Research Article | Vol. 11, No. 1 / 1 January 2020 / Biomedical Optics Express 480
Biomedical Optics EXPRESS

Whole slide imaging system using deep learning-based automated focusing

TATHAGATO RAI DASTIDAR* AND RENU ETHIRAJAN

SigTuple Technologies, Bengaluru, Karnataka 560102, India
*trd@sigtuple.com

Abstract: The auto focusing system, which involves moving a microscope stage along a vertical axis to find an optimal focus position, is the chief component of an automated digital microscope. Current automated focusing algorithms, especially those deployed in cost effective microscopy systems, often cannot match the efficiency of a skilled human operator in keeping a sample in focus. This work presents an auto focusing system that utilises the recent advances in machine learning, namely deep convolutional neural networks (CNN). It improves upon prior work in this domain. The results of the focusing algorithm are demonstrated on an open data set. We describe the practical implementation of this method on a low cost digital microscope to create a whole slide imaging system (WSI). Results of a clinical study using this WSI system are presented. The study demonstrates the efficacy of this system in a practical scenario.

1. Introduction

Diagnosis of diseases using manual microscopic review is still a gold standard in many areas. These include peripheral blood smear analysis (haematology), parasite identification in blood (e.g. malaria), analysis of cancer in tissues (histopathology), study of bacteria and other micro organisms (microbiology), and many more.

Automated digital microscopy systems, also known as whole slide imaging (WSI) systems, partially automate the process of review. They capture digital images of the physical slide and create a "virtual slide". This virtual slide can then be reviewed by multiple experts, enabling both remote review and collaborative review, and also opens up the possibility of automated analysis by artificial intelligence (AI) systems [1,2]. The automated focusing process forms the core of a WSI system, which is responsible for moving the sample under the lens along the vertical axis to bring it to an optimal focus position. Research on automated focusing has been pursued for a few decades [3].

However, in spite of the advancement in performance and reduction in cost of both compute systems and camera systems, the cost of state-of-the-art WSI systems remain high. They are typically used in large tertiary care clinical laboratories or research institutions only. On the other hand, the biggest need for a WSI system is in primary care, especially in non urban areas of developing countries, where there is a lack of trained clinical professionals. If a physical sample is digitised, and then reviewed remotely by an expert (who may be located in a tertiary care centre in an urban area), it can help address the problem of scarcity of medical professionals in primary care. It can also help in better handling of seasonal epidemics like malaria or dengue, and early diagnosis of diseases such as tuberculosis, which are widespread in rural and semi urban areas of developing countries. A fast and affordable WSI system, which can scan a wide range of biological sample types, is thus a pressing need.

A typical WSI system uses a 20X objective lens (0.75NA) to capture images of a biological sample [4]. The depth of field of such lenses are usually less than $1\mu m$. Both the topography of the biological sample, and the glass slide underneath can have depth variations. Thus, the microscope needs to be continuously focused as it moves from one field-of-view to another.

#379780 https://doi.org/10.1364/BOE.379780
Journal © 2020 Received 28 Oct 2019; revised 14 Dec 2019; accepted 16 Dec 2019; published 23 Dec 2019

Biomedical Optics EXPRESS

Auto focusing systems can be broadly divided into two categories – reflection based and image based. In this work, we concentrate on image based auto focusing systems only. Image based auto focusing systems typically capture an image of the sample with a camera, through the objective lens, and then calculate a figure of merit (FoM) to judge the quality of focus. Multiple images are captured along the vertical axis (optical axis), and the image with the best FoM value is taken as the "in focus" image. Different FoM measures have been used in the literature, starting with [3]. Commonly used ones include norm of Sobel operator, variance of Laplacian, norm of Boddeke's operator, etc. [5].

These auto focusing systems typically employ one or more FoM, and tries to find the peak of the FoM. However, the mechanical backlash [6], present in most microscope stages, complicates the positioning process, as positions are never fully reproducible. Further, it was demonstrated in [7] that in some cases the FoM curve can have multiple peaks, not necessarily corresponding to the true focus. Thus, FoM based peak finding may lead to capturing out-of-focus images.

In the recent past, convolutional neural networks (CNNs) [8] have been shown to be effective for several types of computer vision applications. This includes object recognition (e.g. [9,10]), object localisation [11,12] and segmentation [13]. They have been successfully applied for classification [1,2] and segmentation [14] of microscopic images as well.

