Weakly Supervised Airway Orifice Segmentation in Video Bronchoscopy

Ron Keuth\textsuperscript{a}, Mattias Heinrich\textsuperscript{a}, Martin Eichenlaub\textsuperscript{b}, and Marian Himstedt\textsuperscript{a}

\textsuperscript{a}Institute of Medical Informatics, University of Lübeck, Germany
\textsuperscript{b}University Heart Center Freiburg-Bad Krozingen, Germany

ABSTRACT

Video bronchoscopy is routinely conducted for biopsies of lung tissue suspected for cancer, monitoring of COPD patients and clarification of acute respiratory problems at intensive care units. The navigation within complex bronchial trees is particularly challenging and physically demanding, requiring long-term experiences of physicians. This paper addresses the automatic segmentation of bronchial orifices in bronchoscopy videos. Deep learning-based approaches to this task are currently hampered due to the lack of readily available ground truth segmentation data. Thus, we present a data-driven pipeline consisting of a $k$-means followed by a compact marker-based watershed algorithm which enables to generate airway instance segmentation maps from given depth images. In this way, these traditional algorithms serve as weak supervision for training a shallow CNN directly on RGB images solely based on a phantom dataset. We evaluate generalization capabilities of this model on two in-vivo datasets covering 250 frames on 21 different bronchoscopies. We demonstrate that the model is capable to transfer its knowledge to the unseen in-vivo domain, reaching an average error of 4.35 vs 7.98 pixels for detected centroids of airway segmentations by an image resolution of $128 \times 128$. Our quantitative and qualitative results indicate that in the context of video bronchoscopy, phantom data and weak supervision using non-learning-based approaches enable to gain a semantic understanding of airway structures.

Keywords: Video Bronchoscopy, Weakly Supervision, Airway Orifice Segmentation, CNN

1. INTRODUCTION

Video Bronchoscopy (VB) is commonly applied in conjunction with lung diseases. It is a fundamental procedure for diagnosis of lung cancer, enabling biopsy of deep airway tissue.\textsuperscript{1} In addition to that, VB is routinely conducted for monitoring Chronic Obstructive Pulmonary Disease (COPD) patients and clarification of acute respiratory problems at Intensive Care Units (ICU). The navigation within the bronchial tree is challenging and physically demanding for physicians due to homogenous textures and perceptually similar appearance of bronchial orifices and requires long-term experiences.\textsuperscript{2} This is particularly the case without prior CT scans and Electromagnetic Tracking (EMT) systems at ICUs. Airway orifice segmentation which is the main objective of this paper enables image-based guidance, e.g. by providing graphical overlays on top of the VB images. In conjunction with EMT or image-based tracking\textsuperscript{3} these overlays can be accompanied by airway labels w.r.t. generic or interpatient lung models which, however, is not addressed by the approach presented in this paper. The variety of tissue appearance, illumination, image artifacts, secrete and patient anatomy poses a challenge for airway segmentation which we aim to address by deep learning-based approaches. However, this is currently hampered due to a lack of readily available ground truth labels motivating the incorporation of traditional (non-learning-based) methods as weak supervision. In particular, we incorporate an airway phantom dataset collection accompanied by ground truth depth images to generate airway orifice labels for training a deep learning-based segmentation model. Our proposed methods are also developed with a focus on their complexity and runtime, keeping them real-time capable even on low-end devices for intervention guidance.

Further author information: E-mail: ron.keuth@student.uni-luebeck.de, marian.himstedt@uni-luebeck.de
2. RELATED WORK

The illumination model of Phong has been used for an approach considering only the RGB image of the bronchoscopy. Its assumption that the intensity value decreases quadratically with the distance to the light source of the endoscope allows the extraction of intensity- and geometric-based features. The constructed feature space is then clustered by a $k$-means allowing to classify a bronchoscopy image into lumen, wall structure and foreground. Although, this method produces fairly good results considering only the RGB image as an input, it is sensitive to common artifacts during a bronchoscopy such as bubbles or secret.

