PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Characterization of cervigram image sharpness using multiple self-referenced measurements and random forest classifiers

Mayoore Jaiswal, Matt Horning, Liming Hu, Yau Ben-Or, Cary Champlin, et al.

Mayoore Jaiswal, Matt Horning, Liming Hu, Yau Ben-Or, Cary Champlin, Benjamin Wilson, David Levitz, "Characterization of cervigram image sharpness using multiple self-referenced measurements and random forest classifiers," Proc. SPIE 10485, Optics and Biophotonics in Low-Resource Settings IV, 1048507 (13 February 2018); doi: 10.1117/12.2292179

SPIE.

Event: SPIE BiOS, 2018, San Francisco, California, United States

Characterization of cervigram image sharpness using multiple self-referenced measurements and random forest classifiers

Mayoore S. Jaiswal[1], Matt Horning[1], Liming Hu[1], Yau Ben-Or[2], Cary Champlin[1], Ben Wilson[1], David Levitz[2]

[1]Intellectual Ventures Laboratory, 14360 SE Eastgate Way, Bellevue, WA 98007
[2]MobileODT Ltd., Ben Avigdor 8, Tel Aviv 67017, Israel

ABSTRACT

Cervical cancer is the fourth most common cancer among women worldwide and is especially prevalent in low resource settings due to lack of screening and treatment options. Visual inspection with acetic acid (VIA) is a widespread and cost-effective screening method for cervical pre-cancer lesions, but accuracy depends on the experience level of the health worker. Digital cervicography, capturing images of the cervix, enables review by an off-site expert or potentially a machine learning algorithm. These reviews require images of sufficient quality. However, image quality varies greatly across users.

A novel algorithm was developed to evaluate the sharpness of images captured with the MobileODT's digital cervicography device (EVA System), in order to, eventually provide feedback to the health worker. The key challenges are that the algorithm evaluates only a single image of each cervix, it needs to be robust to the variability in cervix images and fast enough to run in real time on a mobile device, and the machine learning model needs to be small enough to fit on a mobile device's memory, train on a small imbalanced dataset and run in real-time.

In this paper, the focus scores of a preprocessed image and a Gaussian-blurred version of the image are calculated using established methods and used as features. A feature selection metric is proposed to select the top features which were then used in a random forest classifier to produce the final focus score. The resulting model, based on nine calculated focus scores, achieved significantly better accuracy than any single focus measure when tested on a holdout set of images. The area under the receiver operating characteristics curve was 0.9459.

Keywords: Cervical cancer, low resource setting, VIA, digital cervicography, image focus, random forest classifier, feature selection, machine learning

1. INTRODUCTION

Cervical cancer is the fourth most common cancer among females worldwide, according to 2012 estimates from the World Health Organization (WHO) International Agency for Research on Cancer (IARC)[1]. There are more than half a million new cases and approximately quarter million deaths annually reported worldwide related to cervical cancer[2]. Approximately 85% of the cases and 87% of the deaths occur in less developed regions of the world like South America, sub-Saharan Africa, and South-East Asia[1].

The main target of cervical cancer screening is to identify, cervical intraepithelial neoplasia (CIN), a premalignant lesion that can progress to cervical cancer if left untreated. In the developed countries, women are screened for cervical cancer by performing routine Papanicolaou (Pap) smear test and HPV testing[3,4]. A positive Pap test is followed up by colposcopy with the option for biopsy, as well as treatment if necessary[5]. In low resource settings (LRS) where there is a higher prevalence of cervical cancer, there is lack of awareness, limited access to healthcare, poorly-trained medical staff, inadequate cancer screening infrastructure, and highly-variable detection practice. In these countries, visual inspection with acetic acid (VIA) is one of the few available screening tests[6]. A clinician performs VIA by using a speculum and a long swab to apply a thin layer of dilute acetic acid (vinegar) to the cervix and then uses a headlamp or

Optics and Biophotonics in Low-Resource Settings IV, edited by David Levitz, Aydogan Ozcan, David Erickson
Proc. of SPIE Vol. 10485, 1048507 · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2292179

flashlight to observe the cervix through the vaginal canal to detect if areas of the cervix have turned white. A positive VIA test often indicates an immediate cryotherapy treatment[6]. VIA has very limited diagnostic accuracy which can lead to challenges in monitoring and evaluating the screening procedure.

