Deep Learning for Assessing Image Focus for Automated Cervical Cancer Screening

Peng Guo[#], Sanjana Singh[#], Zhiyun Xue, Rodney Long, Sameer Antani
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine, NIH
Bethesda, MD, USA

***Abstract*—Cervical cancer is one of the leading causes of women's mortality worldwide. Early diagnosis of precancer, the highest grade of cervical intraepithelial neoplasia (CIN) prior to being called cancer, is critical in improving the survival rate. Visual Inspection with Acetic Acid (VIA) is a visual examination technique that reveals lesions with HPV infection that whiten on exposure to 5% acetic acid. Combined with HPV testing, VIA can be an effective screening technique in low resource settings. Recently, a deep learning method for screening cervical images showed the ability to detect precancer at a rate superior to human experts, particularly in women 25-49 years of age, the critical age for high yield in screening the disease. However, the method's performance depends on having good quality images. Smartphone-based enhanced cervical image assessment is a strong candidate to replace the commonly used VIA technique in rural low-resource settings but is susceptible to poor image quality – particularly out-of-focus images. Thus, detecting sharp images is a critical first step toward accurate screening for cervical precancer.**

We present a deep learning architecture that detects in-focus smartphone cervical images. The method was evaluated on over 4500 images acquired by a commercial cervical image acquisition framework. We examined and compared three types of deep learning networks: an object detection model (RetinaNet); fine-tuned deep learning models (VGG, Inception); and transfer learning models (VGG, Inception Feature extractor + SVM) in evaluating the sharpness of the images. The highest technical image quality assessment accuracy we obtained was 94%.

Keywords— cervical cancer, convolutional neural network, RetinaNet, image focus, SVM, visual inspection with acetic acid, machine learning, transfer learning

I. INTRODUCTION

According to the World Health Organization (WHO), there were about 570,000 new cases of invasive cervical cancer diagnosed in 2018, of which 311,000 women died [1]. Regular screening is imperative in reducing mortality rates, as symptoms are not felt until the late stages of the cancer. Possible screening modalities to address cervical cancer diagnosis include Pap test, HPV test and visual inspection with acetic acid (VIA). Visual inspection with acetic acid (VIA) is an inexpensive and simple alternative to screening modality, but with inadequate performance; VIA is conducted by applying dilute (3%-5%) acetic acid to the cervix in a vaginal speculum exam and then visually inspecting the cervix. By identifying the color changes on the cervix, the health care provider may be able to determine the presence of possible precancerous lesions or cancer [2].

Digital imaging of the cervix has been used for maintaining patient records of screening visits. Recently, a deep learning method was shown to potentially improve or replace VIA with a more accurate automatic screening application [3], particularly in women aged 25-49 years, which is the most high-yield age group for cervical pre-cancer. In this age group the algorithm, called *Automatic Visual Evaluation* (AVE), has been shown to be better than humans in predicting pre-cancer, which is defined as cervical intra-epithelial neoplasia (CIN) above grade 2.

The challenges of making accurate diagnosis with digital imaging technologies applied to cervix images result from several factors that affect image quality, including low lighting conditions, suboptimal usage of the imaging device by the medical technician, and human body movement. One digital cervical imaging device for *Enhanced Visual Assessment* (EVA) of the cervix has been developed by MobileODT (https://www.mobileodt.com). The EVA device uses an embedded smartphone as the data source; EVA is a highly portable platform for cervical image data acquisition and, potentially, early pre-cancer and cancer diagnosis. The images in the EVA dataset that we used for this paper were acquired only to document the patient's visit, not to diagnose. The quality of these images varies widely, since the images did not undergo the quality control that we might expect in a future screening system. Some images were degraded by motion blur or other out-of-focus characteristics, poor lighting, poor camera positioning relative to the target anatomy (i.e., the cervix), or other flaws which might result in inaccurate screening by the AVE system. It should also be noted that these images were not acquired for automated image evaluation using machine learning. Our results in applying deep learning to image focus assessment may provide guidance for improved image capture to support AVE, or similar techniques. For example, a method that is able to automatically classify cervix images for sharpness and filter out "poor" images may address a significant part of the problems in capturing high quality images.