To overcome this problem, the approach by Wang et al. relies on depth images as abstract scene representations. The depth image is binarized with a threshold using maxima detection. The promising results of this method inspired us to use data-driven methods for generating orifice segmentations given depth images. However, we found that the generated segmentations on our phantom dataset tend to underestimate the airway orifice, and therefore we developed our own method. One drawback of such depth-based methods is the need of a depth image. In the context of bronchoscopy, such a depth image cannot be obtained natively via a stereo camera due to the hardware limitation and therefore has to be generated directly from the RGB image. Wang et al. implemented the complex non-linear operation of the domain translation with a Generative Adversarial Networks (GAN). However, the literature emphasizes that the use of GANs without conditioning brings the risk of mode collapse. In the context of bronchoscopy, this could result in incorrect anatomies in the unsupervised generated depth image. To avoid this kind of risk, we decided to directly train a Convolutional Neural Network (CNN) on the RGB images using the airway orifice segmentation extracted from the depth image as ground truth.

3. METHODS

3.1 Datasets

We utilize three datasets for training, validation and testing our method: The dataset Phantom-DS consisting of about 30k RGB and depth images captured within a simplified silicone model. For in-vivo evaluation, we use 125 samples of 20 different bronchoscopies with their annotated segmentations from the public dataset CVC-DS and about 100 frames of the private dataset HAUSER-DS with expert annotations.

3.2 Data-Driven Methods for Instance Orifice Segmentation

Fig. 1 shows the different steps of our proposed pipeline to generate an instance orifice segmentation map from a given depth image. A $k$-means determines the two classes ($k = 2$) of airway and other tissue, considering only the depth distribution. The obtained global airway labels are then used to generate a binary segmentation map and to define the region of interest for the next stage. For determining different airway instances, we low-pass filter the depth image using an efficient box filter ($3 \times 3$ average pooling with kernel size=3) followed by a non-maximum suppression, where the peaks have to be 5% of the image resolution apart from each other. These peaks define the markers for the compactness marker-based watershed algorithm, which runs on the inverted depth image as input. The watershed perfectly models the nature of the instance airway segmentation problem, allowing different depth values of adjacent airways and let their segmentation flow smoothly into each other. The result of the watershed is finally composed together with the global labels obtained by the $k$-means to an instance orifice segmentation map.

The pipeline is real-time capable, running with roughly 130 Hz on a laptop CPU (INTEL i5-7200U). One current limitation of our pipeline is the segmentation of at least one orifice, even if it is a false positive. However, such false-positives tend to cover an unusual huge area, the problem can be solved by defining a heuristic such as a relative threshold over the area covered by one orifice instance or an absolute threshold could be set empirically determining a minimum depth value range for the presence of an airway.
3.3 CNN Architecture and Training

For this paper, we solely focus on the binary orifice segmentation to enable the use of a Lite Reduced Atrous Spatial Pyramid Pooling (Lite R-ASPP)\(^9\) as an efficient CNN architecture for segmentation. We use an encoder pretrained on ImageNet for training. The high texture and illumination variety from the synthetic to the in-vivo VBs introduces a domain gap. To narrow this gap, we apply multiple data augmentation methods. Therefore, we follow a guideline\(^10\) for an intensity value augmentation of the RGB images, randomly choosing transformations like color jitter, quantization and histogram equalization. In in-vivo VB operators frequently rotate endoscopes during navigation, making rotation invariance a real-world requirement for our methods. We achieve that by rotating the images by radians randomly sampled from \([0, 2\pi]\).\(^2\) To prevent the model’s parameter from becoming too complex and thus overfitting the synthetic data, we limit the amount of training epochs to benefit from early stopping, use a weighted cross entropy with label smoothing\(^11\) and weight decay with its AdamW implementation\(^12\) during training. We also found gray scaling the input helps to narrow the domain gap to in-vivo data.\(^13\)

3.4 Evaluation

A well established method to evaluate the overlapping of two segmentations is the Dice Similarity Score (DSC). However, the DSC alone has only limited significance in our context. This is because an airway’s orifice has, unlike e.g. a liver, no clear organ boundaries, resulting in a high inter and also intra observer variability in the in-vivo segmentation ground truths. Fig. 2 visualizes this observer variability in a simulation with different sized segmentations.

