Winner of MICCAI 2019 Best Presentation Award

Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions

Ruibin Ma and Rui Wang, Stephen Pizer, Julian Rosenman, Sarah K. McGill, and Jan-Michael Frahm

University of North Carolina at Chapel Hill

Abstract. Colonoscopy is the most widely used medical technique to screen the human large intestine (colon) for cancer precursors. However, frequently parts of the surface are not visualized, and it is hard for the endoscopist to realize that from the video. Non-visualization derives from lack of orientations of the endoscope to the full circumference of parts of the colon, occlusion from colon structures, and intervening materials inside the colon. Our solution is real-time dense 3D reconstruction of colon chunks with display of the missing regions. We accomplish this by a novel deep-learning-driven dense SLAM (simultaneous localization and mapping) system that can produce a camera trajectory and a dense reconstructed surface for colon chunks (small lengths of colon). Traditional SLAM systems work poorly for the low-textured colonoscopy frames and are subject to severe scale/camera drift. In our method a recurrent neural network (RNN) is used to predict scale-consistent depth maps and camera poses of successive frames. These outputs are incorporated into a standard SLAM pipeline with local windowed optimization. The depth maps are finally fused into a global surface using the optimized camera poses. To the best of our knowledge, we are the first to reconstruct dense colon surface from video in real time and to display missing surface.

Keywords: Colonoscopy · SLAM · Reconstruction · RNN.

1 Introduction

Colorectal cancer is the third most common cancer in men and the second in women worldwide [6]. Colonoscopy is an effective method of detecting and removing pre-malignant polyps.

There is strong evidence to support the assertion that polyps and adenomas of all kinds are missed at colonoscopy (pooled miss-rate 22% [8] among multiple studies). An important cause is that the colonic mucosal surface was not entirely surveyed [5]. However, it is very difficult to detect missing colonic surface from video alone, let alone quantify its extent, because one sees only a tiny fraction of the colon at any given time rather than a more global view. The solution is to build a system to visualize missing colon surface area by reconstructing the streaming video into a fully interactive dense 3D textured surface that reveals holes in the surface if regions were not visualized (Fig. 1). This should be done

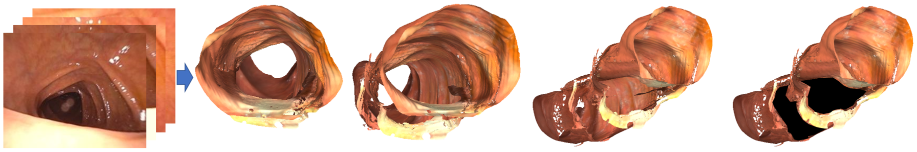

Fig. 1: 3D reconstruction for visualization of missing colonic surface (highlighted in black in the last image, 25% surface), small colon pouches that are occluded by ridges.

in real time so that the endoscopist can be alerted to the unseen surface in a timely manner so that the situation can be remedied.

Hong et al. [4] used haustral geometry to interpolate the virtual colon surface so as to find missing regions. However, their work only provided single-frame reconstruction and haustral occlusion (without fusion), which is inadequate to determine what has been missed during the procedure. Also, there is no inter-frame odometry being used, which could boost reconstruction accuracy. Armin et al. [1] produced a 2D visibility map which was less intuitive than a 3D dense reconstruction. Zhao et al. [15] used Shape From Motion and Shading for dense endoscopy reconstruction but is not real time.

The SLAM (simultaneous localization and mapping) [7, 2, 3] and the Structure-from-Motion (SfM) methods [9] take a video as input and generate both 3D point positions and a camera trajectory. However, besides the fact that most of them do not generate dense reconstructions, they work poorly on colonoscopy images for the following reasons: 1) colon images are very low-textured, which is a disadvantage for the feature-point-based methods, e.g., ORBSLAM [7]; 2) photometric variations (caused by moving light source, moist surface and occlusions) and geometric distortions make tracking (predicting camera pose and 3D point positions for each frame) too difficult; 3) lack of translational motion and poor tracking leads to severe camera/scale drift (Fig. 2) and noisy 3D triangulation.

