Anatomical landmark segmentation in uterine cervix images using deep learning

Guo, Peng, Xue, Zhiyun, Long, L. Rodney, Antani, Sameer

Peng Guo, Zhiyun Xue, L. Rodney Long, Sameer K. Antani, "Anatomical landmark segmentation in uterine cervix images using deep learning," Proc. SPIE 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 1131810 (2 March 2020); doi: 10.1117/12.2549267

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Anatomical Landmark Segmentation in Uterine Cervix Images Using Deep Learning

Peng Guo, Zhiyun Xue, L. Rodney Long, Sameer K. Antani
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine,
National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, USA 20894

ABSTRACT

For automated evaluation of changes on uterine cervix, the external os (here simply os) is a primary anatomical landmark in locating the transformation zone (T-zone). Any abnormal tissue changes typically occur at or within the T-zone. This makes localizing the os on cervical images of great interest for detecting and classifying changes. However, there has been very limited work reported on segmentation of the os region in digitized cervix images, and to our knowledge no work has been done on sets of cervix images acquired from independent data collections exhibiting variabilities due to collection devices, environments, and procedures. In this paper, we present a process pipeline which consists of deep learning os region segmentation over such multiple datasets, followed by comprehensive evaluation of the performance. First, we evaluate of two state-of-the-art deep learning-based localization and classification algorithms, viz., Mask R-CNN and Mask^X R-CNN, on multiple datasets. Second, in consideration of the os being small and irregularly-shaped, and of the variabilities in image quality, we use performance measurements beyond the commonly used DICE/IoU scores. We obtain higher performance, on a larger dataset, as compared with the work reported in the literature, and achieve a highest detection rate of 99.1% and an average minimal distance of 1.02 pixels. Furthermore, the network models we obtained in this study show potential use of quality control for data acquisition.

Keywords: Deep learning, uterine cervical cancer, external os segmentation, automated visual evaluation, Mask-RCNN, Mask^X R-CNN

1. INTRODUCTION

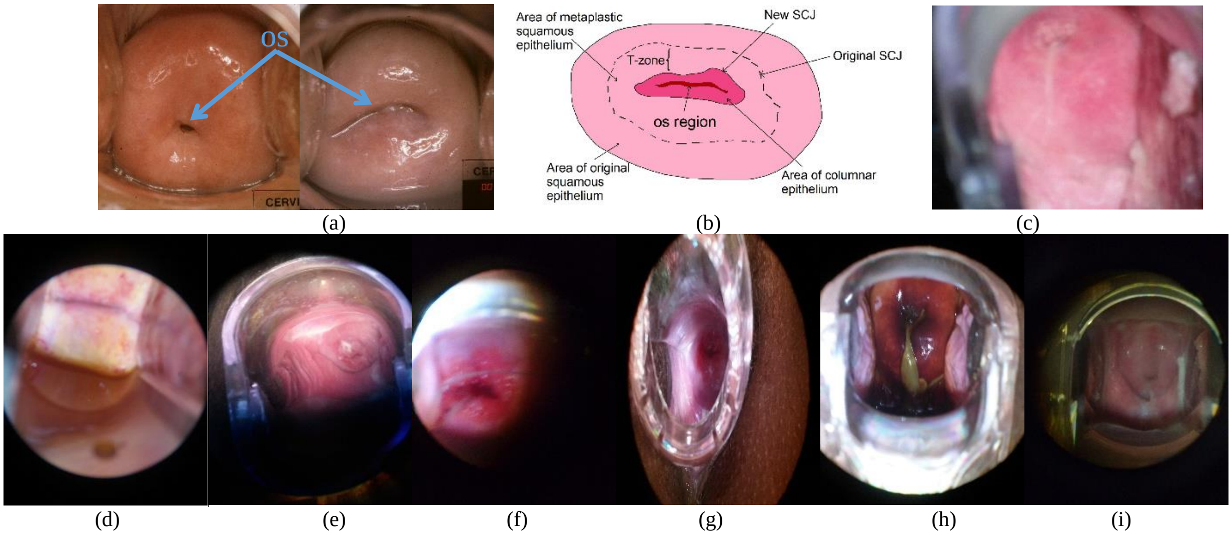
As the fourth most frequent cancer in women, there were 570,000 new cases of cervical cancer in 2018, according to the World Health Organization (WHO) [1]. There are several cervical cancer/pre-cancer early detection (screening) methods, among which VIA is considered to be simpler and less expensive compared with the other screening modalities. However, VIA has rather inter-observer agreement; in [2] it was reported that only a 56.8% complete decision agreement was reached among 20 gynecologists over 948 cases, for example.