As a solution, we also consider the distances of the first moments (centers of gravity) of the individual airway orifice instance segmentations. Moments are scale invariant and therefore well-suited for this use case. We inject the distances of the first moments into our Average Minimum Centroid Distance (AMCD) metric as followed: Having \(N\) airways with their ground truth segmentations and \(M\) predicted segmentations, we calculate their first moments in \(C \in \mathbb{R}^{N \times 2}\) and \(\hat{C} \in \mathbb{R}^{M \times 2}\) respectively. First, we determine the distance \(d_{c_i}\) for each centroid of the ground truth \(c_i\) to the nearest centroid of the prediction \(\hat{c}_j\) using the Euclidean distance

\[
d_{c_i} = \min_{j=1,\ldots,M} ||c_i - \hat{c}_j||_2
\]
with $i \in N$. Then, we obtain the AMCD for the overall image by the mean over the minimal distances of all centroids of the ground truth

$$AMCD = \frac{1}{N} \sum_{i \in N} d_{c_i}.$$  

(2)

Note that the symmetry of the Euclidean distance is lost during the aggregation with the minimum function. Our implementation punishes false-negatives (see Eq. 1) and therefore considers only the recall.

Even we explained the limited significance of the DSC in our context, we decided to include it due to its scientific importance.

4. RESULTS AND DISCUSSION

4.1 Data-Driven Methods for Instance Orifice Segmentation

The evaluation of the proposed method to generate an instance orifice segmentation can only be qualitative due to the lack of paired depth images and segmentation masks. Fig. 3 shows positive examples, which allow the conclusion that depth-based $k$-means can successfully generate the binary segmentation of an airway orifice. The local maximum detection on the smoothed depth image also works robustly and even successfully detects difficult edge cases like it is visible in the last row of Fig. 3. With the local maxima as markers, the marker-based watershed is capable of discriminating adjacent airway instances from each other with their border on the carina of the bifurcation. However, the quality of segmentation is highly reliant on the quality of the depth image and e.g. produce false-negatives in rare cases when the depth profile of an airway is underestimated due to the camera perspectives (see Fig. 4).

4.2 Weakly Supervised Training

The qualitative results in Fig. 5 demonstrate that the model trained on the phantom dataset was able to gain semantic knowledge about airways from noisy ground truth generated based on depth images using our proposed data-driven pipeline. In some cases, it even outperforms this noisy ground truth, particularly when airways were not detected properly due to their low depth profiles (see two left columns in Fig. 5a). The false-positives at the edges in the right column of Fig. 5a are caused by the masked cross entropy loss function, which omits the region outside the circular image during the loss calculation. The data augmentation method was able to close the domain gap from the phantom to the in-vivo datasets, when considering its qualitative results in Fig. 5b and Fig. 5c. Unfortunately, the CNN is incapable of overcoming the limitation of our data-driven pipeline always detecting at least one airway orifice, resulting in false-positive at cases without an airway orifice in the in-vivo data. But we are optimistic that this problem can be solved by implementing one of the described approaches for preventing such false-positives by the data-driven pipeline and increasing the quality of the ground truth in such situations.

Our quantitative results shown in Tab. 1 have only limited significance for the Phantom-DS due to the lack of real ground truth, which is compensated by our automatically generated noisy ground truth. However, the
Figure 3: Good examples for orifice segmentation from given depth images using the proposed data-driven method. From left to right: the RGB image, the smoothed depth images with their local maxima (red marker), the obtained binary segmentation and the airway instance segmentation. The region of the orifice is successfully discriminated from other tissue. Our method is also capable of separating adjacent airways with their segmentation borders on the carina.