Our goal, then, is to investigate state-of-the-art deep learning technologies for automatic assessment of cervical image focus. If our methods are successful, we anticipate that field workers could be effectively advised to recapture an image when the previously taken images are detected as "out of focus" or of low quality. The technique could be also used to develop an automatic image capture utility that assists in taking "good

# These two authors contributed equally to the work

978-1-7281-0848-3/19/$31.00 ©2019 IEEE

quality" photos. Further, these "in focus/good quality" pictures can be input to the AVE algorithm for precancer classification.

A major consideration we have to deal with is that, in realistic circumstances, it is difficult to obtain a reference cervical image for focus evaluation. Thus, we are carrying out what is commonly called a *No Reference Image Quality Assessment* (NR-IQA) task. However, there is little literature regarding the characterization of cervical image sharpness to guide us. In [4], cervix images collected by MobileODT are segmented with a fixed circular frame in the center, and processed to extract features according to focus measure algorithms, then classified with a random forest classifier after feature selection. The algorithm performs well on distinguishing extreme cases of image sharpness, such as "very poor" and "excellent" quality. For more general (non-cervix) images, techniques evaluating image sharpness have been widely investigated [5]. Studies using NR-IQA to classify image quality on more general image datasets report the use of several metrics, such as focus measuring operators, natural scene statistics, adversarial learning, and deep learning methods, to assess the level of image sharpness. Among the deep learning methods, the highest accuracy for NR-IQA tasks has been overwhelmingly attributed to deep convolutional neural networks (CNNs) [5, 6].

In this study, we first performed automatic cervix detection using our pre-trained deep learning model to provide bounding box annotation of the detected cervix regions. The image can be cropped to this region, as necessary, for further processing. Then, we applied a leading object detection model (RetinaNet [7]) and VGG16 as base networks for image sharpness classification. We compared the results from RetinaNet with those from a fine-tuned VGG/Inception as well as with those from a transfer learning model of VGG/Inception as feature extractor followed by L1 feature selection and SVM classifier. Based on our experiments, RetinaNet with ResNet50 "backbone" was the top performing model.

To summarize, this paper makes the following contribution: a) we implemented an object detection model (RetinaNet) for cervical image sharpness assessment, which achieves the top performance; b) we carried out model comparison between RetinaNet and alternative models such as fine-tuned VGG and VGG with feature selection; c) we applied cervix detection techniques which may reduce the time and effort for cervical region annotation; these techniques have potential to pre-process large datasets of cervix images.

II. Methods

A. Dataset and Annotation

We obtained 4525 deidentified images (Fig. 1) collected from 1399 women using MobileODT's EVA system. Each woman could have multiple images taken during a session. As described in [5], these images were annotated according to the focus score output from a trained classifier and the annotations were quantized into two-class labels with manual refinement by MobileODT. Using this method, of the 4525 images, 2170 were labeled as "Not Sharp" and 2355 as "Sharp". For purposes of this paper, the label "Sharp" is semantically associated with being in focus. We propose several approaches that utilize convolutional neural network (CNN) as well as transfer learning methods to evaluate cervical image quality with state-of-the-art results.

(a) (b) (c) (d)

Fig. 1: Examples of MobileODT (EVA) images. (a) and (b) are labeled as "Not Sharp", (c) and (d) are labeled as "Sharp".

