Monocular 3D Reconstruction of the Colon Using CNNs Trained on Synthetic Data

A. Rau[1], F. Chadebecq[1], P. Riordan[2], D. Stoyanov[1]

[1]*Wellcome / EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London*

[2]*Digital Surgery, London, UK*

a.rau.16@ucl.ac.uk

INTRODUCTION

Colorectal cancer is the third most common cancer worldwide and is rising in incidence. Indicators or precursors to cancerous tissue development can be detected as polyps and removed during colonoscopy. However, complete, endoscopic colon investigation is still challenging and often regions of the colon are not fully examined resulting in high polyp miss rates [1]. Improving the endoscopist's ability to detect abnormal tissue through computational or biphotonic techniques and also to navigate within the colon and reference the position of the camera within the anatomy are significant clinical needs. Better navigation within the colon relies on the ability to map the 3D environment and localize the endoscope within it, but while computer vision advances make this possible in many applications, it has yet to be achieved reliably in endoscopic examination.

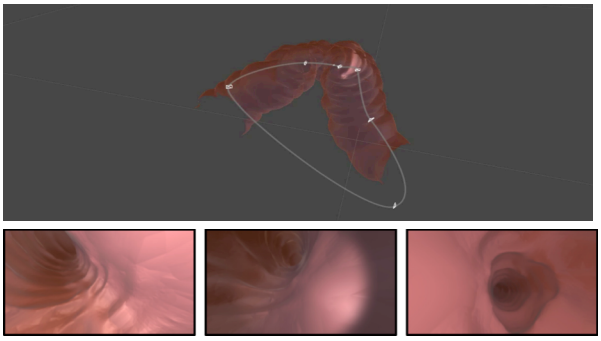

Figure. 1. Colon segment observed within the *Unity* simulation environment (top) with camera trajectory and examples of the endoscopic inside view of the virtual colon (bottom) with different material and lighting configurations generating views from the simulated colon environment.

With recent advances in deep learning, data driven approaches are leading the performance tables in vision based environment mapping. One problem with applying such approaches to endoscopy is that ground truth data is not available to train any CNN models. An appealing alternative, the use of synthetic data, has recently been reported for predicting depth during colonoscopy [2]. After training on simulation, a transformer network, that learns to generate a synthetic representation of real RGB images is used to adapt to real data. Despite promising results the ability of the network to handle real images can still be improved and needs further investigation.

In this paper, we take a different approach and learn directly from the properties that synthetic and real colons have in common, namely, their shape. In synthesizing images with different light/material-configurations, we emphasize the importance of learning a shape prior based on depth instead of using hand modelled assumptions or regularization, as is often used in shape-from-shading, for example. Our experiments with different state-of-the-art algorithms [3,4,5] show that when lighting conditions change, the prediction capability of networks fail and the models lose any understanding of tubular shape. We therefore first learn from depth maps a statistical model that describes the shape of the colon. During training, we then penalize predictions of the CNN that, according to the learned shape model, have a small probability.

MATERIALS AND METHODS

Data Generation: Using the game engine *Unity*, we generate simulated endoscopic renderings of a 3D mesh, which is based on a CT scan of a human colon. We simulate an endoscopic camera with an attached light source that follows a trajectory through the colon (Figure 1). According to a frames/sec rate RBG images and corresponding depth maps, scaled to a depth $\in [0,1]$, are recorded. To obtain a larger data set, we randomly displace and rotate the camera relative to its initial path during render passes.

We generate RGB images using different materials by varying colors and reflection properties (Figure 1). We also vary lighting settings, in particular color, brightness, and angle of the virtual endoscopic illumination. We keep camera parameters constant to ensure the geometry of the colon is consistent and our simulated camera field of view is 140 degrees, consistent with real endoscopes. Our training data consists of nine subsets, each of which is a combination of one out of three material settings and one out of three lightening settings. In total, we generate roughly 11,000 images with depth ground truth.

Shape classification: Distinct from stereo images, where geometrical inference can be drawn from the relation of the position of a landmark in the image pair, estimating depth from a single image is highly ill posed. However, we can benefit from knowledge that the colon has an approximately tubular shape.

Therefore, we first estimate the direction of the lumen in a single image. We use K-means to cluster the depth maps into five groups, which results in for humans distinguishable clusters that depend on the curvature direction of the lumen (Figure 2). Given the clusters, we train a network based on ResNet-50 [6] to classify RGB images into one of the five classes with 89% accuracy. This allows us to classify new images without known depth map and estimate the location of the vanish point.