Figure 4: Negative example for orifice segmentation from a given depth image using the proposed data-driven method. From left to right: the RGB image, the smoothed depth images with their local maxima (red marker), the depth image as point cloud with the labels of the $k$-means and the obtained binary segmentation map. The very left airway (black arrow) is not detected due to its high angle to endoscope perspective, resulting in low depth values.

Table 1: Quantitative results of the model trained on the Phantom-DS. We use the Dice Similarity Score (DSC) and the Average Minimum Centroid Distance (see Eq. 2) in pixel within the image resolution of $128^2$. Please remind the limited significance of the DSC in our context (see Sec. 3.4).

| dataset             | AMCD      | DSC       |
|---------------------|-----------|-----------|
|                     | median    | median    |
| Phantom-DS[test]    | 8.381±12.02 | 4.350 | 0.785±0.163  | 0.840 |
| CVC-DS              | 7.326±5.839 | 5.614 | 0.646±0.189  | 0.671 |
| HAUSER-DS           | 14.442±11.904 | 10.346 | 0.532±0.169  | 0.515 |
Figure 5: Qualitative results of the Lite R-ASPP. The red masks are the predicted segmentation, the blue ones the ground truth. The ground truth of the Phantom-DS is generated from the corresponding depth images using our proposed data-driven method. All images are shown in color for a better visualization, regardless of the gray scaled input of the Lite R-ASPP. See Sec. 4 for a detailed description of the results.
ground truth by the in-vivo dataset was created by human experts and is, beside the high observer variability, reliable. The model trained on the Phantom-DS shows a great domain robustness, considering the resulting AMCD, which is for the in-vivo CVC-DS in the same range as the Phantom-DS. The results indicate that the domain gap to the HAUSER-DS is significantly higher due to its interlacing artifacts (barely noticeable in Fig. 5c) caused by its old, analogous recording technic. The comparison of experimental results on the Phantom-DS and CVC-DS prove the limited significance of the DSC in our context and the motivation for our AMCD. Even the DSC is 10 points lower on the CVC-DS, the AMCD is in a comparable range, going hand in hand with the correct prediction visible in the qualitative results. This is due to the observer variability of the ground truth segmentation, resulting in different diameters of the segmentations depending on the observer.

Considering the quantitative and especially the qualitative results, the training on the Phantom-DS with the noisy ground truth as weak supervision enables the learning of semantic knowledge of airways. As shown in Fig. 5 we demonstrated that this knowledge can be transferred to in-vivo data.

5. CONCLUSION

In this work, we presented a real-time capable pipeline that extracts airway segmentations from a given bronchoscopy depth image using efficient data-driven classical methods serving as supervision for training solely on RGB images. However, this method has some disadvantages: On the one hand it requires a depth image, which is not natively provided by the endoscope due to hardware limitations and therefore has to be generated via a complex non-linear domain translation like a GAN. On the other hand, a data-driven approach is not equal to a semantic understanding. Considering this and due to the lack of robustness to some edge cases, we consider this pipeline alone as not suitable for real-world applications. However, paired with the RGB image of the bronchoscopy, these generated segmentation maps can be used as weak supervision during training of a shallow CNN for airway orifice segmentation. We showed that this model being trained on phantom data gains robust semantic knowledge of airway structures overcoming noisy ground truth on edge cases and is even applicable directly to in-vivo VB due thanks to a substantial data augmentation. With all of this, our proposed method allows the generation of segmentation masks directly on RGB images without the need of manually annotated datasets. We argue that this direct prediction from RGB images is superior to the segmentation approaches on the depth images because it comes without the risks of a domain translation from RGB to depth via GANs, which is mainly due to on the unsupervised manner of the GAN training, which likely causes the generation of wrong anatomies like additional or absent airway branches in the synthesized depth images.

