Ralf Hackner*, Martin Raithel, Edgar Lehmann, Thomas Wittenberg

Deep-learning based reconstruction of the stomach from monoscopic video data

Abstract: For the gastroscopic examination of the stomach, the restricted field of view related to the „keyhole"-perspective of the endoscope is known to be a visual limitation. Thus, a panoramic extension can enlarge the field of vision, supports the endoscopist during the examination, and ensures that all of the inner stomach walls are visually inspected. To compute such a panorama of the stomach, knowledge about the geometry of the underlying structure is required. Structure from motion an approach to reconstruct the necessary information about the 3D-structure from monocular image sequences as provided by a gastroscope. We examine and evaluate an existing deep neuronal network for stereo reconstruction, in order to approximate the geometry of stomach parts from a set of consecutive acquired image pairs from gastroscopic videos.

Keywords: Endoscopy, 3D-reconstruction, deep neural networks, panoramic imaging.

https://doi.org/10.1515/cdbme-2020-3012

1 Introduction

One challenge in diagnostic and interventional endoscopy including gastroscopy (endoscopic examination of the upper GI-tract of esophagus, stomach and duodenum) is the limited 'field-of-view' through an endoscope, also known as 'keyhole view'. During a typical gastroscopic examination, the flexible endoscope is advanced through the patient's mouth into the esophagus, entering the stomach through the cardia and is then moved through the stomach until the duodenum is reached. During the insertion process, the endoscopist usually does not look for details of the surface, but rather concentrates on the rapid insertion of the endoscope into the lumen down to the duodenum. The examination takes place in a second step, when the endoscope is slowly withdrawn through the pylorus, passing the antrum and the body of the stomach. Inside the stomach, the physician turns and bends the endoscope tip for a closer examination of the tissue of surface. Also, the endoscope tip is retroflexed in order to allow an assessment of fundus and cardia.

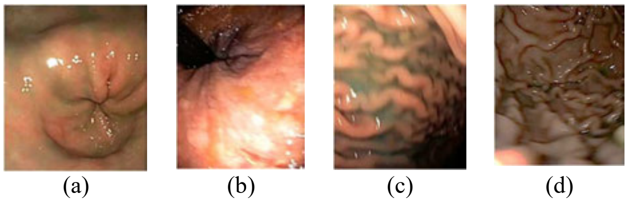

Figure 1: Example of endoscopic images on the stomach: (a) pylorus, (b) antrum, (c) larger curvature, (d) fundus.

Even though in the past two decades, technological improvements in the field of endoscopic imaging have been proposed, such as HD-endoscopy [1,2], near-focus [3,4], wide-angle [5, 6] or magnifying endoscopes [7,8], the limitations of the 'keyhole view' (only small parts of the observed tissue can be seen, even though in high spatial resolution) still remains, cf. Fig 1.

Thus, the endoscopist never perceives a complete view of the stomach wall, and must fuse the already seen scenes together in his/her mind in order to a complete a so-called 'texture map' of the stomach wall. This subjective 'visual impression' and the corresponding clinical findings are later trans- and described textually in the clinical report, and recited from memory in interdisciplinary conferences and councils. The image-based documentation of the upper GI tract is currently limited to single image frames of lesions (not depicting the anatomical context), or videos (needing time for replay and viewing). The above-mentioned 'fused image information of the hollow organ' of the endoscopist is only available in his/her mind and can thus not be used for interdisciplinary discussions or education. Also, anatomical anomalies (e.g. organo-axial torsion), functional alterations (e.g. fundus cascade) or changes in organ diameter, space and hernia are constructed in the endoscopists brain and reported in the finding, but usually need further radiological confirmation by X-ray or CT. In order to compensate the 'key-hole' effect during an endoscopic examination of the GI tract, as e.g. the colon, multiple-view endoscopes such as 'third eye' endoscopes [9, 10] or 'full-spectrum endoscopy' (FUSE)" [11, 12] have been proposed and evaluated. Even though the lesion-detection rates reported

***Corresponding author: Ralf Hackner:** Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany, e-mail: ralf.hackner@fraunhofer.de
Martin Raithel: Malteser Waldkrankenhaus St. Marien, Erlangen
Edgar Lehmann: E&L medical systems, Erlangen
Thomas Wittenberg: Fraunhofer IIS, Erlangen

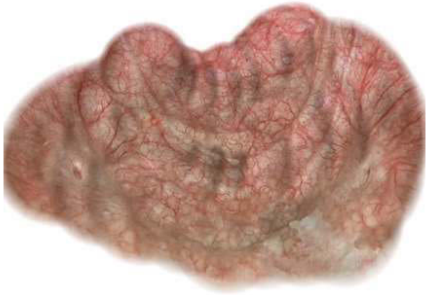

Figure 2: Endoscopic panorama of the urinary bladder floor, stitched together from approx. 50 single views.

for these systems have improved, the acquired image data can still not be used for an image-based documentation of the endoscopic examination and the clinical report of type and location of a finding.