Convolutional neural networks (CNN) have been used for SLAM tasks and predicting dense depth maps [16, 12, 14]. However, these end-to-end networks are subject to accumulated camera drift because there is no optimization used during prediction as in standard SLAM systems. In contrast, there are works

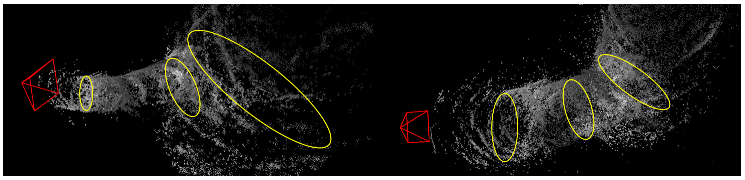

Fig. 2: Left: Sparse point cloud of a chunk of colonoscopy video produced by a standard SLAM pipeline (DSO) [2]; right: Sparse point cloud produced by ours (intermediate result). The cross sections are approximated by yellow ellipses. The diameters of the DSO result are dramatically decreasing (scale drift), which is non-realistic. Our result has a much more consistent scale thanks to the depth maps predicted by the RNN.

that use CNN to improve a standard SLAM system [11, 13]. CNN-SLAM [11] incorporated CNN depth prediction to the LSD-SLAM [3] pipeline to provide robust depth initialization. The dense depth maps are finally fused into a global mesh. Yang et al. [13] used CNN-predicted depth (trained on stereo image pairs) to solve the scale drift problem in Direct Sparse Odometry (DSO) [2]. However, there are neither stereo images nor groundtruth depth for colonoscopy images. Also, training a CNN on colonoscopy images will be difficult due to the aforementioned challenges.

In this paper, we present a deep-learning-driven colonoscopic SLAM system. We develop a recurrent neural network (RNN) to predict both depth and camera poses and combine it in a novel fashion with a SLAM pipeline to improve the stability and drift of successive frames' reconstructions. The RNN training addresses the difficulties of reconstructing from colonoscopy images. The SLAM pipeline optimizes the depth and camera poses provided by the RNN. Based on these optimized camera poses, the depth maps of the keyframes are fused into a textured global mesh using a nonvolumetric method. Our method produces a high-quality camera trajectory and colon reconstruction which can be used for missed region visualization in colonoscopy. The whole system runs in real time.

2 Methodology

2.1 Full pipeline

The full pipeline includes the following steps: 1) Deep-learning-driven tracking: predicting frame-wise depth map and tentative camera pose which are used to initialize the photoconsistency-based tracking; 2) Keyframe selection: upon enough camera motion, creating a new keyframe as the new tracking reference and updating the neural network; 3) Local windowed optimization: the camera poses and sparsely sampled points' depth values of the latest N (e.g., 7) keyframes are jointly optimized; 4) Marginalization: the oldest keyframe in window is finalized, i.e., marginalized from the optimization system; 5) Fusion: using optimized camera pose, the image and the depth map of the marginalized keyframe is fused with existing surface. We will detail item 1 in Sec 2.2, items 2-4 in Sec 2.3 and item 5 in Sec 2.4.

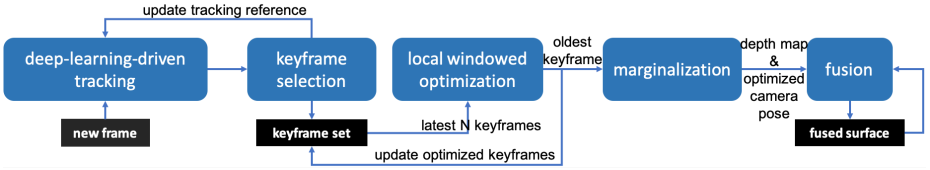
update tracking reference
deep-learning-driven tracking
keyframe selection
local windowed optimization
oldest keyframe
marginalization
depth map & optimized camera pose
fusion
new frame
keyframe set
latest N keyframes
update optimized keyframes
fused surface

Fig. 3: Flow chart of presented deep-learning-driven colonoscopic SLAM system

2.2 Deep-learning-driven tracking

Our deep-learning-driven tracking is developed upon RNN-DP (a **r**ecurrent **n**eural **n**etwork for **d**epth and **p**ose estimation [12]) that predicts a depth map and a camera pose for each image in the video. However, it cannot be directly trained on colonoscopy videos because there is no groundtruth depth available. In addition, the pose estimation network in RNN-DP is trained based on image reprojection error, which is severely affected by the specular points and occlusions in colonoscopy videos. Therefore, in this section we present several new strategies that allow RNN-DP to be successfully trained on colonoscopy videos.

To solve the problem of the lack of groundtruth depth, we used SfM [9] to produce a sparse depth map for each individual colonoscopy video frame. These sparse depth maps are then used as groundtruth for RNN-DP training. We collected 60 colonoscopy videos, each containing about 20K frames. Then we grouped every 200 consecutive frames into a subsequence with an overlap of 100 frames with the previous subsequence. Thereby we generated about 12K subsequences from 60 colonoscopy videos. Then we ran SfM [9] on all the subsequences to generate sparse depth maps for each frame. Following the training pipeline in RNN-DP [12], these sparse depth maps are used as ground-truth for training.