REFERENCES

[1] Ost, D. E., Ernst, A., Lei, X., Kovitz, K. L., Benzaquen, S., Diaz-Mendoza, J., Greenhill, S., Toth, J., Feller-Kopman, D., Puchalski, J., Baram, D., Karunakara, R., Jimenez, C. A., Filner, J. J., Morice, R. C., Eapen, G. A., Michaud, G. C., Estrada-Y-Martin, R. M., Rafeq, S., Grosu, H. B., Ray, C., Gilbert, C. R., Yarmus, L. B., and Simoff, M., “Diagnostic Yield and Complications of Bronchoscopy for Peripheral Lung Lesions. Results of the AQuIRE Registry,” American Journal of Respiratory and Critical Care Medicine 193, 68–77 (Jan. 2016).

[2] Yoo, J. Y., Kang, Y., Park, J. S., Cho, Y.-J., Park, S. Y., Yoon, H. I., Park, S. J., Jeong, H.-G., and Kim, T., “Deep learning for anatomical interpretation of video bronchoscopy images,” Scientific Reports | 11, 23765 (2021).

[3] Wang, C., Oda, M., Hayashi, Y., Villard, B., Kitasaka, T., Takabatake, H., Mori, M., Honma, H., Natori, H., and Mori, K., “A visual SLAM-based bronchoscope tracking scheme for bronchoscopic navigation,” International Journal of Computer Assisted Radiology and Surgery 15, 1619–1630 (Oct. 2020).

[4] Sánchez, C., Bernal, J., Gil, D., and Sánchez, F. J., “On-Line Lumen Centre Detection in Gastrointestinal and Respiratory Endoscopy,” in [Clinical Image-Based Procedures. Translational Research in Medical Imaging], Erdt, M., Linguraru, M. G., Oyarzun Laura, C., Shekhar, R., Wesarg, S., González Ballester, M. A., and Drechsler, K., eds., 31–38, Springer International Publishing, Cham (2014).
[5] Wang, C., Hayashi, Y., Oda, M., Kitasaka, T., Honma, H., Takabatake, H., Mori, M., Natori, H., and Mori, K., “Bronchial orifice segmentation on bronchoscopic video frames based on generative adversarial depth estimation,” in [*Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*], Linte, C. A. and Siewerdsen, J. H., eds., 11598, 20, SPIE (2021). ISBN: 9781510640252.

[6] Shin, Y., Qadir, H. A., and Balasingham, I., “Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance,” *IEEE Access* 6, 56007–56017 (2018).

[7] Visentini-Scarzanella, M., Sugiura, T., Kaneko, T., and Koto, S., “Deep monocular 3D reconstruction for assisted navigation in bronchoscopy,” *International Journal of Computer Assisted Radiology and Surgery* 12, 1089–1099 (July 2017). Publisher: Springer Verlag.

[8] Neubert, P. and Protzel, P., “Compact Watershed and Preemptive SLIC: On Improving Trade-offs of Super-pixel Segmentation Algorithms,” in *2014 22nd International Conference on Pattern Recognition*, 996–1001, IEEE (Aug. 2014). ISSN: 1051-4651.

[9] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., and Adam, H., “Searching for MobileNetV3,” in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 1314–1324 (Oct. 2019). arXiv: 1905.02244v5.

[10] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V., “Autoaugment: Learning augmentation strategies from data,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019*, 113–123 (June 2019). arXiv: 1805.09501v3 ISBN: 9781728132938.

[11] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., “Rethinking the Inception Architecture for Computer Vision,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-Dec*, 2818–2826 (2016). arXiv: 1512.00567 ISBN: 9781467388504.

[12] Loshchilov, I. and Hutter, F., “Decoupled weight decay regularization,” in *7th International Conference on Learning Representations, ICLR 2019*, (2019). arXiv: 1711.05101.

[13] Sganga, J., Eng, D., Graetzel, C., and Camarillo, D. B., “Autonomous Driving in the Lung using Deep Learning for Localization,” (July 2019). arXiv: 1907.08136.