Using the urinary bladder as an example of a hollow organ to be assessed and examined with an endoscope, it has recently been shown, that it is possible to provide a so-called 'endoscopic panorama' or 'endoscopic map' of the bladder directly during the endoscopic examination in real time [13, 14], which can afterwards be used for an image-based documentation of the bladder-examination, see Figure 2.

Nevertheless, for the computation of panoramic images of the stomach from endoscopic image sequences, hardly any work is known. *E.g,.Liu et al* [15] have introduced an approach to create a gastric panorama. To address the problem of the 6-DOF freedom of the endoscope tip, additional electric magnet tracker was mounted to the endoscope tip, to track the motion and orientation of the tip inside the hollow with external sensors placed around the patient. *Ali* [16] suggested the computation of a panorama around the pyloric antrum using optical flow between successive image frames for tracking. Capsule endoscopic sequences have been used to create image panoramas [17-19] of the stomach, where the control of the imaging process is limited as the capsules cannot be controlled. In our previous work to obtain panoramic images of the stomach only subsections were considered and the 3D-geometry was neglected [20].

Hence, in this work we evaluate deep neural networks with a transfer learning approach to stitch a set of monoscopic gastric images to obtain a panorama view of the stomach and furthermore provide 3D information of the stomach wall.

2 Materials

Image data was obtained from the department of gastroenterology of the Malteser Waldkrankenhaus Erlangen. We used 8 anonymized video sequences of gastroscopic examinations (Olympus GIF HQ180). To obtain a reference for the evaluation of our 3D-reconstruction, a silicon stomach phantom was used. The phantom was scanned with a 3D hand scanner (Systems Cubify Sense 3D) from the outside. To obtain image sequences from the phantom's inside, a portable gastroscope (Storz Gastro Pack 2504 30) was applied. To determine corresponding landmarks on the in- and outside of the phantom, small ball magnets were used as beacons, pairwise placed on the in- and outside at defined positions during the 3D scan, as well as during the recording of the image sequences.

3 Methods

Several approaches to reconstruct depth from image data exist. Beyond reconstruction based on stereo disparity information (not available in our case), it is possible to reconstruct depth from other features, like motion, illuminance, textures or sharpness. In our work we apply the depth estimation Motion Network (DeMoN) of *Ummenhofer et al.* [21]. This network relies on motion disparity and has originally been trained with 16,152 real and synthetic street and indoor scenes (such as sun3d [22]). The resulting weights are publicly available under [23]. The DeMoN network consist of a sequence of various encoder-decoder networks, grouped into a so-called 'bootstrap', an 'iterative' and a 'refinement' network applied on the image pairs [21]. The first two subnets are pairs of encoder-decoder networks. The first one computes the optical flow of an image pair while the second calculates depth and camera motion. The 'iterative' net is applied recursively to refine the results of the previous iteration. The last component is a single encoder-decoder network generating the final up-sampled and refined depth map [21] a resolution of 256x192 pixels. Based on the depth map and the corresponding original image, it is possible to create a point cloud, describing the 3D structure of the observed scene.

The key questions of this experiment were as follows: (a) Can this DeMoN network be used in principle for the 3D-approximation of monoscopic endoscopy data (without retraining of the network); and (b) how good are the achieved results? To answer these questions, two experiments were conducted, namely on one hand feeding the DeMoN network with various image pairs of gastroscopy data, and qualitatively evaluating the outcome, and secondly inserting image pairs from the phantom data and comparing the results with the 3D-scan of the phantom.

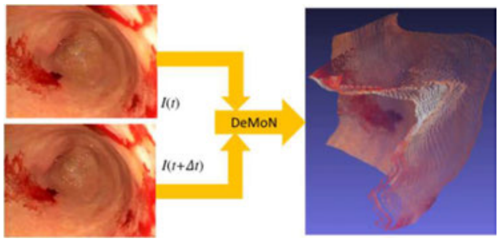

Figure 3: A pair of consecutive images with interval of Δt is inserted into the DeMoN network, which yields an outcome of a 3D-mesh.

