to determine the position of the camera, reduces the average distance of the model points to the ground truth compared to the results from single reconstructions alone. As can be seen in figure 4, the error, in case of the optical tracking, seems to accumulate at the border of the reconstructed object. This

|          | Polaris  | ICP      | Mixed    | Single Frames |
|----------|----------|----------|----------|---------------|
| Porcine  | 4.21 ± 0.63 | 10.37 ± 4.31 | 10.94 ± 4.47 | 6.88 ± 1.67   |
| Silicone 1 | 1.89 ± 0.19 | 5.04 ± 3.82  | 3.79 ± 1.6 | 2.88 ± 2.01   |
| Silicone 2 | 3.64 ± 0.31 | 6.73 ± 4.11  | 6.14 ± 2.57 | 5.36 ± 0.99   |

Table 1. The RMS error between the mosaicked models and the ground truth models in mm and the standard deviation.
can be attributed to fewer viewpoints collected for the border than for the center, meaning fewer samples were available to smooth outliers.

Using either one of the ICP based methods though increases the error. This can be attributed to the homogeneous structure of the livers used, since ICP tends to converge in local minima. Figure 5 illustrates an example where the ICP based methods failed.

To determine if the models created by our approach are suitable for registering a preoperative model in absence of soft tissue deformation, we transformed the model for Silicone 1 with random rigid transformation multiple times. We then did a rough registration of the model to the ground truth laser scan with8 and then fine-tuned using an ICP. The average distance error for 600 random transformations was 13.19 mm ± 23.39, with 90% having an error less than 10 mm. In comparison, using only a single frame reconstruction had an error of 89.92 mm ± 20.48.

The integration of a reconstruction into the TSD volume took on average 0.122 seconds, the average run-time for the segmentation was 0.045 seconds and the overall run-time, including the 3D reconstruction required 0.252 seconds, implying a frame-rate of about 4 fps.

4. DISCUSSION

In this paper, we presented an approach that makes it possible to reconstruct and segment organs from multiple viewpoints online during laparoscopic surgery. We have clearly demonstrated that mosaicking multiple reconstruction reduces the distance error when compared to single reconstruction. Furthermore we have also shown that using a mosaicked model for rigid registration produce a significant smaller error.

Three methods for mosaicking from a moving camera were evaluated, showing that using an optical tracking system produces more accurate results than two ICP-based methods when considering homogeneous structures.

Future research will focus on accounting for a dynamic scene, as so far only static scenes were considered, meaning that soft tissue deformation was not taken into account. Since it was shown that ICP does not work well with homogeneous structures, evaluating other registration algorithms should be considered to lessen the dependency on an optical tracking system.
ACKNOWLEDGMENTS

The present research was conducted within the setting of the research training group 1126 and the SFB/Transregio 125 funded by the German Research Foundation. It is furthermore sponsored by the European Social Fund of the State Baden-Wuerttemberg.

REFERENCES

[1] Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., et al., “Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery,” Medical image analysis 17(8), 974–996 (2013).

[2] Röhl, S., Bodenstedt, S., Suwelack, S., Kenngott, H., Müller-Stich, B. P., Dillmann, R., and Speidel, S., “Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration,” Medical Physics 39, 1632 (2012).

[3] Schroff, F., Criminisi, A., and Zisserman, A., “Object class segmentation using random forests,” 54.1–54.10 (2008). doi:10.5244/C.22.54.

[4] Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohli, P., Shotton, J., Hodges, S., and Fitzgibbon, A., “Kinectfusion: Real-time dense surface mapping and tracking,” in [Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality], ISMAR ’11, 127–136, IEEE Computer Society, Washington, DC, USA (2011).

[5] Newcombe, R. A. and Davison, A., “Live dense reconstruction with a single moving camera,” in [Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on], 1498–1505 (2010).

[6] Haase, S., Bauer, S., Wasza, J., Kilgus, T., Maier-Hein, L., Schneider, A., Kranzfelder, M., Feumer, H., and Hornegger, J., “3-d operation situs reconstruction with time-of-flight satellite cameras using photogeometric data fusion,” in [Medical Image Computing and Computer-Assisted Intervention MICCAI 2013], Mori, K., Sakuma, I., Sato, Y., Barillot, C., and Navab, N., eds., Lecture Notes in Computer Science 8149, 356–363, Springer Berlin Heidelberg (2013).

[7] Blais, G. and Levine, M. D., “Registering multiview range data to create 3d computer objects,” Pattern Analysis and Machine Intelligence, IEEE Transactions on 17(8), 820–824 (1995).

[8] Rusu, R. B., Blodow, N., and Beetz, M., “Fast point feature histograms (fpfh) for 3d registration,” in [Robotics and Automation, 2009. ICRA’09. IEEE International Conference on], 3212–3217, IEEE (2009).