To avoid the error from specularity (saturation), we computed a specularity mask M^t_{spec} for each frame based on an intensity threshold. Image reprojection error at saturated regions are explicitly masked out by M^t_{spec} during training.

Colonoscopy images also contain severe occlusions by haustral ridges, so a point in one image may not have any matching point in other images. The original RNN-DP did not handle occlusion explicitly. In order to properly train it on colonoscopy video, we compute an occlusion mask M^t_{occ} to explicitly mask out image reprojection error at occluded regions. The occlusion mask is determined by a forward-backward geometric consistency check, which was introduced in [14].

Our improved RNN-DP outputs frame-wise depth maps and tentative camera poses (relative to the previous keyframe). They are used to initialize the photoconsistency-based tracking [2] that refines the camera pose.

2.3 Keyframe management and optimization

In this subsection, we will briefly review how a vanilla SLAM pipeline (DSO) works and then introduce how RNN-DP interacts with the system.

Besides (deep-learning-driven) tracking, the other three main modules of the SLAM system are keyframe selection, local windowed optimization and marginalization. The SLAM system keeps a history of all keyframes. The latest keyframe is used as the tracking reference for the incoming frames. In the keyframe selection module, if the relative camera motion or the change of visual content (measured by photoconsistency) is large enough, the new frame will be inserted into the keyframe set. It will then be used as a new tracking reference.

When a keyframe is inserted, the local windowed optimization module is triggered. The local window contains the latest 7 keyframes. From each of these keyframes, 2000 2D active points are sampled in total, preferring high-gradient regions. Each active point is based on exactly one keyframe but is projected to other keyframes to compute a photometric error. By minimizing the total photometric loss, the camera poses (7×6 parameters) and the depth values of the sampled points (2000 parameters) are jointly optimized. In addition, to tolerate global brightness change of each keyframe, two lighting parameters per frame are added to model the affine transform of brightness. The purpose of the sampling is to enable efficient joint optimization by maintaining sparsity.

After optimization, the oldest keyframe is excluded from the optimization system by marginalization based on the Schur complement [2]. The finalized reconstructed keyframe is to be fused into the global mesh.

The SLAM system is improved using our RNN-DP network. In the keyframe selection module, when a new keyframe is established, the original DSO used the dilated projections of existing active points to set the depth map for this keyframe, which is used in the new tracking tasks. The resulting depth map is sparse, noisy and is subject to scale drift. In our method we set the depth map for this keyframe using the depth prediction from the network. Our depth maps are dense, more accurate and scale consistent. As a result, it makes the SLAM system easier to bootstrap, which is known to be a common problem for SLAM. On the other hand, the SLAM system also improves the result of raw RNN-DP predictions by optimization, which is very important to eliminate accumulated camera drift of RNN-DP. In summary, this is a win-win strategy.

Our RNN-DP network is integrated into the SLAM system. Its execution is directed by the keyframe decisions made by the system. After tracking, the hidden states of the RNN-DP remain at the stage of the latest keyframe. They are updated only when a new keyframe is inserted.

2.4 Fusion into a chunk

The independent depth maps predicted by the RNN-DP need to be fused into a global mesh. We use a point-based (nonvolumetric) method called SurfelMeshing [10]. It takes a RGB+depth+camera sequence as input and generates a 3D surface. Since SurfelMeshing requires well-overlapped depth maps, we add a preprocessing step to further align the depths.

Windowed depth averaging: the fusion module keeps a temporal window that keeps the latest 7 marginalized keyframes. In parallel, the depth map of the 6 old keyframes are first projected to the latest keyframe. Second, the new keyframe replaces its depth with the weighted average of the projected depth maps and its current depth. The weights are inversely proportional to time intervals. The average depth is used for fusion. This step effectively eliminates the non-overlapping between depth maps at a cost of slight smoothing.

The fusion result (a textured mesh) is used for missing region visualization and potentially for region measurement.

3 Experiments

Our algorithm is currently able to reconstruct a colon in chunks when the colon structure is clearly visible. The end of a chunk is determined by recognizing a sequence of non-informative frames, e.g., frames of intervening material or bad lighting, whose tracking photoconsistencies are all lower than a threshold. The chunks we reconstructed are able to visualize the missing regions. We provide quantitative results estimating the trajectory accuracy and qualitative results on the reconstruction and missing region visualization.