Jiang *et al*. [15] explore the use of CNNs for the purpose of microscope auto focusing. Given a sample image anywhere in the focus stack, the trained CNN should be able to estimate the optimal distance to be moved vertically (either up or down) to reach the optimal focus position. This approach of using a CNN to predict focus distance has the potential of significantly improving the automated focusing speed of a microscope, as it uses a single image anywhere in the Z axis to predict the direction and distance to the optimal focus position, thus avoiding the iterative trial-and-error approach employed by all standard focusing algorithms [5] which use a FoM to estimate image quality. It also has the potential of not getting trapped at false peaks [7] exhibited by most FoM measures in certain sample types. The training and test data set used for training and validating their CNN are publicly available. The test data consists of two different types of samples, prepared with different staining protocols at different sites. It was observed in [15] that the trained CNN has relatively poor performance on the test set, when trained with only the RGB images, especially on the test set with different protocol of preparation. To counter this, they propose the usage of *spectral domain* and *multi domain inputs*. Pinkard *et al*. [16] also explore the use of CNNs for the purpose of focus distance estimation. This work also highlights the lack of generalisation ability across sample types.

A different approach to using CNNs for predicting focus distances is proposed in [7]. Instead of a single image from the focus stack, this method uses two images captured at a fixed vertical distance from one another. The pixel-wise difference in intensity of these two images is fed to the CNN for predicting the defocus distance. This method was shown to achieve state of the art results on the data set provided in [15], without having to use multi-domain inputs. Though it has the extra overhead of capturing an extra image to estimate the focus distance as compared to other approaches [15,16], the enhanced accuracy compensates for the overhead.

This paper presents a method for automated focusing of a digital microscope which harnesses the recent advancements in convolutional neural networks (CNN). It improves upon the earlier work in this particular field [7,15]. The key contributions of this paper are:

- We propose a novel pre-processing step and a weighted loss function for focus distance prediction using a standard CNN architecture suitable for edge devices with limited compute resource.
- Using this technique, we exhibit state-of-the accuracy on a public data set [15].
- We propose a robust algorithm to implement this focusing method on a real life low cost digital microscope, which works around the non idealities like backlash [6], empty areas

Biomedical Optics EXPRESS

on a slide, etc. The cost of the proposed digital microscope is significantly lower than the typical price of state-of-the-art WSI systems, and thus ideal for use in primary care.

- We showcase the use of this low cost microscope, with our focusing algorithm, in a clinical study of microbiology samples. The sample type was chosen with rural and semi urban areas of developing countries in mind, where there is a wide prevalence of diseases caused by bacteria. In the study, concordance is measured between the results of physical slide review and digital slide review by specialised medical professionals. The study shows excellent concordance between the two methods of review.

The paper is organised as follows: Section 2 describes the proposed method. Section 3 presents experimental results on an open data set. Section 4 describes the implementation of the method on a cost effective digital microscope system. Section 5 provides details of the clinical study and its results. Finally, Section 6 concludes the paper.

2. Proposed method

2.1. Data set

The data set [15] consists of several microscopic fields of view of biological samples, captured with a 20X lens and a 5 megapixel camera. Though images with different sources of illumination are available in the data set, for the purpose of this work we consider only the images captured with a white light emitting diode (LED) illumination. For each field-of-view (FOV), approximately 40 images are captured with defocus distance varying from $-10\mu m$ to $+10\mu m$ in steps of $0.5\mu m$. A total of 128,699 patches of size 224×224 are available for training. The data is split into a 102,960 training set and a 25,739 validation set, ensuring that the two sets contain images from *different* samples. The test set consists of two types of samples – one prepared with the same staining protocol as the training set (698 images), and the other prepared using a different staining protocol at a different site (1,313 images).

2.2. Focusing algorithm and the difference image

2.2.1. Focusing algorithm

We propose a method which uses *difference* between two images in the focus stack to predict the distance to the optimal focus position. The steps of the method can be described as follows:

1. At every new field-of-view, the microscope stage will be moved to a known position which is likely to be *below* the focus plane. Let us call this position as z_1. An image (I_1) is captured at z_1. The vertical position of the best image at the previous field-of-view can be taken as an estimate of the focal plane position at the current field-of-view, and z_1 can be chosen to be a known distance below this estimated position.

2. The stage is then moved *upwards* by a known distance Δ to a new position $z_2 = z_1 + \Delta$. Again, an image (I_2) is captured at position z_2. The distance Δ should be greater than the depth-of-field of the microscope which is typically around $1\mu m$.

3. A pixel-wise *difference image* I is computed as $I = I_2 - I_1$.

4. The difference image I is given as input to a deep CNN to predict the optimal focus position. I is pre-processed as described in Section 2.2.3 before being fed to the CNN model.