In previous work we carried out a deep learning based automatic screening process for digitized cervix images, called Automatic Visual Evaluation (AVE) [3]. This method shows potential to improve or replace visual assessment with acetic acid (VIA). However, AVE performance over sets of cervix images acquired from varying geographical locations, with potentially varying imaging devices, environmental factors, and imaging procedures, could be highly variable. These datasets may differ with respect to image quality factors such as absence of the anatomical region of interest (cervix), or other factors such as illumination or focus; this in turn could lead to erroneous assessments. In our recent work addressing automatic assessment of image focus, inability to determine the cervix region is one of the major sources of error in [4]. To resolve this problem, the os provides a valuable anatomical fiduciary in cervix images that can be used to support or enhance the cervix localization. The os is usually clearly visible as the opening in the center of the ectocervix (cervical canal) (Figure 1a). Accurate segmentation of the os can be helpful in overcoming previously reported error cases. Furthermore, the fine contour of the os plays an important role in identifying the transformation zone (T-zone) which is of great importance in cervical cancer screening [5]. As shown in Figure 1b, the T-zone is defined as the area between the new squamocolumnar junction (SCJ) and the original SCJ which is located 1) at or very close to the os at birth during premenarchar years; or 2) at variable distances around the os during reproductive age. As important anatomical landmarks, they are of great significance 1) in cervical cancer screening for defining the cervix type [6]; and, 2) are considered for determining the viability of cervix images for use in AVE [3], since images with non-visible T-zone are not considered for AVE processing.

Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications,
edited by Po-Hao Chen, Thomas M. Deserno, Proc. of SPIE Vol. 11318, 1131810

However, automated segmentation of the os region over datasets sourced from different providers is challenging for several reasons: 1) the os region or even the entire cervix looks different from image to image; within a dataset captured at a single geographic locations this could be caused by configuration differences of image-capture devices, variabilities in following operational procedures among field workers, variabilities in illumination at the specific time of collection, and other reasons; for images in different datasets (collected at different geographical sites), the image collection devices themselves may be different, and there may even be visual appearance differences due to differences in cervix microbiota between geographical sites. Some of this variability is illustrated below (Figure 1c-f). 2) The size and shape of the os varies in a single person with age, child-carrying history and hormonal state. It could appear to be a small circular opening, or a wider and more elongated irregular shape (Figure 1a). 3) Distractors in the image can also affect segmentation accuracy of the os region. These include: presence of pubic hair, intra-uterine devices or the speculum, blurriness, or lighting issues (Figure 1d-i).

There are a limited number of technical publications on os region segmentation. One research group reported an average minimal distance of around 10 pixels from automatically computed os and manually marked os boundaries, for 87/101 (87%) cervix images, using a geometric measure of local concavity to identify os [7]. In [8] the same algorithm was applied and the same measurement of 10-pixel minimal distance was obtained on two extended datasets of 123/158 (79%) and 99/120 (83%) cervix images. In [9], a Harris detector was implemented with a sliding window approach, where features of anatomical landmarks were extracted and matched between different images. The purpose of the study [9] was to do registration of os regions between images, the quantitative results were reported as an overall landmark matching rate (but not obtained specifically from the os region segmentation). Additionally, some peripheral research efforts were reported in [10, 11, 12] on applying feature-based segmentation of ROIs (region of interests) in cervix images, however the specific topic of os segmentation with its standalone evaluation on multiple, heterogeneous datasets was not addressed.


OS
Area of metaplastic squamous epithelium
New SCJ
T-zone
Original SCJ
os region
Area of original squamous epithelium
Area of columnar epithelium

(a) (b) (c)

(d) (e) (f) (g) (h) (i)

Figure 1: Examples of digitized cervix images. (a) example of os in different shapes. (b) illustration of cervix anatomy and definition of T-zone. (c-i) cervix images with variability in quality and os visibilities.

Following successes in application of deep learning to medical imaging [3, 4, 13], the goal of this work is to apply the techniques to automatic segmentation of the os region. In this work, we use three heterogeneous datasets collected with a variety of image acquisition devices and from different geographical regions. In general, large intra- and inter-dataset variety can be found in our datasets with respect to lighting, image focus, ROI color, visibility, size, shape irregularity and other factors. Our goal is to create an os segmentation method that is robust across such datasets. We split these three datasets into *strongly annotated* (that is, having contour ground truth) and *weakly annotated* (with only ROI bounding box ground truth) data. We apply two strategies: 1) we train the Mask R-CNN architecture [14] on our strongly annotated data. As a strongly supervised method [15], Mask R-CNN requires its training data to be strongly annotated; 2) then we also train a "partial-supervision" architecture Mask^X R-CNN [15] which can take both strongly and weakly annotated data as input. This network is trained in two stages: in the first stage, the weakly annotated data is used and, in the second stage, the strongly-annotated data.

In our performance evaluation, we provide classic DICE coefficients and Intersection-over-Union (IoU) scores and additional measures which may be more appropriate for this segmentation task, including: 1) detection rate: the number of valid predictions that the model can make in a set of test images, 2) centroid distance: the distance between the predicted region centroid and the ground truth centroid, and 3) minimal distance: also used in [7, 8], to described the minimal distance between the automatically marked and manually marked contours. We quantitatively evaluate the segmentation prediction obtained from both models by comparing all the scores calculated. Based on the obtained experimental results, Mask R-CNN with ResNet101 [16] backbone is the best performing model when all of the training data is strongly annotated, but Mask^X R-CNN yields comparable results by using just a subset of strongly annotated images which are used in training Mask R-CNN, plus weakly annotated images. The rest of the paper is organized as follows: Section 2 describes the details of datasets used in this study, Section 3 describes the two deep learning architectures; Section 4 and 5 present the experiments, the results and the discussion.