3.1 Trajectory Accuracy

To evaluate the trajectory accuracy, we compare our method to DSO [2] and RNN-DP [12]. Since there is no groundtruth trajectory for colonoscopic video, to generate high quality camera trajectories in an offline manner, we use colmap [9], which is a state-of-the-art SfM software that incorporates pairwise exhausted matching and global bundle adjustment. These trajectories are then used as "groundtruth" for our evaluation.

Evaluation metrics. We use the absolute pose error (APE) to evaluate global consistency between the real-time system estimated and the colmap-generated "groundtruth" trajectory. We define the relative pose error E_i between two poses $P_{gt,i}, P_{est,i} \in \mathrm{SE}(3)$ at timestamp i as

$$E_i = (P_{gt,i})^{-1} P_{est,i} \in \mathrm{SE}(3) \tag{1}$$

The APE is defined as

$$APE_i = ||trans(E_i)|| \tag{2}$$

where $trans(E_i)$ refers to the translational components of the relative pose error. Then different statistics can be calculated on the APEs of all timestamps, e.g., the RMSE:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} APE_i^2} \tag{3}$$

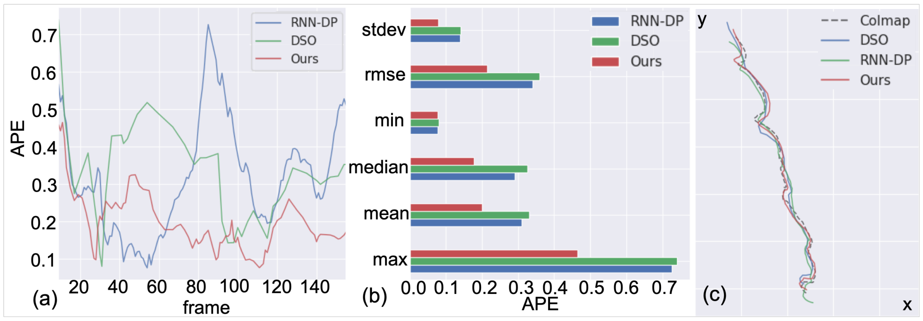

Fig. 4: Evaluation result on one colonscopy sequence. (a) APE of the three approaches across the whole sequence. (b) Statistics based on APE. (c) A bird's-eye view of the full trajectories.

Method	rmse	std	min	median	mean	max
RNN-DP	0.617	0.253	0.197	0.518	0.560	1.229
DSO	0.544	0.278	0.096	0.413	0.465	1.413
Ours	**0.335**	**0.157**	**0.074**	**0.272**	**0.294**	**0.724**

Table 1: Average statistics based on the APE across 12 colonoscopic sequences

Fig. 4 shows evaluation results on one colonoscopic sequence. Fig. 4.a compares the absolute pose error (APE) of the three approaches on the example sequence: our result (red) has the lowest APE at most times. Fig. 4.b shows APE statistics of the three approaches: our result is better than the other two approaches. Fig. 4.c shows the trajectories of the three approaches together with the grountruth. Table 1 shows the statistics of Fig 4b but averaged across 12 colonoscopic sequences: we achieve the best result on all the metrics.

3.2 Reconstructions and Missing Regions

Fig. 5 shows two high-quality examples of fused surfaces. The two chunks are dense and textured. It also shows the incremental fusion process of the first example. The snapshots are captured in real time.

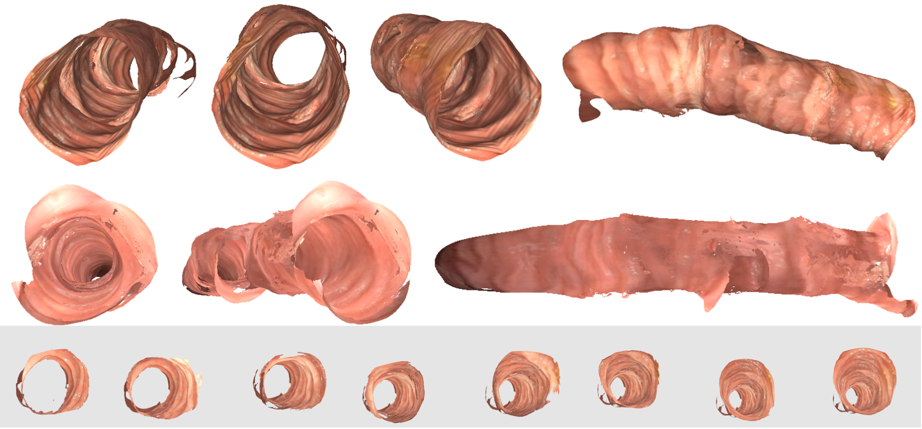

Fig. 5: Rows 1 and 2 each show the reconstruction of a colon chunk from multiple points of view. They have 12% and 10% surface missing. Row 3 shows the incremental fusion of the row 1 example.