Digital imaging of the cervix (digital cervicography) can improve the efficacy of cervical cancer screening by enabling automatic diagnosis applications, remote consultations, and effective user training. In this study, we used the enhanced visual assessment (EVA) system developed by MobileODT. The EVA system consists of a smartphone based mobile colposcope (Figure 1), an application to control the smartphone, a secure cloud-based image portal to store and review images, and additional decision support features to help augment VIA procedures in LRS[7].

Figure 1. (a) The EVA system and its components: a smartphone, a light source, a lens and a neck strap. (b) EVA system being used in a field clinic. (c) The framing ring on the CervDx application.

The quality of images collected by the EVA in LRS often varies significantly due to most practitioners' limited proficiency in using smartphones, challenging screening environment, and nervous patients for whom it may be their first cervix exam. Additionally, cervical imaging is inherently difficult due to the low light conditions, glare, and the desire for images with a higher zoom than typically handled by smartphones. Collecting at least one good quality image per patient is sufficient. An algorithm that could automatically evaluate the quality of the image could help maintain the usefulness of cervical images. Such an algorithm could be used to instruct the user to take another image or assist them to improve the image quality. The proposed algorithm would need to be simple to use and to run on a smartphone. In this paper, we propose such an algorithm for integrating into the EVA software.

Evaluating the image quality is a challenging task for an algorithm. Any computed focus score depends on the image content. Independent of true focus, a cervix with many features or high contrast will typically score higher than one with fewer features or low contrast. Image noise and saturation levels also impact image quality scores. Another problem is that this application requires a "no reference" focus score. Human eyes and digital cameras can quickly defocus and refocus to evaluate sharpness relative to the same field at different focal distances before settling on the best focus. The proposed algorithm would need to classify absolute image quality, rather than the best image quality from a stack of similar images. In this work, we propose a solution that automatically chooses the best focus features and uses a random forest classifier to deliver a score that is able to categorize images by sharpness.

2. METHODS

A random subset of images from the MobileODT image portal collected in LRS was analyzed for focus by manually annotating them into "very poor", "poor", "good" and "excellent" categories. Figure 2 shows examples of digital cervigrams in the dataset, and each sub-figure represents a category of image sharpness ordered from low to high, left to right. It was found that 22% of the images were "very poor", 51% were "poor", 21% were "good" and only 6% were excellent. Around 73% of the images are unacceptable for further use due to poor quality.

Figure 2. Examples of images captured using the CerxDx app by practitioners. (a) – (d) illustrate the various levels of image sharpness from low to high.

There are many proposed image quality assessment algorithms in the literature. Some algorithms require a reference image[8,9] and others do not[10,11,12,13]. Since obtaining a high-quality reference image in the proposed setting is impossible, only methods without external reference images are considered in this work. We calculated focus scores of 28 widely used algorithms[13] and compared these to manual annotations of sharpness. Although many scores correlate with manual interpretation, none of the methods separate the focus categories well enough. Figure 3 shows the plots of calculated focus measure vs manual focus score for an example of six focus methods that were evaluated. Since quality assessment methods rely on image content to deliver a score, the score may depend on image contrast and not on actual image focus. For example, Brenner score[13] for a high contrast but the blurry image in Figure 2(a) is similar to a low contrast but sharp image in Figure 2(d).

Recently deep learning methods such as convolutional neural networks have attained state-of-the-art results in many computer vision problems, including image quality assessment[14]. Due to the combination of high computational power required to run standard deep learning models and the need for very large datasets to train such models, deep learning based methods were not explored for the proposed algorithm.

2.1 Cervix detection

Cervigram images contain the cervix, speculum, vaginal wall, and occasionally other body parts. Since our task is to evaluate whether or not the cervix in the image is in focus, we first segment the cervix in the image. This could be done by annotating the os or using a cervix detection algorithm to automatically segment it. We found cervix detection methods to be computationally expensive. With the introduction of the framing ring feature to the EVA system, the cervix fills a significant portion of most images captured with EVA[15]. A framing ring is a static ring image overlay displayed on the EVA screen, which guides the user to the proper framing of the cervix. This makes images zoomed in and the cervix occupies most of the image. As a result, we assume that the cervix is located at the center of the image.

