Widening Siamese Neural Networks for Stereo Matching in Colonoscopy

P. Brandao[1], E. Mazomenos[1], A. Rau[1], M. Janatka[1], D. Stoyanov[1]

[1]*Centre for Medical Imaging, University College London*
patrick.brandao.15@ucl.ac.uk

INTRODUCTION

Colorectal cancer is the third most common cancer worldwide [1] and, although the fourth deadliest form of cancer, early diagnosis can increase the survival rate significantly. As the quality of colonoscopy diagnosis highly depends on the proficiency of the operator, computer-aided systems have been researched aiming to improve the detection rate of polyps during the procedure. While segmentation algorithms have improved the accuracy of polyp detection on observed surfaces, 3D reconstruction of the colon has the potential to unveil unscanned areas and to provide instructions to the operator guiding the colonoscopy in a manner that guarantees the observation of the entire inner surface of the colon.

3D reconstruction from stereo images has been actively researched for many decades. Recent work has focused on improving stereo matching cost functions which describe the correspondence between two pixels algebraically. However, the performance of stereo algorithms depends on the choice of the cost-function, which might not be ideal for different environment conditions. Despite the progress in the field, accurately finding stereo correspondences in colonoscopy/surgery images is still very challenging because of occlusions, reflective surfaces, repetitive patterns, and textureless or low detail regions-.

Recent research has aimed to discard hand-crafted cost-functions and hand-selected parameters entirely. First approaches proposed to train Convolutional Neural Nets (CNNs) to compare two input images and output a similarity measure [2] or to directly output a depth map [3].

We apply similar Deep Learning methods but with small considerations for colonoscopy. We aim to improve the 3D reconstruction of the colon so we can provide better diagnostic information in clinical settings.

MATERIALS AND METHODS

We construct our stereo matching network with a two branch Siamese architecture by layering 7 blocks of 3 × 3 2D convolutions with 64 neurons, batch normalization and a Rectified Linear Unit (ReLU). The parameters in both branches are shared. The last layers are added without batch normalization and ReLU operations.

Endoscopic procedures present challenges for stereo matching due to the lack of discriminative features in the environment. We address this issue by incorporating pooling operations within our model, which allows a much wider global receptive field [4]. We use two max-pooling layers stacked between every other pair of convolution blocks, widening the global receptive field from 15 to 28 pixels. This allows the CNN to extract more visual cues and helps more accurate matching, especially in textureless regions or areas of aperture problems. We compensate the loss of detail from down-sampling by using transpose convolution (deconvolution) layers. We implement the same amount of 2 strided 3 × 3 deconvolutions as the number of max-poolings before computing any correlation metric. An illustration of our architecture is presented in Figure 1.

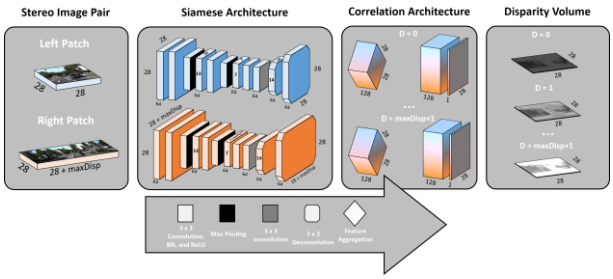

Figure 1 - Representation of our 7-layered stereo matching CNN.

We use an inner product layer as a correlation metric between feature vectors extracted from the Siamese branches like Luo et al. [3]. The operation is computationally efficient, fast and differentiable, which allows backpropagation during training. This layer allows the CNN to learn feature extractors that minimize the inner product of two corresponding points.

We train our models with stereo image pairs from the KITTI datatset [5], where the true displacement of a sparse number of pixels is known. We randomly extract small patches from the left and the right image. This allows to sample diverse training batches while being memory efficient. We treat each disparity value as a mutually exclusive classification problem. The values outputted from the correlation step are used to compute a softmax loss. All parameters are learned with stochastic gradient descent and gradients are backpropagated using the standard Adam optimization [6].

The proposed CNN was trained for 75K iterations with an initial learning rate of 1e−3. We sampled batches by randomly extracting 32 28 × 28 patches from the left image and 32 28 × 156 patches from the right one. The CNN model was implemented in Tensorflow [7] and run on a NVIDIA Titax-X GPU.