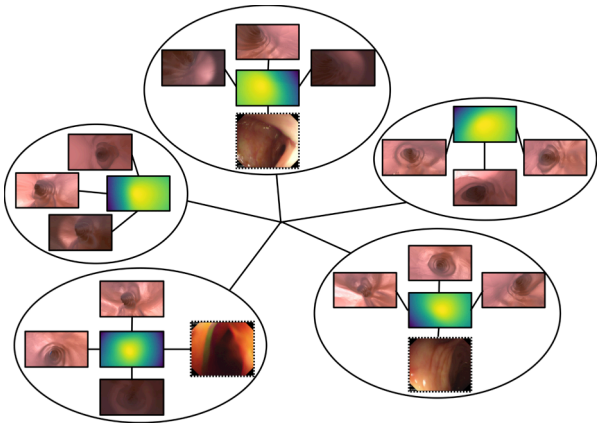

Figure. 2. Mean depth maps of five clusters with examples from the training set. Yellow indicates areas of high depth while blue indicates areas of low depth. Real images that were assigned to the clusters using our classification network, are indicated through a dashed frame.

Depth Estimation: Our architecture is based on two components (Figure 3). We predict the depth based on ResNet-50 followed by a sequence of upscaling layers according to [7] and simultaneously impose a shape that adheres to a statistical model describing the distribution over the depth maps in the training set. A simple approach is using the Maximum Likelihood estimate assuming a Gaussian distribution. To this end, we compute the mean depth of each cluster. During training, we pass the cluster index to the loss function and compute the squared difference between the initial depth estimation and the mean of the given cluster. The final loss function takes into account how close the estimated depth is to both, ground truth and expected shape.

We train our network on six out of the nine subsets in our data set, leaving out all sets that were derived from one of the three lightening settings. The remaining three sets serve as our test set. This allows us to analyze the robustness of the learned model towards changes in illumination. We train each network for 15,000 iterations with a batch size of 32, using Adam optimizer with a learning rate of 10^{-4}. While training takes 18h, depth prediction during test time takes both networks 0.11 sec per image on one NVIDIA TITAN Xp GPU.

RESULTS

While training on ResNet-50 yields a mean distance between ground truth depth and prediction of 0.112 on the test set, our networks yields an error of 0.110, where the maximum depth in the training set is rescaled to 1. Estimating the scale, this roughly corresponds to a mean distance between estimation and ground truth of 9.0 mm on ResNet-50 vs. 8.8 mm on our network. Although the total error on the test set is similar for both networks, we can observe a different source for mistakes. In particular, our network performs better on images that are close to the mean of the cluster (Figure 4a) but fails on images that are outliers (Figure 4b) where the network falsely tries to enforce a tubular shape.

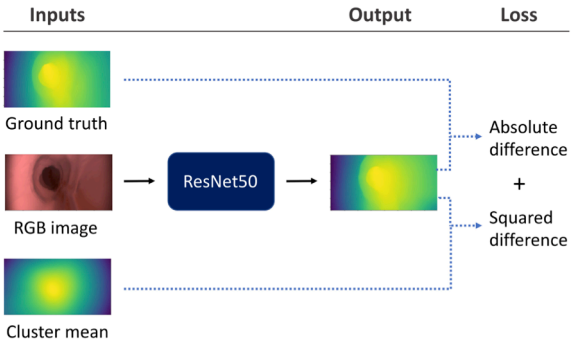

Figure. 3. Network architecture for depth estimation combining the ResNet-50 output to a classifier.

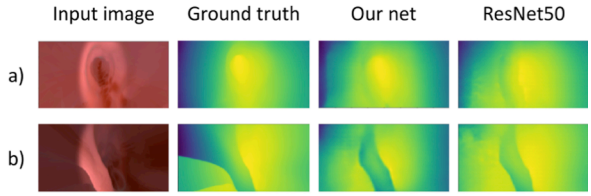

Figure. 4. Comparison results between the output of our model and the direct output of ResNet-50.

CONCLUSION AND DISCUSSION

In this paper we present a method that allows us to train a Convolutional Neural Network to predict depth from a single image without the need of real ground truth data. We describe a procedure to create training data in a virtual environment and propose a network architecture that attempts to limit the drawbacks of using synthetic data by enforcing shape consistency. However, our underlying shape model is trivial and fails to cover the variety of shapes found in the test set, which results in a mean distance of several millimeters between ground truth and prediction. Our future work will focus on a more elaborate statistical model incorporating the joint distribution of nearby pixels instead of considering each pixel independently.

REFERENCES

[1] Van Rijn, Jeroen C., et al. “Polyp miss rate determined by tandem colonoscopy: a systematic review.” The American Journal of Gastroenterology 101.2 (2006): 343.

[2] Mahmood, Faisal, and Nicholas J. Durr. “Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy.” Medical Image Analysis (2018).

[3] Visentini-Scarzanella, Marco, et al. “Deep monocular 3D reconstruction for assisted navigation in bronchoscopy.” CARS (2017).

[4] Laina, Iro, et al. “Deeper depth prediction with fully convolutional residual networks.” 3DV IEEE (2016).

[5] Godard, Clément, Oisin Mac Aodha, and Gabriel J. Brostow. “Unsupervised monocular depth estimation with left-right consistency.” CVPR (2017).

[6] He, Kaiming, et al. “Deep residual learning for image recognition.” CVPR (2016).

[7] Laina, Iro, et al. "Deeper depth prediction with fully convolutional residual networks." 3DV IEEE (2016).