B. Cervix Detection

EVA images may contain not only the cervix, but also the vaginal wall and, if not sufficiently zoomed, include the speculum in the field of view. The anatomy of interest is the cervix, thus the bounding box information of the cervix region is crucial. With the bounding box localization of the cervix region, we can restrict the quality assessment to our area of interest (cervix) instead of including irrelevant regions. To accomplish this task, EVA images (not included in the above dataset) were manually marked with cervix bounding boxes, and used to train the object detection models based on Faster R-CNN [8]. There are two main goals in generating training data with the cervix detection model: the first goal is to generate localization annotation (cervix bounding box) since the cervix coordinates are required for training RetinaNet; the second goal is to crop images to cervix regions only for use as input to other models in this study (for example, the fine-tuned VGG model). We also manually inspected and refined the automatic detection results to ensure high quality of training data in this study. Fig. 2 shows the samples of the automated detection of the cervix region.

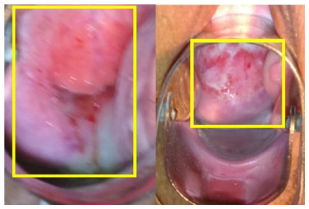

Fig. 2: Examples of cervix detection images from pre-trained Faster R-CNN model. Bounding boxes in yellow are selected and the four vertices are saved as required by RetinaNet as well as for cervix region cropping.

C. RetinaNet

As a single stage object detection structure in deep learning, RetinaNet [7] shares the similar concept of "anchor proposal" with [8]. It uses a feature pyramid network [9] where features on each of the image scales are computed separately (lateral connections) and then summed up through convolutional operations (top-down pathways). Furthermore, with the proposed "focal loss" function, RetinaNet is reported to have better performance handling class imbalance and focusing on hard, misclassified examples. We make the final decision by using the largest predicted score and the associated bounding box.

D. VGG

We utilized the ImageNet pre-trained weights to initialize the fine-tuned VGG network. The layers before the last convolution layer, the third convolutional layer in the fifth block, were frozen, after which weights were set to be trainable. Besides fine-tuning the VGG model, our pre-trained VGG network was also used as a feature extractor. At the point of extraction from the first fully connected layer, each image sample had been transformed into a 4096-digit feature vector. As shown in Fig. 3, we then applied L1-based feature selection to reduce the dimensionality of the extracted features to train an SVM classifier. The results are listed as "Feature Extraction + L1 + SVM" in Table 2. Only the non-zero coefficients were then selected, after which an SVM classifier was trained with radial basis kernel on the selected features to classify images as "Sharp" or "Not Sharp".

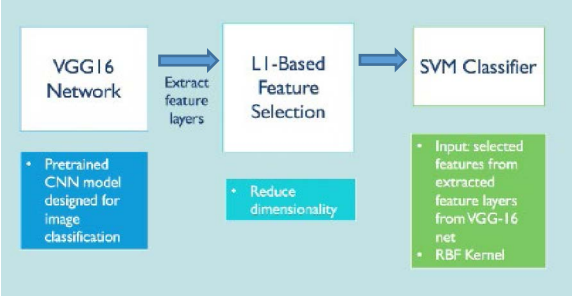

Fig. 3: Method flowchart for the VGG+L1+SVM method

E. Inception v3 convolutional Neural Network

The Inception v3 model was another deep convolutional neural network that we used. Similar to our work with the VGG model, we utilized the Inception v3 framework to perform fine-tuning as well as feature extraction.

III. Experiments

A. Data Preparation

The entire data set was split into a training set and a testing set, in an 80/20 ratio. Instead of randomly splitting all the data, we firstly sorted the dataset according to the 1399 unique subject IDs, then split the images according to unique subject IDs only, so that images of the same woman were all in either training or testing, regardless of what class labels they have.

The reason to prepare the data in this way is to avoid data leakage between training and testing. We prepared a balanced training set where "Sharp" images and "Not Sharp" images occur in an approximate 1:1 ratio. Among these 3679 training images, 1923 were labeled as "Not Sharp" and 1756 as "Sharp". For RetinaNet, original images are directly used as input images; for fine tuning VGG/Inception and the VGG/Inception extractor + L1 feature selection + SVM, all the images are cropped into cervical regions, as described in the previous section.

B. Results

For each training and testing scenario, we trained the VGG based models, Inception v3 based models using a batch size of 32 and the RetinaNet model using a batch size of 1, over 300 epochs. Evaluation matrices were calculated to compare the results between different models, including accuracy, precision and recall.