We choose Δ as $2\mu m$ for training the model on the open data set. $2\mu m$ is greater than the depth-of-field of the microscope used for generating the data set (approximately $0.8\mu m$) and yet it is not big enough to cause significant change in image properties.

Biomedical Optics EXPRESS

During training, for an image I_1 in the training set with defocus distance z_1, we choose I_2 as the image at defocus distance $z_1 + \Delta\mu m$ (if it exists in the data set). The image I is then computed as $I = I_2 - I_1$ and the *corresponding defocus distance* is taken as z_2. If I_2 does not exist in the data set, I_1 is not used for training. The same convention is used for test images as well.

2.2.2. Properties of the difference image

Using the difference of two images has several advantages: It eliminates, to a large extent, the coarse colour information in the image, which is known to be a major source of over fitting [2] on staining characteristics peculiar to a site. It also emphasises local variations, which are important to gauge the defocus distance, and suppresses global features. This lends it more generalisation ability across sample types, as only the *edges* of structures are learnt by the CNN, not the characteristics of the features of a sample type. Further, the *defocus direction* is encoded in the difference image, if z_1 and z_2 are on the same side of the optimal focus position. This is illustrated in Fig. 1, where one can notice a distinct difference in the nature of the object edges for the two different defocus directions – one being below and the other above the optimal position. It was shown in [7] that CNN prediction accuracy using the difference image significantly outperforms predictions based on a single image only. Hence, we use the same scheme in this work as well.

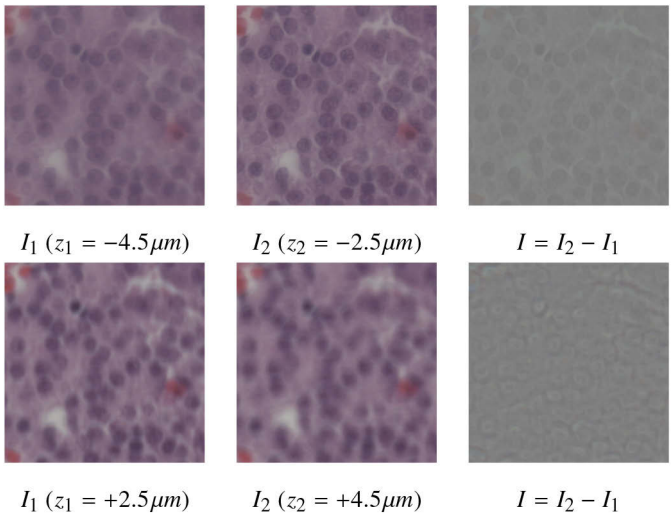


Fig. 1. Two examples of the difference image. The top row shows two images with defocus -4.5 and $-2.5\mu m$, and the difference of the two. Similarly, the second row shows images with defocus 2.5 and $4.5\mu m$, and their difference. Notice the distinct difference in the properties of the edges between the two difference images.

2.2.3. Image pre-processing

The image I is pre-processed as follows before being fed to the CNN:

1. The images I_1 and I_2 are smoothed with a median filter of size 3×3 to reduce local noise prior to the difference operation. Multiple filter sizes were tried. The 3×3 size produced best results.
2. The difference image I is again smoothed with the same filter.
3. A channel-wise contrast normalisation is done on the smoothed difference image. Each channel in the image is centred to 0 (by subtracting the mean of the channel), and divided by the corresponding standard deviation.

Biomedical Optics EXPRESS

2.3. Model architecture

Recently, many "light weight" CNN architectures have been proposed [10,17]. They have low computational cost and memory footprint. This makes them suitable for inference tasks on edge devices with low compute power. For this work, we use the MobileNetV2 [10] architecture to create our CNN. This application is developed with an edge device (an automated microscope) as the target. Hence the choice of the base architecture. This work uses the MobileNetV2 network, except the topmost classification layer. The output of the last feature map of MobileNetV2 is flattened and fed into a dense layer with a single output and no activation function.

2.4. Model training

The model training process introduced in this work helps us surpass the existing state-of-the-art results on this data set described in [7], although both methods use the same CNN architecture. This is described next.