2. IMAGE DATA

In this work, we use three different datasets: the Kaggle dataset, the CVT dataset and the ALTS dataset. For each of the datasets, we prepare two groups of annotations, one strongly annotated with os region contours and a second weakly annotated group with the os delimited only with bounding box. For purposes of our experiments, we generated the bounding boxes (for which we had no data source) by plotting a rectangle enclosing the contour annotations (for which we had a data source). The reason of doing this is to meet the different training data requirements of Mask R-CNN (strongly supervised) and Mask^X R-CNN (weakly supervised) as described in Section. 1. Note that for each dataset, we use "_mask" to denote the mask (contour) annotations and "_box" to denote the bounding box annotations.

2.1 Kaggle Datasets

This dataset consists of 1899 images (1448 for training, 451 for testing), which were collected by the MobileODT company [17] using their Enhanced Visual Assessment (EVA) device designed for digital cervical imaging. This dataset was made available as part of a Kaggle competition for classifying images into three cervix types; the os region contours were created and made available by Fernandes [18]. These images in this dataset exhibit a wide variation in quality; we consider it to be the most challenging dataset in this work. Also, this dataset contains the largest number of images with os region contours annotated, thus we mainly trained both the Mask R-CNN and Mask^X R-CNN networks using images from this dataset. From this "Kaggle dataset" we derived our two subsets $A_{\_mask}$ and $A_{\_box}$.

2.2 CVT Datasets

The Costa Rica Vaccine Trial (CVT) is a longstanding collaboration sponsored and funded by the National Cancer Institute (NCI) (N01-CP-11005). It was supported by the National Institutes of Health (NIH) Office of Research on Women's Health and conducted in agreement with the Ministry of Health of Costa Rica [19, 20]. Women participants received the HPV16/18 vaccines, and were referred to colposcopy if the lesion persisted or high grade cytology was found at any time during the trial. The images in this dataset were acquired with permission obtained during enrollment in the study and digitized. Among them, 230 images were selected and marked with os region contours by a human expert. From this CVT set we derived our two subsets $B_{\_mask}$ and $B_{\_box}$.

2.3 ALTS Datasets

The images in this dataset refer to the atypical squamous cells of undetermined significance / low-grade squamous intraepithelial lesion (ASCUS/LSIL) Triage Study (ALTS), which was reported in [21]. The study was also conducted by the NCI, as a randomized clinical trial in the United States aiming at determining the optimal management plan for low-grade cervical abnormalities. From the original dataset, we took 120 images which have manual annotations which were labeled by medical experts in cervix oncology by drawing boundaries around the os region, using the Boundary Marking Tool (BMT) developed by the National Library of Medicine (NLM). Since the patient ID associated with each image is not available and randomly splitting the dataset would result in data leakage, we only use this dataset for test (not for training). From this ALTS set we derived our two subsets $C_{\_mask}$ and $C_{\_box}$.

3. METHODS

3.1 Mask R-CNN and Mask^X R-CNN

In this paper, we apply two state-of-the-art deep learning networks: Mask R-CNN [14] Mask^X R-CNN [15]. Both are designed to accomplish *object instance segmentation* in images and have been applied to public datasets such as COCO and obtained satisfactory performance [22]. The architectures of these two deep learning networks have some similarity at the network module level. However the "masks" (segmentation predictions) are accomplished differently, using different network structures.

In Mask R-CNN and most other methods for object instance segmentation, strongly annotated data is required. Thus the model(s) cannot be used to predict accurate target contours if only the object bounding boxes are available. However, the Mask^X R-CNN addresses this "data annotation gap" by taking advantage of both types of annotations, bounding box and mask (contour). In Mask^X R-CNN, the masks not only can be predicted after learning from the feature extraction module purely in the "mask branch" of the network, but also can be obtained by "converting" bounding boxes to contours by learning the necessary weight transformation. Thus we 1) train Mask R-CNN using images with os region contours, and 2) train Mask^X R-CNN using the same images which are trained in Mask R-CNN, but part with os region bounding boxes and part with contours. The details of both methods are presented next.

Mask R-CNN adopts the same two-stage procedure used in Faster R-CNN [23], in which: 1) the first stage consists of a Feature Pyramid Network (FPN) [24] that extracts the features in different scales for region proposals; this pyramid feature architecture may aid in improving accuracy and speed compared with single-scale feature map structures; and 2) the second stage consists of network modules that predict binary masks, bounding boxes and class labels of the objects based on those regional features obtained from the first stage. Moreover, a novel technique of *RoIAlign* [14] is applied to compute the exact values of the input features at four regularly sampled locations in each evenly split RoI bin. For the detailed description on Mask R-CNN, please refer to [14]. Since this supervised architecture requires strongly annotated images as training data, we focus on training and testing Mask R-CNN using datasets with os region contour ground truth.