Once the cervix is found, the image is rescaled using the zoom factor of the image. Then an *MxM* thumbnail centered on the cervix is extracted. This produces scale independent and centered thumbnails for further evaluation.

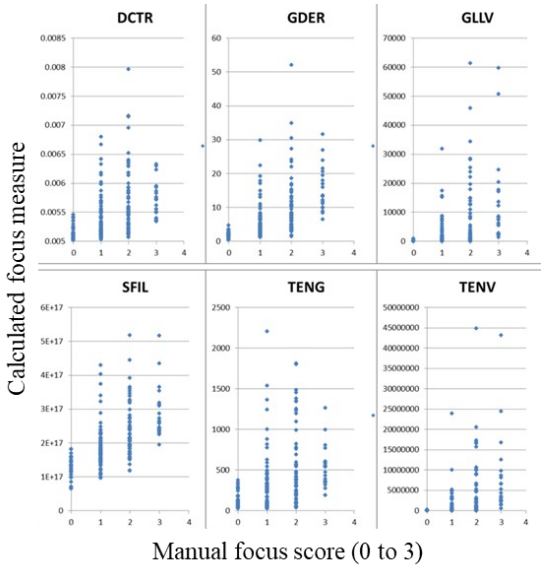


Figure 3. Various focus feature measures[13] vs manual focus score. The features compared in the figure (clockwise) are DCT reduced energy ratio (DCTR), Gaussian derivative (GDER), Gray-level local variance (GLLV), Steerable filters (SFIL), Tenengrad (TENG) and Tenengrad variance (TENV). Note that none of the feature measures separate well by manual focus values.

2.2 Feature extraction

Images were preprocessed with a Gaussian low pass filter to reduce noise. Features were then computed on the cervix thumbnails using focus algorithms[13]. High contrast images score higher on focus measure algorithms than low contrast images irrespective of focus level. To mitigate this problem, each image was compared to a blurred version of itself. When a sharp image is blurred the difference is significant, but when an already blurry image is again blurred the difference is small[12]. To achieve this, the preprocessed images were blurred using a Gaussian low pass filter with kernel size σ=8. Focus measures were computed on this blurred image. The ratio of the focus measures between the preprocessed and blurred image were used as features for the random forest classifier.

2.3 Feature selection

We calculated focus scores of the 28 focus algorithms[13]. Calculating all 28 focus features would increase the complexity of the algorithm and likely not run in real-time on a smartphone. Also, some features are correlated and redundant. So, the number of features must be reduced.

We introduce a weighted feature selection method in Equation (1) to quantitate the predictive value of the focus measures,

$$feature\ weight = \frac{1}{5}\sum\left(\frac{1}{t}, |w_{lasso}|, |w_{lr}|, N - r_{mRMR}, w_{MI}\right), \quad (1)$$

where $1/t$ is inverse time taken to compute the focus measure in 2000 thumbnails. $|w_{lasso}|$ is the absolute LASSO[16] feature weights. $|w_{lr}|$ is the absolute logistic regression[17] feature weights. r_{mRMR} is a ranking of features based on maximum dependency, maximum relevance, and minimum redundancy[18]. N is the total number of features to be ranked. $|w_{MI}|$ is weights based on mutual information between feature and label. All the components were normalized by the total number of features being considered, so that each component gets equal weight in the metric.

Each focus algorithm is ranked in descending order of feature weights and then the top *n* focus measures are selected. In our experiments, we use the top nine feature measures. The chosen focus measures were computed on all images in the dataset and used to train and test the random forest classifier.

2.4 Random forest classifier

We use feature selection method in section 2.3 to find the most predictive focus measures. In the training phase, focus features of the selected algorithms are computed as described in section 2.2. Then a weighted random forest classifier[19] is trained using the focus features. The "weighted" random forest weighs the minority class (in-focus images) to match the majority class (out-of-focus images) in the dataset. The hyper-parameters, number of ensemble learning cycles, the maximal number of decision splits, and minimum number of leaf node observations, are chosen using Bayesian optimization[20] on the validation set.

In the testing or deployment phase, selected focus features for each image are computed. Subsequently, the trained random forest classifier is used to predict the probability of focus using the selected focus features.