2.4.1. L1 loss

A least square regression loss function (L2 loss) was used in previous work [7,15] to train the deep CNN. However, it has been noted in the literature that a least square loss can cause instability in the model training process during the initial epochs, when gradients are high. To counter this, many previous works, especially those for object localisation in images [12,13], propose the use of a smooth L1 loss function for regression tasks, which is less susceptible to outliers. We use a smooth L1 loss function $\mathcal{L}_1$ which is defined as:

$$\mathcal{L}_1(x) = \begin{cases} \frac{1}{2}(|x| - 0.25) & \text{if } |x|>0.5 \\ \frac{1}{2}x^2 & \text{otherwise} \end{cases} \tag{1}$$

Our observation is that this loss function significantly speeds up the model training process compared to [7].

2.4.2. Sample weights

In addition to the smooth L1 loss, we introduce a weight component in the loss function. We observed that the model predictions are usually better for images which have well defined objects and features (cells, or boundaries between the tissue and background glass). On the other hand, prediction quality was poorer for images which were mostly empty, or had no distinguishing features. This is not easily apparent in the open data set, as it was curated to remove most of the empty images. However, in a real life scenario, such empty areas are common. Our aim here is to increase the penalty for a poor prediction on an image with ample features (and thus probably more clinically significant), while reduce the penalty for smoother images with less features.

The *standard deviation* of the pixel values in an image is a good indicator of the amount of features in it. Furthermore, we observed that the standard deviation of a field-of-view remains relatively unchanged, irrespective of how well focused the image is. Here, we use the standard deviation of the pixel values of either I_1 or I_2 as the weight for that sample.

In addition to giving more importance to images with well defined features, we want the focusing system to be more accurate when the optimal focus distance is nearer. The motivation is driven by considerations for the practical implementation of this system. When the predicted distance is large, we use a two step approach. The stage first moves by a shorter distance in the predicted direction. Then a second estimation is done. The second estimation occurs at a position which is expected to be closer to the optimal focus position. On the other hand, we want to avoid a re-estimation when the predicted distance is small. Thus, we want the prediction quality to be better for smaller distances.

Biomedical Optics EXPRESS

While training, the distance-wise weight w for a sample is defined as:

$$w = \frac{1}{\log(|d| + 1) + 1} \tag{2}$$

where d is the ground truth distance of the sample to the optimal focus position expressed in microns. The logarithm of $|d|$ is used so that the growth of the denominator is sub-linear. Introducing this term makes little difference while evaluating on the open data set, but causes significant improvements in the practical scenario.

The final loss function for a minibatch is thus:

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sum_{i=1}^{N} w_i s_i \mathcal{L}_1(y_i - \hat{y}_i)}{\mathbf{w}^\top \mathbf{s}} \tag{3}$$

Here, N is the minibatch size, $\mathbf{y}$ is the vector of ground truth focus distances for the minibatch, $\hat{\mathbf{y}}$ is the vector of predicted focus distances, $\mathcal{L}_1$ is the smooth L1 loss function, $\mathbf{s}$ are the standard deviations of the patches, and $\mathbf{w}$ are the distance-wise weights.

2.4.3. Multi phase training

We employ a two phase training process. In the first phase, we include only the standard deviation weight terms for the sample and leave out the distance wise weights. The motivation for this is to allow the model to learn features for all defocus distances. The model is trained with stochastic gradient descent (SGD), with momentum 0.9, initial learning rate 0.001 and batch size of 32. The learning rate is reduced by half every 20 epochs. We allow the training to run till convergence and save the model from the epoch which gave the lowest loss on the validation set.

In the second phase, we take the best model from the first phase and retrain it by introducing the distance wise weight term in the loss function. The initial learning rate is taken as the learning rate in the first phase at the epoch which yielded the best model. The learning rate annealing scheme remains the same. We let the training run till convergence in this phase also.

3. Experimental results on the open data set

In this section, we present the experimental results on the open data set provided by [15]. Two versions of the model – the output of both phases of the two phase training method – were tested, and the results are compared with both [15] and [7]. As mentioned in Section 2.4.3, the first phase model is trained with L1 loss and only the standard deviation of the image patches as weights. The second phase model also includes the distance-wise weight term in the loss function. The test images are larger in size (1224×1024). They are split into tiles of size 224×224 pixels. The *weighted average* of the predicted focus distance for these tiles, using the standard deviation of pixel values of each patch as the weights, is taken as the focusing distance of the overall image.

A machine with 6-core Intel Xeon 2.6GHz processor, 60GB RAM and a single NVIDIA Tesla K80 GPU with 12GB memory was used for this work. Software included the Linux operating system (Ubuntu 16.04), NVIDIA Cuda 9.0, CUDNN 7.7. The Keras deep learning package (version 2.2.4) was used with Tensorflow (version 1.11.0) backend.