Mask^X R-CNN is built also based on Faster R-CNN but adds a mask processing module which consists of a small convolutional neural network (weight transfer network) and a Multi-Layer Perceptron (MLP), together known as "mask head". The predictions of the mask head are fused by adding the two scores into a final output. Taking this network property into consideration, we train the Mask^X R-CNN in a two-stage manner. In the first stage, only the bounding box ground truth is used to train a Faster R-CNN. In the second stage, the mask head is trained with the bounding box head and feature layers remaining frozen. In other words, we use the full dataset with bounding box ground truth to train in the first stage and use contour annotations to train the mask head in the second stage in order to help the mask head in learning to convert the bounding box prediction weights to finer contour predictions. After the Faster R-CNN is trained in the first stage, we, in the second stage 1) use the contour annotations of the same images, in order to compare the performance of both network on the same image data; 2) use the contour annotations of an image subset of the training set in first stage, in order to observe the performance of Mask^X R-CNN of utilizing weakly annotated data. Also we investigate the number of contours used in our experiment for Mask R-CNN to observe the robustness of the model with respect to the size of strongly annotated data. For example, 1) we train the network using training split of $A_{\_box}$ in the first stage and then use all the $A_{\_mask}$ in the second stage; 2) we train the network using training split of $A_{\_box}$ in the first stage and then use a fraction of $A_{\_mask}$ in the second stage. Finally, we evaluate the network on the test splits of datasets for both training scenarios.

3.2 Implementation Details

We resized images to have shorter edge of 532 pixels, maintaining aspect ratio. We used mini-batch size of 2 images per GPU and an RoI mini-batch size of 512. For the base network we used ResNet101 [16] with one Nvidia GPU. In training, an output RoI is considered positive if it has IoU with a ground-truth box of at least 0.6; otherwise, it is considered negative. We trained the networks for 120,000 iterations with a learning rate starting at 0.001 and decaying by a factor of 10 at iterations 40,000 and 80,000. Initialization was done with ImageNet pre-trained weights. We applied only scaling augmentation.

4. EXPERIMENTS

4.1 Measurements

As performance measures we use: 1) DICE/IoU score, 2) detection rate (DR) 3) centroid distance (CD) and 4) minimal distance (MD). The widely-used DICE and IoU scores are the difference between the predicted mask(s) (region enclosed within the contour) and the ground truth mask(s). However, based on the examination of many os segmentation visual results and the corresponding DICE/IoU values, we found that 1) there are results which visually appear to be satisfactory but have low DICE/IoU scores (Figure 2a). Due to the small area of the os region, a small mis-prediction, sometimes even one single pixel off to the ground truth contour, would make a big difference on DICE/IoU score; 2) the os regions are labeled by human expert(s), which might be subjective; 3) for some images which have certain quality issues or in which the os region is hardly visible, the segmentations are unsatisfactory.

To deal with this problem, we included the three additional performance measures listed above. To calculate detection rate, a segmentation was considered "detected" if 1) the predicted mask had a network-output confidence value larger than 0.9; and, 2) the predicted mask had overlap with the ground truth mask (DICE or IoU > 0). In addition, "centroid distance" and "minimal distance" (adopted from [7]) were calculated, as illustrated in Figure 2b.

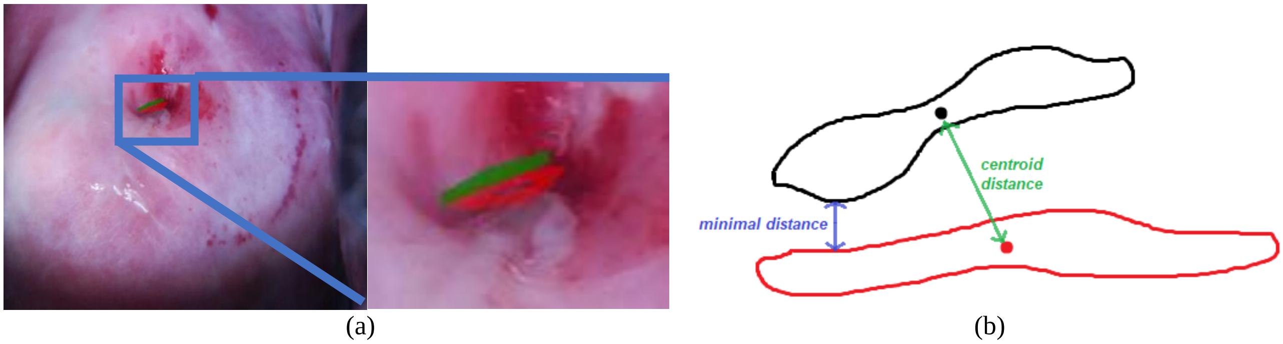


Figure 2: Example of segmentation and illustration of minimal distance (MD) and centroid distance (CD). (a) A segmentation prediction example with DICE score of only 0.01. The "red" contour denotes the prediction and the green contour denotes the ground truth. The prediction visually looks really good since it's very close to the ground truth, however, since the os region is of very small area and they do not intersect greatly, the prediction has a small DICE score which is commonly considered to be extremely bad segmentation. (b) The illustration of the centroid distance (CD) and minimal distance (MD) are calculated. The "black" contour and dot denote the predicted contour and its centroid, and the "red" contour and dot denote the ground truth contour and its centroid, respectively.