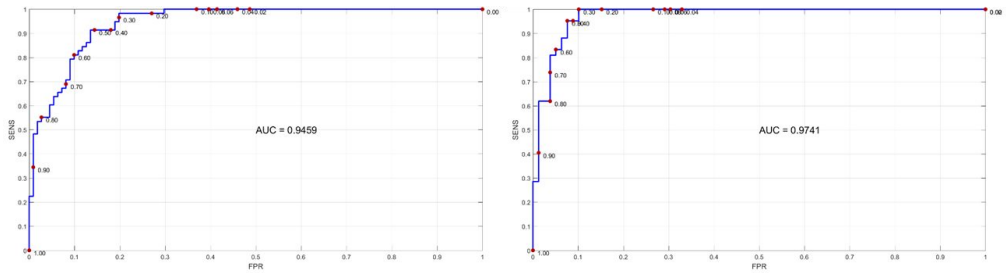
3. RESULTS

3.1 Dataset

The algorithm to evaluate image sharpness was developed using a dataset of 2083 training images, 169 validation images and tested on a holdout set of 4019 images. An annotation tool was developed to manually score the sharpness of images on a scale of 0 ("very poor") to 3 ("excellent"). The annotators were shown the images in random order and were instructed to choose a focus score from 0-3 for each image depending upon their evaluation. Example for each image quality category is shown in Figure 2. Example images were displayed on the annotation tool so that the annotator can conveniently compare the image to be annotated with samples in each image category. Since images had different zoom levels, the cervices were rescaled to fit the same size in the annotation tool. Images with scores 2 and 3 were considered to be eligible for digital evaluation. Each image was annotated by at least three expert image analysts. The average of the image analysts' score was used as ground-truth focus labels.

3.2 Results

For each test image in the dataset, the focus features were extracted using the chosen focus methods as outlined in section 2. Then the sharpness probability was predicted by passing the focus features through the trained random forest classifier. We plot the ROC curve (plot of sensitivity vs false positive rate at various probability thresholds) for the holdout images. Figure 4 (a) shows the plot for all the holdout images, and (b) shows the plot for all holdout images manually annotated as "very poor" and "excellent". The area under the curve AUC) for all test images is 0.9459 and for extreme cases is 0.9741. This shows that our algorithm is capable of separating sharp images from blurred images with high certainty.



(a) (b)

Figure 4. The ROC curve of the focus score algorithm plotted on the test data. The area under the curve (AUC) is shown in each plot. (a) Entire test dataset. (b) Only test images manually labeled as "very poor" and "excellent".

The algorithm is 91.38% sensitive and 86.49% specific for a probability threshold of 0.5. The threshold was selected to balance between sensitivity and specificity. It is necessary to reduce the number of blurred images passing the algorithm because that would affect later use of the image. Though it is possible to instruct the practitioner to take another image, it adds time and effort in an already resource-constrained environment. So the number of false positives and false negatives should be reduced.

Incorrectly predicted images were manually examined. Most false positive images and some false negative images had pubic hair, intra-uterine devices (IUDs), or parts of a swab present at a different focal distance than the cervix. Motion blur, glare and the os being too close to the edge of the image also caused the algorithm to flag the image as positive when these generated sharp, high-contrast features in the image. These incorrect predictions could be reduced by either rejecting images with non-cervix objects in the center of the image or including more examples of such images into the train set. Upon close examination, it was found that some images in the data set were mislabeled, leading to a small fraction of the incorrect predictions.

We performed another experiment where the random forest classifier was trained with all the focus features instead of the selected subset. The AUC of this classifier on the holdout set was 0.9556. Though this classifier used 3 times more focus features as the proposed algorithm, the performance improved by only 0.0097. Having fewer features to compute enables the proposed algorithm to run in real-time on the limited computational power of a smartphone.

4. CONCLUSION

Low-quality images captured by practitioners are a limiting factor for the performance of digital cervicography devices including the EVA. In this paper, we propose an algorithm to evaluate the sharpness of images. We compared various focus scores of the preprocessed image and the corresponding blurred image and used a trained random forest classifier to assign the probability of an image being out of focus. The algorithm has high sensitivity and specificity on a holdout set of images. It can test images in real-time on a smartphone. In future work, we will optimize the algorithm to run on smartphones in real time, integrate the algorithm into the EVA software, and assess its performance on images captured in the field.