In regard to training and testing time, testing time was negligible for all models, as expected. For training, RetinaNet is structurally more complicated compared with the VGG-based and Inception models so loading more graph and weights information should be expected to consume more training time. Running times for the models are given in Table 1. However, based on our classification results (Table 2), higher overall performance is obtained from RetinaNet compared with other tested models, which indicates that we trade off training time to classification performance.

Table 1: Running Time

Model	Training Time/ Epoch (s)	Testing Time/ Image (s)
RetinaNet (ResNet50)	1838.7	0.31
RetinaNet (VGG19)	1726.4	0.24
Fine-tuned VGG	849.2	0.20
Fine-tuned Inception	704.2	0.17

RetinaNet with ResNet50 as the backbone was 98% sensitive and 85% specific to the "Not Sharp" class at a probability threshold of 0.5, and achieved 94% for accuracy. With the same probability threshold, RetinaNet with VGG19 as the backbone achieved 90% for accuracy. The ROC curve for the test set in RetinaNet is shown in Fig. 4 (Red dots denote the points of different probability thresholds).

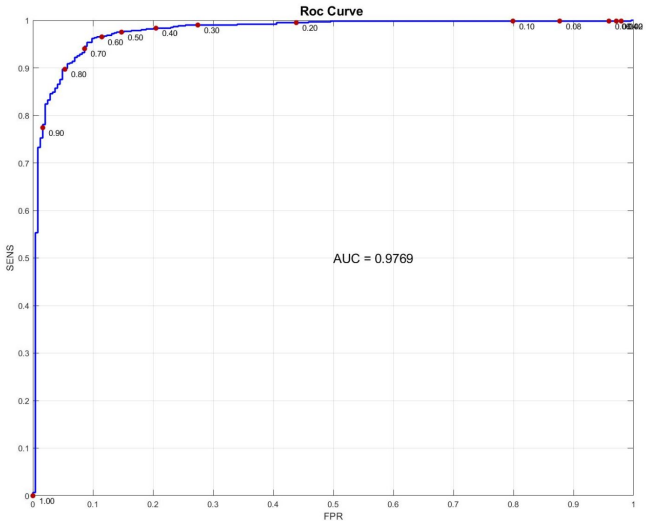

Fig. 4: ROC of RetinaNet with Resnet50 as the backbone. Red dots denote the points of different probability thresholds.

For the VGG architecture, we utilized transfer learning. After significant network fine-tuning over several epochs, the highest test set accuracy achieved was 92.43% with F1 score of 0.87. Several other models were also tested with comparable results. In Table 2, we list the results of other methods we implemented. For the method utilizing the pre-trained VGG model with L1-based feature selection, for each test set image, the features were first extracted using the VGG architecture and

then reduced to 1381 features by selecting the most predictive features, which were used as input for an SVM classifier. L1 feature selection slightly helps in the customized VGG architecture by achieving accuracy of 0.90 compared with 0.88 without L1 feature selection.

The Inception v3 model was also utilized as a feature extractor for the task of cervix image quality assessment. The feature selection reduced the dimensions of the input of the SVM from 2048 to 1088. Similar to the VGG model, the L1-based feature selection brought improvements to most performance metrics, boosting the test set accuracy to 91.49% with F1 score of 0.85. We also fine-tuned the Inception network by freezing the first 249 layers and training the rest of the layers, which ultimately achieved 86.17% accuracy and F1 score of 0.75.