4.2 Experiments

We trained Mask R-CNN with $A_{\_mask}$ and tested each of the models on datasets of A, B and C. We used the following 5 training and testing scenarios: 1) Train with $A_{\_mask}$, test on test split of A; 2) train with $A_{\_mask}$, test on dataset B; 3) train with $A_{\_mask}$, test on dataset C; 4) train with $B_{\_mask}$, test on test split of A; and, 5) train with $B_{\_mask}$, test on test split of C. By observing the results of scenario 1 we can evaluate the performance of Mask R-CNN in fully supervised learning within the same dataset. To evaluate the performance of this network across datasets, we use the scenario 2-5 above. Results are presented in Table 1.

Table 1. Results for training Mask R-CNN with dataset A and B. Best scores are highlighted.

	Testing Dataset	*Training Dataset :* A			
		Avg. DICE/IoU	*DR*	*Avg. CD*	*Avg. MD*
1	A	0.36/0.24	0.891 (402/451)	12.71	7.78
2	B	0.35/0.23	**0.991 (228/230)**	**5.08**	**1.13**
3	C	**0.37/0.25**	0.983(118/120)	9.8	1.2
		Training Dataset : B			
4	A	0.22/0.14	0.712 (321/451)	42.54	18.00
5	C	0.40/0.27	0.967 (116/120)	13.79	2.47

The top part of Table 1 shows results of steps 1)-3): training on dataset A with strongly annotated data, and testing on datasets A, B, and C. The 2nd row shows that three of the highest scores were achieved when testing on dataset B.

The bottom part of Table 1 shows results of steps 4)-5): training on dataset B with strongly annotated data, and testing on datasets A and C. This testing on dataset A, trained with B (line 4 of Table 1), yielded results which were worse in all four measures, as compared to the results obtained by testing on dataset A, training with A (line 1 of Table 1). The testing on dataset C (line 5 of Table 1) trained on B was worse in three of the four scores, as compared to testing on C, trained on A (line 3 of Table 1); increased performance was only shown for DICE/IoU.

For Mask^X R-CNN we trained dataset A but with minor differences compared with what was done for Mask R-CNN. For example, for the training of 1), 2) and 3) above, we trained the first-stage Faster R-CNN with $A_{\_box}$ and second-stage mask heads with all the images in $A_{\_mask}$; then we tested the trained model on the test splits of A, B and C. We also performed quantitative evaluation of the results and compared each of these 3 testing results of Mask^X R-CNN with the corresponding results obtained from Mask R-CNN. By doing this, 1) we observe the performance difference between the two deep learning architectures on the same task; and, 2) we get a "sanity check" on the correctness of test results from the two architectures. Results obtained using Mask^X R-CNN are presented in Table 2.

Table 2. Results for training Mask^X R-CNN with dataset A. Best scores are highlighted.

Testing Dataset	*Training Dataset :* A			
	Avg. DICE/IoU	*DR*	*Avg. CD*	*Avg. MD*
A	0.31/0.20	0.894 (403/451)	13.21	9.44
B	0.35/0.23	**0.991 (228/230)**	**5.58**	**1.02**
C	**0.37/0.25**	0.983 (118/120)	9.93	2.3

We observed that: 1) by comparing Table 2 and the top half of Table 1, the testing results are very similar across network architectures (Mask R-CNN or Mask^X R-CNN). This is consistent with our expectation that, when all of the (strongly annotated) images in $A_{\_mask}$ are used to train the second stage of Mask^X R-CNN, the performance should be similar with that obtained from the Mask R-CNN model trained with all of these same images in $A_{\_mask}$; and, 2) the general performance results when training with dataset A are better than those obtained when training with dataset B (line 4 and 5 of Table 1). These findings match with the visual observation we made during our review of the datasets, that dataset A, compared with dataset B and C, appears to have more complex examples and is more representative of cervix variability and appearance of the os. We carried out a final experiment to investigate the effect that varying the number of strongly annotated images has on the Mask^X R-CNN classification. We ran three tests, where we reduced the number of $A_{\_mask}$ images used in the second stage to 70%, 60%, and 50%, respectively of the total $A_{_mask}$size and randomized this data. We tested against datasets A, B, and C, just as for Mask R-CNN. The segmentation results are visualized in Figure 3 with both ground truth and predicted contours marked, and the performance scores are presented in Table 3.

Table 3. Results for training Mask^X R-CNN with subset of $A_{\_mask}$.