Results of the model on the test data set prepared with both staining protocols is presented in Table 1. Here, "same protocol" refers to the test set which was prepared at the same site, using the same staining method as the train data set. "Different protocol" refers to the test set which was prepared at a different site using a different staining protocol. As can be seen, both the models from the current work far exceeds the previously reported performance. The improvement is more marked for the test set prepared with a different protocol. This indicates a significant enhancement in the generalisation ability of the model. Results on individual samples are not presented for the sake of brevity.

Biomedical Optics EXPRESS

Table 1. Focusing errors (absolute difference between predicted and ground truth focusing distance for an image) obtained using RGB only images for training. The figures for [15] refer to results on incoherent illumination images only, whether RGB only or multi domain. Figures represented as mean error ± standard deviation of error.

Method	Focusing error on same protocol (μm)	Focusing error on different protocol (μm)
Jiang *et al*. [15]	0.46 ± 0.34	0.53 ± 0.59
Our previous work [7]	0.22 ± 0.25	0.36 ± 0.37
First phase model	0.20 ± 0.18	0.25 ± 0.27
Second phase model	**0.19 ± 0.18**	**0.25 ± 0.26**

The improvement of the second phase model over the first phase model is marginal at best. This is because the data set only contains defocus distances within $\pm 10 \mu m$. In the practical scenario, however, we see a major advantage of using the second phase model (see Section 4.3.3).

4. Implementation on a real life digital microscope

In this section, we describe the implementation of this focusing algorithm on a real life low cost automated digital microscope.

4.1. Description of the hardware

The hardware consists of the following components:

- **Compute system**: A mini-ATX motherboard with Intel i5 quad core processor, 8GB RAM, NVIDIA GPU with 4G GPU memory, running Ubuntu Linux (v16.04).
- **Optics system**: Consists of an optical tube (40X or 100X Plan Achromat objective and 10X eyepiece), and Abbe Condenser with white LED source. A 13MP USB 3.0 camera is used.
- **Hardware control**: A small PCB designed to receive USB commands and drive motors and LED.
- **XYZ slide stage**: The XYZ platform is built using commercially available low-cost ball screws and stepper motors, along with some machined parts. The main movement specifications are:
  - $100 \mu m$ of X and Y axes positioning accuracy (including the mechanical backlash)
  - $0.5 \mu m$ of Z axis positioning accuracy in single direction movements
  - Up to $20 \mu m$ of movement backlash in Z axis while reversing direction

The bill-of-material (BoM) of the above hardware is less than USD 7,000 when components are purchased at retail prices, significantly lower than the price of a state-of-the-art WSI system.

While we implement our algorithm on this particular microscope and use it for a clinical study, the algorithm is not specific to this device only. The same algorithm can be easily implemented on any device consisting of an optical column with a digital camera, an electronically controllable XYZ stage with sufficient precision, and a compute unit.

4.2. Data generation and model training

Training data was generated with several types of physical slides obtained from a clinical laboratory. All samples were anonymised before use. Sample types include peripheral blood smear, stained semen, histopathology and microbiology. Data was generated by stepping the

Biomedical Optics EXPRESS

stage upward with step size of approximately $1\mu m$, and capturing the image at each position, at multiple locations on the slide. A 40X objective lens was used. The central part of the image was cropped, downsampled by $4X$, and then split into 4 patches of size 224×224, each of which were used as independent samples. Identification of the sharpest image in a stack is done automatically using the variance of Laplacian FoM. A manual review is done of the identified sharpest image, to eliminate false peaks.

Unlike $\pm 10\mu m$ as in the open data set, we generated data for approximately $\pm 25\mu m$ for each stack. The final training set consists of approximately 200,000 training samples 50,000 validation samples. A model with the architecture as described in Section 2.3 was trained with this data, with weighted L1 loss and two phase training. MobileNetV2 provides a tunable parameter α to control the size of the feature maps, and thus the overall model. Default value of α is 1. It was shown in [7] that a smaller value of α can significantly reduce the memory footprint of the model without affecting prediction accuracy by much. For this model, we use $\alpha = 0.5$.

For our experiments on the open data set, Δ (the distance between the pair of focus estimation images) was chosen as $2\mu m$ (see Section 2.2.1). However, the choice of a relatively small Δ also meant that the dynamic range of the difference image I is very small, often comparable to the random noise range, since I_1 and I_2 are very similar to each other. This made the trained model very sensitive to compression artefacts [7]. Here, we experimented with different values of Δ, from 2 to 5 microns. We observed that the prediction accuracy on our validation set increases with increasing Δ, and the sensitivity to noise also reduces. However, there are practical difficulties in using a larger Δ in a real microscope. A large Δ might cause many instances where z_1 is below the optimal focus position and z_2 is above. This will involve change in direction of the stage, and thus result in backlash effects. Thus, we chose $\Delta = 3\mu m$ for the real life implementation.