There are multiple reasons for missing regions. Two important ones are lack of camera orientations to the full circumference of parts of a colon and haustral occlusion. These two reasons are respectively illustrated in Fig. 6 and Fig. 1. For the four chunks shown in this paper the missing area fraction was notable: 25%, 12%, 10%, and 33% respectively, as verified on the video by our colonoscopiist co-author, Dr. McGill.

Limitations and future work We currently reconstruct in chunks because the tracking will fail upon very large camera motion or deformation. Loop closure is not included in our current system; it could be useful for backward motion. Making the tracking more robust to large deformation and adding loop closure are two future directions.

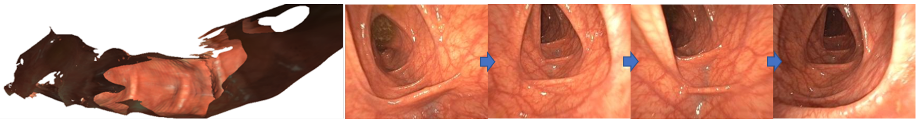

Fig. 6: A part of the colon chunk is missing (33% surface) due to the lack of camera orientations. This can be verified by checking the respective video frames (the upper part of the colon was not seen). However, this might not be realized during a colonoscopy.

4 Conclusion

We developed a deep-learning-driven dense SLAM system for colonoscopy. It is the first to reconstruct chunks of a colon as fused surface from a video sequence (vs. existing single-frame methods) in real time. The reconstructions can be used for the visualization of missed colonic surfaces that lead to potential missed adenomas. Our technical contributions include 1) a recurrent neural network that predicts depth and camera poses for colonoscopic images; 2) integrating the recurrent neural network into a standard SLAM system to improve tracking and eliminate drift, and 3) fusion of colonoscopic frames into a global high-quality mesh. Clinically, it should help endoscopists to realize missed colonic surface and resect more pre-cancerous polyps.

References

1. Armin, A., Chetty, G., De Visser, H., Dumas, C., Grimpen, F., Salvado, O.: Automated visibility map of the internal colon surface from colonoscopy video. International Journal of Computer Assisted Radiology and Surgery **11** (08 2016). https://doi.org/10.1007/s11548-016-1462-8
2. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence (Mar 2018)
3. Engel, J., Schöps, T., Cremers, D.: Lsd-slam: Large-scale direct monocular slam. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 834–849. Springer International Publishing, Cham (2014)
4. Hong, D., Tavanapong, W., Wong, J., Oh, J., De Groen, P.C.: 3d reconstruction of virtual colon structures from colonoscopy images. Computerized Medical Imaging and Graphics **38**(1), 22–33 (2014)
5. Hong, W., Wang, J., Qiu, F., Kaufman, A., Anderson, J.: Colonoscopy simulation. In: Proc.SPIE (2007)
6. Jemal, A., Center, M.M., DeSantis, C., Ward, E.M.: Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiology and Prevention Biomarkers **19**(8), 1893–1907 (2010)

7. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Trans. Robotics **31**(5), 1147–1163 (2015)
8. C van Rijn, J., B Reitsma, J., Stoker, J., Bossuyt, P., van Deventer, S., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: A systematic review. The American journal of gastroenterology **101** (02 2006)
9. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
10. Schöps, T., Sattler, T., Pollefeys, M.: Surfelmeshing: Online surfel-based mesh reconstruction. CoRR **abs/1810.00729** (2018), http://arxiv.org/abs/1810.00729
11. Tateno, K., Tombari, F., Laina, I., Navab, N.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 00, pp. 6565–6574 (July 2017)
12. Wang, R., Pizer, S.M., Frahm, J.M.: Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
13. Yang, N., Wang, R., Stueckler, J., Cremers, D.: Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In: ECCV (2018)
14. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR. pp. 1983–1992 (2018)
15. Zhao, Q., Price, T., Pizer, S., Niethammer, M., Alterovitz, R., Rosenman, J.: The endoscopogram: A 3d model reconstructed from endoscopic video frames. In: MICCAI (2016)
16. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: CVPR (2017)