Portion of $A_{\_mask}$ used	*Testing Dataset :A*			
	Avg. DICE/IoU	*DR*	*Avg. CD*	*Avg. MD*
70%	0.28/0.18	0.887 (400/451)	15.59	10.12
60%	0.25/0.16	0.873 (394/451)	18.37	12.69
50%	0.24/0.13	0.851(384/451)	20.17	15.88
	Testing Dataset :B			
70%	0.33/0.20	0.957 (220/230)	7.0	1.3
60%	0.28/0.16	0.913 (210/230)	8.8	2.8
50%	0.28/0.13	0.8870(204/230)	9.9	4.2
	Testing Dataset :C			
70%	0.35/0.23	0.975 (117/120)	10.9	2.8
60%	0.31/0.18	0.958 (115/120)	12.8	3.8

50%	0.28/0.16	0.942(113/120)	14.9	5.2

When we train the second stage of the Mask^X R-CNN with less strongly annotated images, we can observe that the overall performance of testing C dataset drops with the number of strongly annotated images going down from 70% of the original set to 50%. The degradations that we observe in testing datasets A, B, and C, as the amount of strongly annotated images decreases, may be tolerable for some applications, and indicates further study.

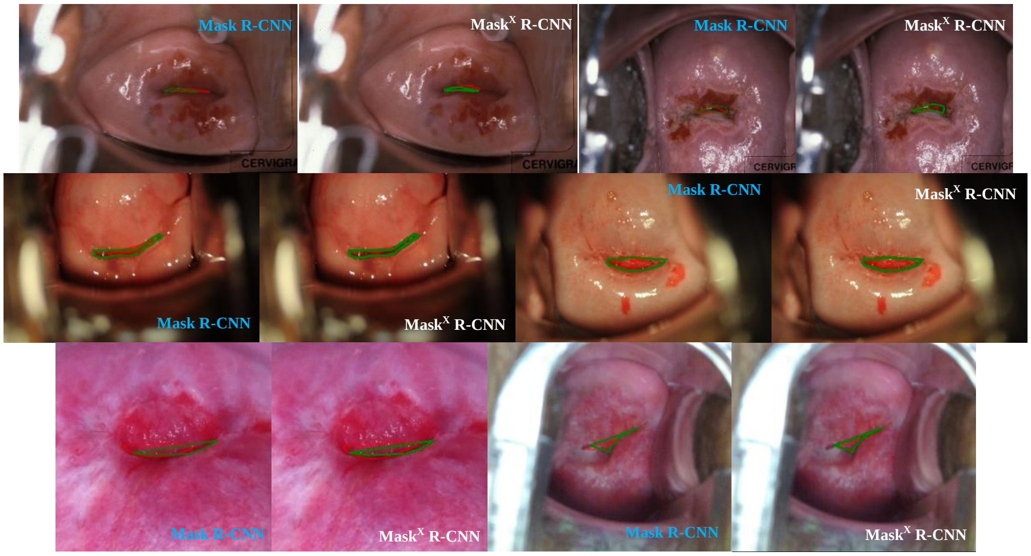


Figure 3: Segmentation examples from both Mask R-CNN and Mask^X R-CNN with full set of $A_{\_mask}$. The green lines are the ground truth marked by human experts, red lines are model predictions. Images in 1st row are results from ALTS dataset; images in 2nd row are results from CVT dataset; images in 3rd row are results from Kaggle dataset.

4.3 Discussion

We have investigated both Mask R-CNN and Mask^X R-CNN networks with respect to os segmentation performance. We tested both networks with strongly annotated data inputs (in the case of Mask^X R-CNN, bounding boxes were also input). We used three datasets, A, B, and C, each split into training and testing subsets. We achieved our best performance when training a Mask R-CNN with dataset A and testing on B, with an average of about 1 pixel of minimal distance from predicted contours to the ground truth contours among 99% of the test images in dataset B. Our results outperform the results reported in [7] with smaller minimal distance value (more accurate segmentation) on a larger number of images. We can also achieve suboptimal but comparable results with Mask^X R-CNN using the same datasets. We also observed better performance in testing dataset C, trained on A, as compared to testing dataset C, trained on B. This reinforced our visual observations that dataset A has more complex, varied examples, and may provide more robust training data.

We also investigated effect of using fewer numbers of strongly annotated images, as compared to weakly annotated images (in Mask^X R-CNN). We found that with Mask^X R-CNN, os contours can be predicted with some, but perhaps acceptable degradation, when we reduced our original strongly annotated training set by as much as 50%. This is a promising finding in 1) possibly using more weakly annotated images for applications in segmenting the os region where few input contours are available; 2) automatically creating more mask annotations with the trained model on datasets without contour annotations.