Table 2: Testing result statistics ("Not Sharp as Positive")

		Test Set Accuracy	F1 Score	Precision	Recall
RetinaNet	**ResNet50**	**0.94**	**0.96**	**0.94**	**0.98**
	VGG19	0.90	0.90	0.93	0.88
VGG	Fine tuning	0.92	0.87	0.88	0.86
	Feature extraction + L1 + SVM	0.90	0.84	0.81	0.87
Inception	Fine tuning	0.86	0.75	0.80	0.70
	Feature extraction + L1 + SVM	0.91	0.85	0.84	0.86

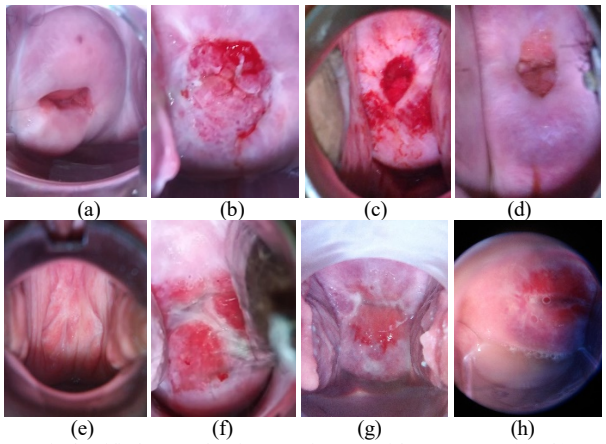

Fig. 5: Misclassified examples from RetinaNet with ResNet50. (a), (b), (c), (d) are the "Not Sharp" images predicted as "Sharp"; (e), (f), (g), (h) are the "Sharp" images predicted as "Not Sharp".

Misclassified images were also manually examined (samples are shown in Fig. 5), among which false positive images and some false negative images have the presence of patients' hair, intra-uterine devices (IUDs), or partial presence of the speculum. These factors sometimes confound the classification when they overlap the cervical region. We conjecture that incorrect predictions could be reduced by collecting more correct examples for the training set or refining the current dataset with uniform and clear criteria for selecting training images. For the long-term purposes of collecting high quality data, training for the field worker could also be considered. We should note that there remain a residual number of problematic images that are mislabeled for unknown reasons.

IV. Conclusion

Low-quality images captured by practitioners are a limiting factor for the performance of cervix screening with digital imaging technology. In this study, we implemented and fine-tuned multiple deep learning architectures, including RetinaNet(a one-stage detection model), and fined-tuned VGG and Inception-based models. We prepared the training and testing datasets to exclude data leakage between the two and used automatic cervix detection results as the basis of our ground truth annotations. We compared these deep learning models for classification of "Sharp" and "Not Sharp" images. RetinaNet outperformed the other models with the tradeoff of additional training time for higher classification performance, as compared with fine-tuned VGG which is the second best model. From analyzing the misclassified images, we obtained clues suggesting the reason for some misclassifications. To capture better quality images, a clear and complete standard must be followed in data acquisition. As next steps, we plan to continue optimizing the algorithm with respect to classification performance and training time.

Acknowledgment

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. In addition, we thank MobileODT for providing us the image data and Dr. Mark Schiffman of the National Cancer Institute for his diligent review of the manuscript.

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: 30-Jan-19]

[2] L. Tao et al., "Cervical Screening by Pap Test and Visual Inspection Enabling Same-Day Biopsy in Low-Resource, High-Risk Communities," *Obstetrics & Gynecology*, vol. 132, no. 6, pp. 1421–1429, Dec. 2018.

[3] L. Hu et al., "An observational study of deep learning and automated evaluation of cervical images for cancer screening," *Journal of the National Cancer Institute*, Jan. 2019. doi: 10.1093/jnci/djy225.

[4] M. Jaiswal et al., "Characterization of cervigram image sharpness using multiple self-referenced measurements and random forest classifiers," in *Proc. 10485 Optics and Biophotonics in Low-Resource Settings IV*, 2018.

[5] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *ICIP*, 2016.

[6] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. Bovik, "Deep Convolutional Neural Models for Picture-Quality Prediction: Challenges and Solutions to Data-Driven Image Quality Assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[7] TY Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in *ICCV*, 2017.

[8] S. Ren, K. He, R. Girshick, and J. Sun, " Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in NIPS, 2015.

[9] TY Lin, "Feature pyramid networks for object detection," in *CVPR*, 2017