3.1 Reconstruction

For the first set of experiments, we extracted 48 image pairs from the above mentioned gastroscopic video streams of the stomach. With an interval of Δt = 0.3 to 0.7 seconds, a light motion disparity between image pairs can be observed in all cases. To reduce errors caused by the endoscopic wide angle sensor, a fisheye calibration algorithm was used to undistort all images. All images were manually cropped to the region of interest and subsampled to the required resolution of 256 x 192 pixels. These image pairs are presented to the DeMoN network as input, and yield in result a 3D mesh, describing the 3D approximation of the underlying scene in the stomach, see Fig. 3. In Figure 4, different views of the DeMoN based 3D-reconstruction of the pylorus from a monocular image pair is provided as example. Due to missing ground truth, the reconstructed shapes (see Figures 3 and 4) were manually inspected for plausibility. In most cases, the reconstructed geometry matched the expectations. Nevertheless, dark regions are not correctly reconstructed in most cases, meaning the underlying images do not yield enough structural information for the network to match properly.

3.2 Evaluation

For evaluation purposes we used the acquired endoscopy data as well as the external 3D-scan from the silicon stomach phantom (cf. Section 2). In order to align and match the reference 3D scan of the phantom with the DeMoN-based 3D-reconstruction from the monocular endoscopy data, we used the small ball magnet markers visible in the both data sets. Fig. 5 shows a cross section through the 3D-reconstructed artefact (red) aligned with the 3D-reference (blue) and its position on the 3D reference scan (right).

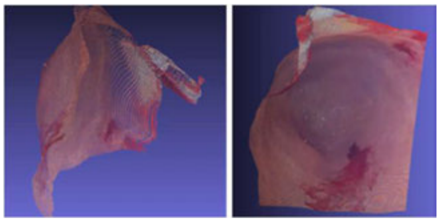

Figure 4: Different views of a 3D reconstruction of the pylorus (exit of the stomach) from monocular gastroscopy data.

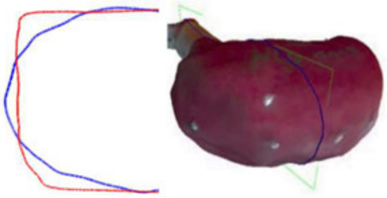

Figure 5: Left: Cross section through the 3D reference scan of the phantom (blue) and the 3D reconstruction (red); Right: related position of the cross section in the stomach phantom.

4 Discussion

In most cases, the reconstructed 3D-structures of the stomach match the geometry expected by the reviewer. Structures in the foreground are separated quite well from structures in the background. Even small and delicate structures such as the stomach folds in the greater curvature have been reconstructed in a plausible manner, as depicted in Fig. 6. Bad illuminated (dark) regions, which have a greater distance to the tip of the endoscope (and thus the light source) cause problems. In some of these cases the DeMoN reconstructs the dark structures falsely close to the observer (see example in Fig. 7). This can explained by the fact, that there are no comparable samples in the original training data of in- and outdoor scenes, This data has its primary light source somewhere far *above* the scene and not coaxial with the image sensor as in endoscopic data. Thus the illumination is rather homogenous in most of the test data.

Due to the missing ground truth for the endoscopic image data, an alternative method for the evaluation of the reconstruction was used. The phantom data mentioned afore can provide us some information about the achieved precision by comparing the reference 3D scan of the phantom with the obtained reconstruction from the DeMoN network. Since the reference scan has been made from the phantom's outside, and the reconstruction is done from inside, the scan is not a perfect ground truth, but gives a close approximation to the real structure, except that the inside features more details. As depicted e.g. in Fig. 5, the cross-sectional shape of the reconstructed

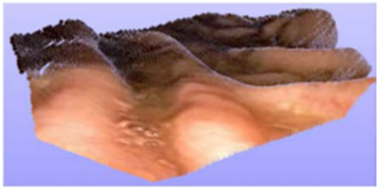

Figure 6: 3D-resconstructed folds on the greater curvature

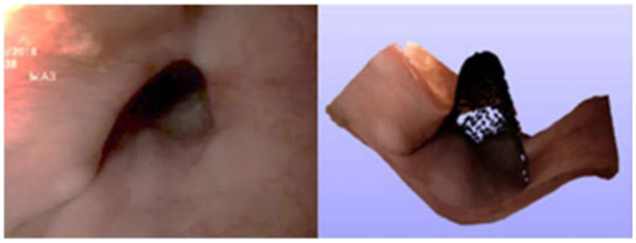

Figure 7: Reconstruction error due to a bad illuminated region (pylorus): original image (left); reconstructed point cloud (right).

structure matches the basic shape of the 3D scan, but there are obvious differences in the curvature and shape. Also the stomach folds of the phantom are not reconstructed. The opposite is true for the real world data, where these details were clearly expressed in the reconstructed material, as depicted in Fig. 6. For the intended purpose of panoramic imaging of the stomach, this is most likely not a problem, because only rough knowledge of the geometry is necessary here to find a suitable projection method. Other applications, such as measuring distances are not yet possible with the quality we obtained.