We also studied the images in which 1) the models make no prediction; 2) the predicted contour totally mis-matches with the ground truth. These images are denoted as "error cases as visualized in Figure 4. In the "error cases" where the models make predictions that have no overlap with defined ground truth (Figure 4g-m), over 95% of the predicted contours are located visually very close to the manually marked contour. We believe it is reasonable to attribute some of these errors to "label noise" or "label errors". These include: 1) manual annotations we can reasonably assume to be marked at the wrong location (Figure 4g-j); 2) manual annotations with the os region too severely constrained; in some cases predictions include the correct area, but there is still no overlap between the prediction and the ground truth. We also examine the "error cases" where the model make no prediction (Figure 4a-f), and hypothesize that they result from factors such as: 1) the presence of visual distractors such as pubic hair, intra-uterine devices (IUDs), the speculum, and parts of human hands; 2) a full view of the os region is not available: in some cases we found that the cervix region sometimes occupies less than 50% of the entire image and the os region is an even smaller part of the entire image; 3) the image is of degraded quality issues due to blur, bad focus, vaginal discharge, or low lighting conditions. In these cases, the images are "declined" by the models based on the features learned from the annotated os region, and are assigned low probabilities (confidence values) to be recognized as the region of interests. On the other hand, these error cases can be used to help in image acquisition and provide assistance in image quality analysis and control.

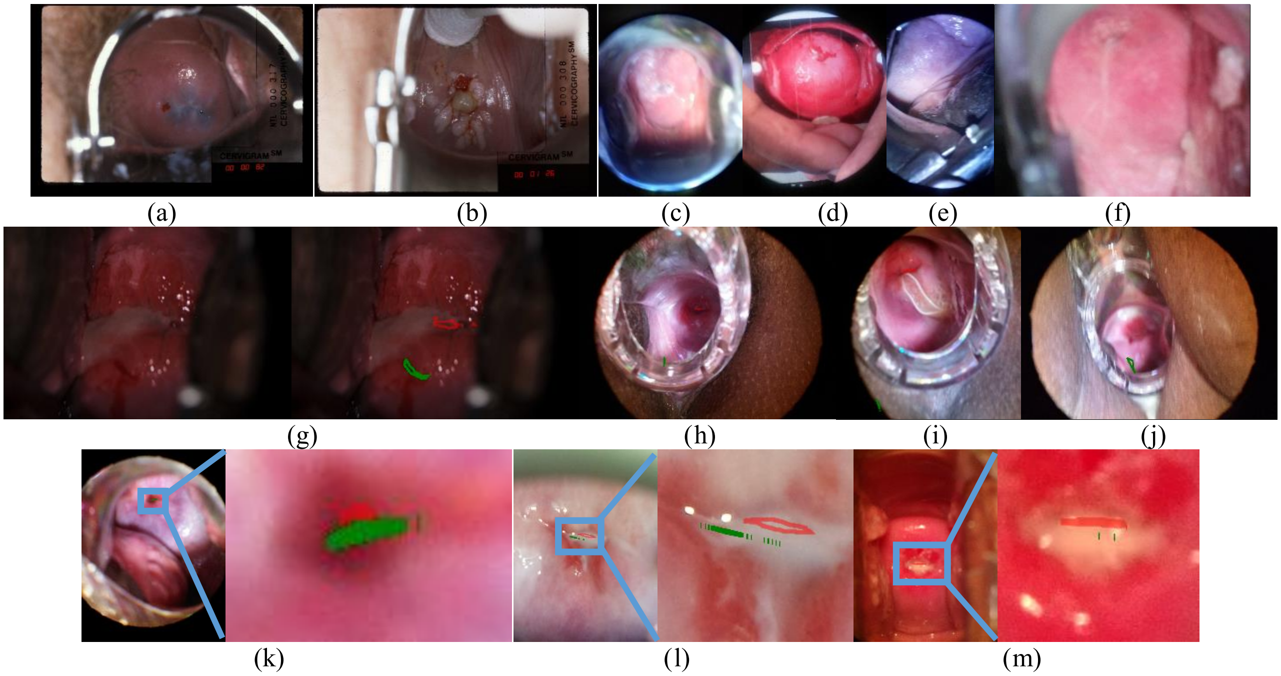


Figure 4: Image samples of "error cases". The green lines are the ground truth marked by human experts, red lines are model predictions. For the image examples in the 1st row, no prediction is made. These images have various quality issues. Images in the second row may be reasonably considered to be improperly annotated. The first two images in the 2nd row are the before-after comparisons of one image. In the 3rd row, prediction regions are enlarged for clearer illustration: though the predicted contours are close to the ground truth contour, they have no region of intersection.

5. CONCLUSION

Os region segmentation is important in improving automated cervical cancer screening performance as well as other cervical cancer related tasks. In this study, we evaluated the os segmentation performance of two state-of-the-art deep learning architectures: Mask R-CNN and Mask^{X} R-CNN. We tested on three available datasets: Kaggle MobileODT, CVT and ALTS. We conducted 1) homogeneous ("within dataset") training and testing, where we trained with one dataset and tested the model on the test split of the same dataset; and 2) heterogeneous ("across datasets") data training and testing, where we trained with one dataset and tested the model on the test split of a different dataset. In these tests we achieved a highest detection rate of 99.1% and an average minimal distance of 1.02 pixels (Table 1, Line 2). Moreover, we investigated the effect of training Mask^{X} R-CNN with: (1) the full set of strongly annotated data (mask annotations) and (2) a small fraction (as little as 50% of full set) of strongly annotated; we achieved a best detection rate of 0.942 and an average minimal distance of 5.2 pixels on testing dataset C using only 50% of $A_{\_mask}$. In analysis of the cases where the models "refuse" to make any prediction, we found that the model is capable of "declining" images with quality issues which leads to failure to detect the os ROI. This capability of automatically filtering out low-quality images can be used to contribute to higher quality in data acquisition. In our future study, we plan to continue optimizing the algorithm with respect to segmentation performance and test on larger scale and more varied datasets.