ACKNOWLEDGEMENTS

Funding provided by The Global Good Fund I, LLC (www.globalgood.com). The authors acknowledge the support provided by Noni Gachuhi, David Bell, and Amir Bernat.

REFERENCES

[1] Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., Bray, F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on 12/15/2017.

[2] Siegel, R. L., Miller, K.D., and Jemal, A., "Cancer statistics, 2016." CA: a cancer journal for clinicians 66.1 (2016): 7-30.

[3] Massad, L.S., Einstein, M.H., Huh, W.K., Katki, H.A., Kinney, W.K., Schiffman, M., Solomon, N., Wentzensen, N., Lawson, H.W., "2012 Updated Consensus Guidelines, for the Management of Abnormal Cervical Cancer Screening Tests and Cancer Precursors", J Low Gen Tr Dis 17, S1-S27 (2013)

[4] Singer, A., Monaghan, J.M., and Quek, S.C., Lower Genital Tract Precancer: Colposcopy, Pathology, and Treatment, 2nd Ed. Blackwell, Oxford (2000).

[5] WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. Geneva: World Health Organization, 2013.

[6] Campos, N.G., Castle, P.E., Wright, T.C., Kim, J.J., "Cervical cancer screening in low resource settings: A cost-effectiveness framework for valuing tradeoffs between test performance and program coverage", Int. J Cancer 2015;137: 2208-19.

[7] Peterson, C., Rose, D., Mink, J., and Levitz, D., “Real-Time Monitoring and Evaluation of a Visual-Based Cervical Cancer Screening Program Using a Clinical Decision Tree Algorithm”, Diagnostics (2016).
[8] Larson, E.C., Chandler, D.M., "Most apparent distortion: full-reference image quality assessment and the role of strategy," Journal of Electronic Imaging 19(1), 011006 (1 January 2010). http://dx.doi.org/10.1117/1.3267105
[9] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., “Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612. (2004).
[10] Mittal, A., Moorthy, A. K., and Bovik, A. C., “No-reference image quality assessment in the spatial domain,” IEEE Transactions on Image Processing, (2012), 21(12), 4695-4708.
[11] Marziliano, P., Dufaux, F., Winkler, S., and Ebrahimi, T, “A no-reference perceptual blur metric”, In Image Processing. 2002. Proceedings. 2002 International Conference on (Vol. 3, pp. III-III). IEEE.
[12] Crete, F., Dolmiere, T., Ladret, P., and Nicolas, M., “The blur effect: perception and estimation with a new no-reference perceptual blur metric”. In Human vision and electronic imaging (2007, January) (Vol. 12, No. 6492, p. 64920).
[13] Pertuz, S., Puig, D., & Garcia, M. A., “Analysis of focus measure operators for shape-from-focus,” Pattern Recognition, (2013) 46(5), 1415-1432.
[14] Bianco, S., Celona, L., Napoletano, P., and Schettini, R., “On the use of deep learning for blind image quality assessment”, Journal of Signal, Image and Video Processing, Springer (2017). DOI: 10.1007/s11760-017-1166-8
[15] Horning, M., Ben-Or, Y., Jaiswal, M., Champlin, C., Gachuhi, N., Liming, H., Rosenberg, Y., Levitz, D., “A digital framing ring for stabilizing cervix location in digital cervicography images”, 2017 ASCCP Poster Presentation, Journal of Lower Genital Tract Disease, (2017), Vol. 21. No. 2
[16] Tibshirani, R., “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society. (1996). Series B (Methodological), 267-288.
[17] Ng, A. Y., “Feature selection, L 1 vs. L 2 regularization, and rotational invariance,” In Proceedings of the twenty-first international conference on Machine learning, (2004, July). (p. 78). ACM.
[18] Peng, H., Long, F., and Ding, C., “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on pattern analysis and machine intelligence, (2005). 27(8), 1226-1238.
[19] Breiman, L., Random forests. Machine learning, (2001), 45(1), 5-32.
[20] Kramer, S. C., and Sorenson, H. W., “Bayesian parameter estimation,” IEEE Transactions on Automatic Control, (1988). 33(2), 217-222.