4.3. Implementation of the focus algorithm

In this section, we describe various optimisations that were implemented to make the algorithm work reliably on a real life system.

4.3.1. Initial focus plane identification

When the scan starts, we first estimate an approximate location of the focal plane for the slide. This is done by moving the stage upwards in short steps, from a low starting position which is guaranteed to be below the focal plane. An image is captured at each step. The position at which the sharpest image (as measured by the maximum of the variance of Laplacian) is obtained, is taken as the approximate focal plane location.

4.3.2. Avoiding backlash

Once the focal plane is identified at the first field-of-view as described above, for all subsequent positions in the slide we use the focusing algorithm shown in Section 2.2.1 with some important changes to work around the effects of backlash [6].

Backlash occurs when a stepper motor changes direction. The actual distance moved, right after a direction change, is not equal to (and usually less than) the commanded distance. When the stage moves to a new field-of-view, the focal plane position at the previous field-of-view is taken as a rough estimate of the focal position at the new field-of-view. Let us call this estimate as the "pivot" position. The position z_1 (where the first estimation image I_1 is captured) is selected as a known distance below this "pivot" position. Going from z_1 to z_2 ($= z_1 + \Delta$) involves a change in direction as the stage had to first move downwards to reach z_1 and then subsequently up to reach z_2. This causes backlash error, and the achieved position z_2 will likely not be exactly $z_1 + \Delta$. This can cause the CNN model predictions to be inaccurate.

To avoid this backlash error, we introduce an extra step at the beginning. The stage is first moved to a position z_0 which is *lower than* z_1, and then moved upward by a known distance to

Biomedical Optics EXPRESS

reach z_1. Due to backlash, the actual distance between z_1 and z_0 may be variable. But since there is no change in direction involved while moving from z_1 to z_2, there is no backlash error, and the difference between them is guaranteed to be Δ. The position z_0 should be chosen such that the distance from z_0 to the "pivot" is greater than the maximum backlash error of the stage, to completely eliminate any effect of backlash while moving from z_1 to z_2. The maximum backlash error can be estimated at the time of setting up the device for the first time, using the method described in [6].

Another source of backlash error is when the optimal focus position, as predicted by the CNN, is *below* z_2. In that case, the stage has to change direction and move down, thereby causing backlash. To counter this, whenever the predicted position is below z_2, we move the stage down by a distance more than twice the predicted distance, and repeat the focus estimation process.

While focusing at any new field of view, the estimated focus position from the previous field of view is taken as the "pivot position". For our implementation, we chose z_0 to be $22\mu m$ below the pivot position. This is greater than the maximum Z backlash of the device (approximately $20\mu m$). z_1 is chosen to be $8\mu m$ below the pivot position, and Δ is taken as $3\mu m$. The value of z_1 is chosen based on the performance of the CNN model on the training data – by identifying the initial distance range for which the predictions are most accurate. These values remain constant for all fields of view.

4.3.3. Weighted average-based prediction and prediction quality estimation

In the focus distance estimation process, images I_1 and I_2 are captured and the difference image I is computed. Then, the central patch is taken, down sampled by 4X, and split into 4 patches of 224×224 (similar to the training data generation process). The model is invoked independently on each of these 4 patches, which leads to 4 independent predictions of the optimal focus distance. We use the *weighted average* of the predictions as the final predicted distance, where the weights are the standard deviations of the patches, computed from either I_1 or I_2.

Unlike a classification network, a regression network does not output a prediction probability which can act as a measure of the model's confidence. We use the divergence of the 4 predicted distance values as a measure of confidence. The standard deviation of the 4 predicted values is computed. If the standard deviation is greater than a threshold (we use $1\mu m$), then the model prediction is taken to be "uncertain". In such cases, the stage is moved in the predicted direction, but by half the predicted distance. Then the distance estimation process is repeated.

We also take advantage of the greater accuracy of the model for nearer distances (due to the two phase training process). If the predicted distance magnitude is greater than $12\mu m$, we treat the prediction as uncertain. Here, too, the stage is moved in the predicted direction by half the predicted distance, and the estimation process is repeated.