ACKNOWLEDGEMENT

This research is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications. Also, we thank Dr. Mark Schiffman of the National Cancer Institute for his kind review of the manuscript. We also are grateful to MobileODT for providing us an image dataset.

REFERENCES

[1] World Health Organization, “Human papillomavirus (HPV) and cervical cancer,” World Health Organization, Jan. 24, 2019. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer, accessed on: July 18th, 2019.

[2] Jeronimo, J., Massad, L. S., Castle, P. E., Wacholder, S. and Schiffman, S., National Institutes of Health (NIH)-American Society for Colposcopy and Cervical Pathology (ASCCP) Research Group. “Interobserver agreement in the evaluation of digitized cervical images,” Obstetrics and gynecology, (2007).

[3] Hu, L., et al., “An observational study of deep learning and automated evaluation of cervical images for cancer screening,” Journal of the National Cancer Institute, (2019).

[4] Guo, P., Xue, Z., Long, R. and Antani, S., “Deep learning for assessing image focus for automated cervical cancer screening,” presented at the IEEE International Conf on Biomedical and Health Informatics. Chicago, USA, (2019).

[5] Sellors, J.W. and Sankaranarayanan, R., [Colposcopy and Treatment of Cervical Intraepithelial Neoplasia: A Beginner’s Manual], International Agency for Research on Cancer, (2003).

[6] Jordan, J., Singer, a., Jones, H. and Shafi, M., [The Cervix], Wiley, 23-29, (2009).

[7] Zimmerman, G., Gordon, S. and Greenspan, H., “Automatic landmark detection in uterine cervix images for indexing in a content-retrieval system,” 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 1348-1351 (2006).

[8] Greenspan, H., Gordon, S., Zimmerman, G., Lotenberg, S., Jeromino, J., Antani, S. and Long, R., “Automatic detection of anatomical landmarks in uterine cervix images,” IEEE Trans Med Imaging 28(3), 454-468 (2006).

[9] García-Arteaga and Kybic, J., "Automatic landmark detection for cervical image registration validation," Proc. SPIE 65142S, (2007).

[10] Xue, Z., Long, R., Antani, S., Neve, L., Zhu, Y. and Thoma, G., “A unified set of analysis tools for uterine cervix image segmentation,” Comput Med Imaging Graph 34(8), 593–604 (2010).

[11] Gordon, S. and Greenspan, H., “An agglomerative segmentation framework for non-convex regions within uterine cervix images,” Image and Vision Computing 28(12), 1682-1701 (2010).

[12] Alush, A., Greespan, H. and Goldberger, J., “Automated and interactive lesion detection and segmentation in uterine cervix images,” IEEE Trans Med Imaging 29(2), 488-501 (2010).

[13] Cervix Segmentation paper – TBD

[14] He, K., Gkioxari, G., Dollár, P. and Girshick, R., “Mask R-CNN,” Proc. IEEE International Conference on Computer Vision, (2017). arXiv: 1703.06870.

[15] Hu, R., Dollár, P., He, K., Darrell, T. and Girshick, R., “Learning to segment every thing,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2018). arXiv: 1711.10370.

[16] He, K., Zhang, X., Ren, S. and Sun, J., “Deep residual learning for image recognition,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015). arXiv: 1512.03385

[17] MobilODT [Company], website: https://www.mobileodt.com/

[18] Fernandes. K. and Cardoso, J. S. “Ordinal Image Segmentation using Deep Neural Networks,” International Joint Conference on Neural Networks, 1-7, (2018).

[19] Herrero, R., et al., “Rationale and design of a community-based double-blind randomized clinical trial of an HPV 16 and 18 vaccine in Guanacaste, Costa Rica,” Vaccine, 26(37), 4795-4808, (2008).

[20] Herrero, R., et al., “Prevention of persistent Human Papillomavirus Infection by an HPV16/18 vaccine: a community-based randomized clinical trial in Guanacaste, Costa Rica,” Cancer Discov, 1(5), 408-419, (2011).

[21] Schiffman, M. and Adrianza, M. E., “ASCUS-LSIL Triage Study. Design, methods and characteristics of trial participants,” Acta Cytol, 44(5), 726-742, (2000).

[22] Lin, T., et al, “Microsoft COCO: Common Objects in Context,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015). arXiv:1405.0312v3

[23] Ren, S., He, K., Girshick, R. and Sun, J., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149, (2016).

[24] Lin, Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, a., “Feature Pyramid Networks for Object Detection,” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 936-944, (2014).