5 Conclusion

With our approach, we were able to yield 3D-approximations of stomach parts based on pairs of monocular gastroscopic images. The quality of the obtained depth maps is sufficient for a subsequent generation of 3D-panoramas in most of our test cases, even though the DeMoN network was originally trained with datasets from a different domain. Only small motions of the gastroscope in a short time interval ($\Delta t = 0.3$ to 0.7 seconds) were necessary to extract sufficient information for a 3D-reconstruction. Thus, the effect of possible ego-motion of the stomach walls can be neglected. In some rare cases, obvious reconstruction errors appeared. These errors mostly appear in scenarios with bad illumination conditions, which have so far been neglected in the training data.

Retraining the DeMoN network with native endoscopy data is envisioned in the next step to improve the reconstruction quality. To obtain adequate training data for the disparity, disparity estimations based on real stereo-endoscopy images shall be used. As currently stereo recordings from gastroscopic are not available, stereo sequences from laparoscopic surgeries shall be used, which feature similar conditions with respect to illumination and texture.

Author Statement

Parts of the research has been funded by the Bavarian Research Society (BFS) under the project GastroMapper (AZ-1349-18).

References

[1] Rey et al. New aspects of modern endoscopy. World J Gastrointest Endosc 2014; 6(8):334-44.

[2] Neumann et al. Advanced endoscopic imaging to im-prove adenoma detection. World J Gastro Endo 2015; 7(3): 224-9.

[3] Waldner et al. Imaging of mucosal inflammation: Current technological developments, clinical implications, and future perspectives. Front Immunol 2017; 8: 1256.

[4] Szura et al. Two-stage optical system for colorectal polyp assessments. Surg Endosc 2016; 30: 204–214.

[5] Pellisé et al. Impact of wide-angle, high-definition endoscopy in the diagnosis of colorectal neoplasia: A randomized controlled trial. Gastroent 2008; 135 (4) 1062-8.

[6] Deenadayalu et al. 170° wide-angle colonoscope: effect on efficiency & miss rates. Am J Gastroent 2004; 99, 2138-42.

[7] Chai et al. Magnifying endoscopy in upper gastroenterology for assessing lesions before completing endoscopic removal. WJ Gastroent 2012; 8(12): 1295-307.

[8] Mabe et al. An educational intervention to improve endoscopist's ability to correctly diagnose small gastric lesions using magnifying endoscopy with narrow-band imaging. Ann Gastroenterol 2014; 27(2) 149-55.

[9] Patel et al. The endoscopy evolution: the superscope era. Frontline Gastroenterol 2015; 6(2)101-7.

[10] Rubin et al. Expanding the view of a standard colonoscope with the Third Eye® panoramic cap. 2015; 21(37):10683-7.

[11] Ratone et al. Impact of full spectrum endoscopy (Fuse, EndoChoice®) on adenoma detection: a prospective French pilot study. Ann Gastroent. 2017; 30(5): 512-7.

[12] Gralnek et al. Standard forward-viewing colonoscopy vs. full-spectrum endoscopy: international, multicentre, randomised, tandem colonoscopy trial. Lanc Oncol 2014; 15(3):353-60.

[13] Bergen T. Real-time endoscopic image stitching for cystoscopy. PhD Thesis, Univ. Koblenz-Landau: 2017.

[14] Kriegmair et al. Digital mapping of the urinary bladder: potential f. standardized cystoscopy reports. Urol 2017; 104:235-41.

[15] Liu et al. Global and local panoramic views for gastroscopy: An assisted method of gastroscopic lesion surveillance. Trans Biomed Eng 2015; 62(9) 2296-307.

[16] Ali. Total variational optical flow for robust and accurate bladder image mosaicing. PhD Thesis, Univ. de Lorraine: 2016

[17] Fan et al. 3D-reconstruction of wireless capsule endoscopy images, Proc's EMBC: 2010; 5149-52.

[18] Turan et al. Sparse-then-dense alignment-based 3D map reconstruction method for endoscopic capsule robot. Mach. Vis. & App. 2018; 29(2) 345-359.

[19] Maciura, Bazan. Granular computing in mosaicing of images from capsule endoscopy. Nat. Comp. 2015; 14(4) 569-77.

[20] Hackner et al. Panoramic endoscopy of the stomach: first results from phantom and patient data. Proc's CURAC 2018; 10-15.

[21] Ummenhofer et al. DeMoN: depth and motion network for learning monocular stereo", Procs' CVPR, 2017; 5038-5047.

[22] Xiao et al. SUN3D: A database of big spaces reconstructed using SfM and Object Labels. ICCV 2013; 625-1632

[23] DeMoN: Depth and Motion Network, *https://github.com/lmb-freiburg/demon.*