5. Results of a clinical study

5.1. Introduction

In this section, we present the results of a clinical study using the automated digital microscope described earlier. The clinical area of study was chosen as microbiology. The reason for the choice is that most commercial WSI systems are typically not used for microbiology, while digitisation and remote review has great applicability in this field. This is especially true in developing countries which have a wide prevalence of diseases caused by bacteria.

The gram stain, developed by Hans Christian Gram in 1884, has the widest application in any bacteriology laboratory. The staining procedure differentiates most bacteria into two types – gram positive and gram negative. Gram-negative bacteria are stained pink and gram-positive bacteria are stained purple with the gram stain. The gram stain can be performed on smears made from sputum, urine, faecal specimens and pus swabs.

Biomedical Optics EXPRESS

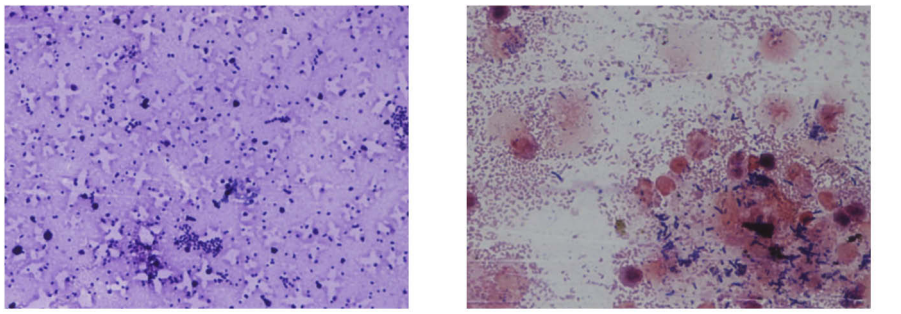
Gram Stain is analysed to detect the presence and number of micro organisms, their description, epithelial cells and pus cells. The findings are reported as per the protocol shown in Table 2. Manual microscopic analysis is typically performed with 100X objective lens in oil immersion. The grade is based on the reading of more than 10 fields.

Table 2. Grading methodology used while reporting findings on gram stain smears.

Grading	Grade description	Number of cells	Number of bacteria
1+	Rare	<1	<1
2+	Few	1–5	2–10
3+	Moderate	6–10	11–50
4+	Many	>10	>50

5.2. Study design

In this study, 32 gram stained smear slides were chosen from the normal workload of a clinical laboratory. The samples were anonymised, i.e. patient identifiers were removed. The slides were scanned using our automated digital microscope with a 40X objective lens. A few hundred fields-of-view were captured for each slide, but 30 were chosen randomly per slide for clinical review. A few example images from the digital microscope are shown in Fig. 2. Additionally, the slides were also digitised manually using a standard microscope, at both 40X and 100X objectives, and a CCD camera, by an operator skilled in the operation of a microscope. In the manual mode, 30 random fields-of-view were captured per slide. Due to lack of resources, only 29 out of 32 slides could be imaged manually. Two certified medical professionals then reviewed, in a blinded manner, the physical smears under a microscope, as well as the digital images from the three different sources listed above. The review consisted of identification and characterisation of gram positive cocci and gram negative bacilli. The purpose of the study was to measure the concordance between manual review and the three different modes of digitisation.



Gram positive cocci Gram positive cocci and gram negative bacteria

Fig. 2. Two example fields of view from two slides captured through the automated microscope. The image on the left shows Gram positive cocci, while the image on the right shows both gram positive cocci and gram negative bacteria.

5.3. Results

Concordance of findings with observer 1 and observer 2 are shown in Tables 3 and 4 respectively. A reading by an observer is taken to be concurring if the report from the manual review of the smear by the same observer matches that from the review of the digital images. It is taken to be non concurring otherwise. In multiple cases, the review of the digital images led to the

Biomedical Optics EXPRESS

identification of certain organisms which were not detected in manual review. Such cases were taken as concurrences. In the below tables, the number of such cases are captured in the column titled "Cases where digital review surpassed manual review". The concurrence of the automated digitisation process surpassed that of the manual digitisation for both the observers. In cases where an organism was detected in manual review, but not seen in the digitised slides, we observed that the organisms were imaged, but they did not occur in the randomly selected 30 images that we used.

Table 3. Comparison of the three digitisation methods against the first observer.

Observer 1	Concurrences	Non concurrences	Cases where digital review surpassed manual review	% Concurrences
Automated 40X	28	4	6	**87.5**
Manual 40X	24	5	5	82.75
Manual 100X	23	6	5	79.31

Table 4. Comparison of the three digitisation methods against the second observer.