The final disparity map computed by the CNN can be used with standard stereo point triangulation to project the 2D images to a 3D space.

RESULTS

As a preliminary study, we validate our method using a PVA-C colon phantom manufactured from a 3D model

of a human colon. Data was acquired using a stereo camera from a *da Vinci Surgical System*. The camera was calibrated and stereo image pairs were rectified before being processed by our CNN.

An example of a 3D reconstruction using the disparity map computed with our Siamese architecture is presented in Figure 2.

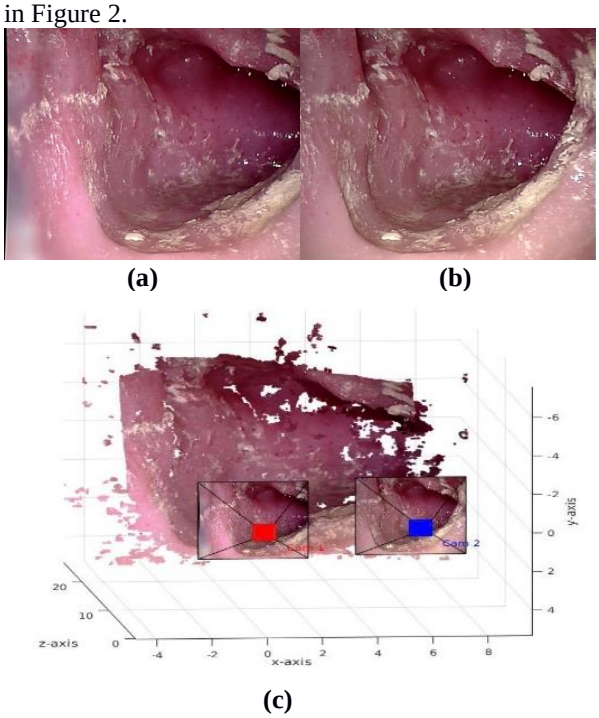

Figure 2 - (a) left and (b) right stereo image. (c) Example of 3D reconstruction of a stereo image pair using a disparity map computed by our CNN.

CONCLUSION AND DISCUSSION

Acquiring ground truth for stereo reconstruction of our colon phantom is complicated by spatial restrictions. Because of this, we limit our analysis to a qualitative evaluation.

Figure 2 shows that the 3D reconstruction faithfully reflects the original phantom environment. The tubular structure is maintained and the shape of the visible folds is also present. Considering that every single pixel is matched individually, without any type of spatial regularization or post-processing, a very dense and cohesive point cloud was obtained. This shows how powerful our Siamese architecture is for the image matching task. This is even more impressive considering that the model was trained to match natural outdoor street view images.

Despite the positive result, a significant amount of points are incorrectly reconstructed. Again, because no spatial consistency is guaranteed, pixels that are mismatched create a series of isolated points in the 3D space. Moreover, because the CNN outputs a dense disparity map, occluded pixels are also reconstructed. This creates gaps in the colon surface and a miss-reconstructed "cloud" of points corresponding to occluded pixels. One example of this is the occluded area from the fold in the right top corner.

Standard post-processing pipelines that minimize spatial irregularities can still be applied to our method in order to minimize the reconstruction error. Furthermore, our model can also be expanded to allow spatial regularization to be learned from data, following the same principle suggested by Kendall et al. [8].

In our future work, we intend to develop an unsupervised training method that would allow matching by learning environment specific features, independently of the existence of ground truth, one of the main limitations in clinical applications.

REFERENCES

[1] Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. CA: a cancer journal for clinicians, 65(2), 87-108.

[2] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research, 17(1-32) 2016.

[3] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5695–5703, 2016.

[4] Haesol Park and Kyoung Mu Lee. Look wider to match image patches with convolutional neural networks. IEEE Signal Processing Letters, 2016.

[5] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[6]Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980,2016.URL:http://arxiv.org/abs/1412.6980.

[7] Martín Abadi, Ashish Agarwal, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

[8]Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. arXiv preprint arXiv:1703.04309, 2017.