Observer 2	Concurrences	Non concurrences	Cases where digital review surpassed manual review	% Concurrences
Automated 40X	30	2	3	**93.75**
Manual 40X	26	3	4	89.66
Manual 100X	27	2	4	93.10

Another observation to be noted here is that the report from the manual microscopic review of the smears from the two observers concurred in **24 out of 32** cases only. This high level of inter observer variability highlights the challenges associated with the task, and the need for digitisation to enable easier multiple reviews, and automated analysis in future.

As part of this exercise, we measured the speed of image capture with this device. The device could capture images to completely cover 15mm×15mm region of the slide in approximately 8 minutes using a 40X objective lens, and between 2 and 2.5 minutes using a 20X objective lens. This makes it comparable in performance with commercially available WSI systems. We were also able to scan different other types of biological samples, e.g. bone marrow and body fluids with this device, even though such samples were not used for training the CNN. However, we have not performed any formal clinical validation on such samples yet.

6. Conclusion

This paper presented a new method for applying deep learning for focus distance estimation in automated digital microscopy. The method is shown to be superior to results in existing literature in terms of generalisation error over multiple staining protocols, as demonstrated on a publicly available data set. The practical implementation of this method on a low cost automated digital microscope was presented, which includes methods to work around hardware non idealities like backlash. The low cost of the hardware component makes this system ideally suited for use in resource constrained scenarios in developing countries, where there is a dearth of both quality medical equipment and certified medical professionals in non-urban areas. We presented the results of a clinical study in the field of microbiology, which shows good concordance of the manual and digital image review process. This showcases the effectiveness of the digitisation method, and its utility for standardising the reporting workflow for gram stain smears.

Future work will include more extensive clinical studies with different types of biological samples and larger sample sizes. We will also investigate further into improving the CNN

Biomedical Optics EXPRESS

performance, and gaining deeper understanding on exactly what the CNN is learning when it learns to predict focus distance.

Funding

SigTuple Technologies Private Limited.

Disclosures

The authors declare that there are no conflicts of interest related to this article.

References

1. D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-assisted Intervention* (Springer, 2013), pp. 411–418.
2. D. Mundhra, B. Cheluvaraju, J. Rampure, and T. R. Dastidar, "Analyzing microscopic images of peripheral blood smear using deep learning," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, (Springer, 2017), pp. 178–185.
3. J. F. Brenner, B. S. Dew, J. B. Horton, T. King, P. W. Neurath, and W. D. Selles, "An automated microscope for cytologic research a preliminary evaluation," J. Histochem. Cytochem. **24**(1), 100–111 (1976).
4. E. Abels and L. Pantanowitz, "Current state of the regulatory trajectory for whole slide imaging devices in the usa," J. Pathol. Inform. **8**(1), 23 (2017).
5. R. Redondo, G. Cristóbal, G. B. Garcia, O. Deniz, J. Salido, M. del Milagro Fernandez, J. Vidal, J. C. Valdiviezo, R. Nava, and B. Escalante-Ramírez, "Autofocus evaluation for brightfield microscopy pathology," J. Biomed. Opt. **17**(3), 036008 (2012).
6. F. R. Boddeke, L. J. Van Villet, and I. T. Young, "Calibration of the automated z-axis of a microscope using focus functions," J. Microsc. **186**(3), 270–274 (1997).
7. T. Rai Dastidar, "Automated focus distance estimation for digital microscopy using deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019).
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Adv. Neural. Inf. Process. Syst. **25**(2), 1097–1105 (2012).
9. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv e-prints (2014).
10. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," *IEEE Conference on Computer Vision Pattern Recognition*, pp. 4510–4520 (2018).
11. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision Pattern Recognition* pp. 779–788 (2016).
12. S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* (2015), pp. 91–99.
13. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE conference on computer vision (ICCV)* pp. 2980–2988 (2017).
14. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Int. Conf. on Med. Image Computing Computer-assisted Intervention*, pp. 234–241 (2017).
15. S. Jiang, J. Liao, Z. Bian, K. Guo, Y. Zhang, and G. Zheng, "Transform and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging," Biomed. Opt. Express **9**(4), 1601–1612 (2018).
16. H. Pinkard, Z. Phillips, A. Babakhani, D. A. Fletcher, and L. Waller, "Deep learning for single-shot autofocus microscopy," Optica **6**(6), 794–797 (2019).
17. F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size," arXiv e-prints (2016). ArXiv:1602.07360.