Computer-Automated Malaria Diagnosis and Quantitation Using Convolutional Neural Networks

Courosh Mehanian[1], Mayoore Jaiswal[1,2], Charles Delahunt[1,2], Clay Thompson[3], Matt Horning[1], Liming Hu[1], Shawn McGuire[1], Travis Ostbye[1], Martha Mehanian[1], Ben Wilson[1], Cary Champlin[1], Earl Long[4], Stephane Proux[5], Dionicia Gamboa[6], Peter Chiodini[7], Jane Carter[8], Mehul Dhorda[9], David Isaboke[8], Bernhards Ogutu[10], Wellington Oyibo[11], Elizabeth Villasis[6], Kyaw Myo Tun[12], Christine Bachman[13], David Bell[13]

[1]Global Good Research, [2]University of Washington, [3]Creative Creek Software, [4]LSHTM, [5]SMRU, [6]UPCH, [7]HTD, [8]Amref, [9]WWARN, [10]Kemri, [11]University of Lagos, [12]DSMA, [13]Global Good Fund
{cmehanian, mjaiswal, cdelahunt, mhorning}@intven.com

Abstract

The optical microscope remains a widely-used tool for diagnosis and quantitation of malaria. An automated system that can match the performance of well-trained technicians is motivated by a shortage of trained microscopists. We have developed a computer vision system that leverages deep learning to identify malaria parasites in micrographs of standard, field-prepared thick blood films. The prototype application diagnoses P. falciparum with sufficient accuracy to achieve competency level 1 in the World Health Organization external competency assessment, and quantitates with sufficient accuracy for use in drug resistance studies. A suite of new computer vision techniques—global white balance, adaptive nonlinear grayscale, and a novel augmentation scheme—underpin the system's state-of-the-art performance. We outline a rich, global training set; describe the algorithm in detail; argue for patient-level performance metrics for the evaluation of automated diagnosis methods; and provide results for P. falciparum.

1. Introduction

Automated detection of malaria in field-prepared blood films is a challenging computer vision task with potential benefit for millions of people. Half of the world's population are at risk for contracting malaria, with an estimated 212 million cases in 2015 [1]. A majority of the 429,000 deaths from malaria in 2015, mostly of young children, are attributable to *P. falciparum*. Four other *Plasmodium* species—*P. vivax, P. ovale, P. malarie,* and, rarely, *P. knowlesi*—also infect humans [2].

Microscopy continues to be regarded as a standard for malaria diagnosis and quantitation [3], in part because it can be used to detect other infectious diseases [4], has low incremental cost, is widely available, can measure parasite density, and can identify malaria species. Microscopy can detect low density infections if enough blood is scanned, but this is time-consuming, difficult, and tedious due to the low density and small size of parasites as well as the abundance of similar non-parasite objects, as illustrated in Figure 1. To be effective, microscopy needs well-trained staff for consistent slide preparation and examination. In areas with poor quality control, microscopy can produce inaccurate results [5] resulting in inappropriate treatment.

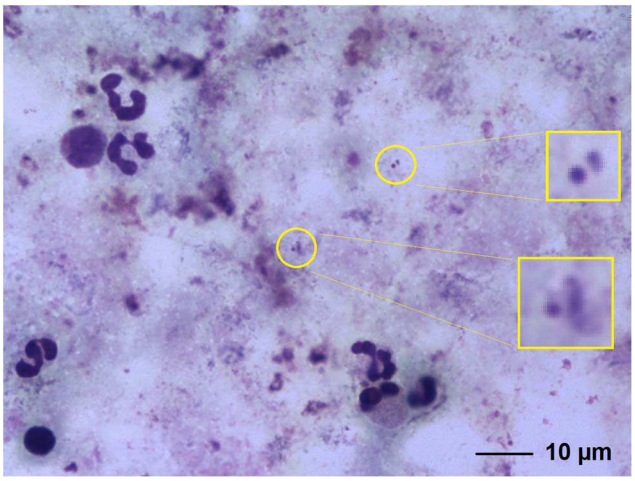

Figure 1. Typical thick film microscope image. This field-of-view image contains only two parasites indicated by yellow circles with enlargements at the right.

In addition, evolving drug resistance is an increasing concern. The World Health Organization (WHO) encourages regular monitoring of antimalarial efficacy in malaria-endemic countries [6]. Microscopy remains the most field-practical tool to accurately monitor response to therapy, due to its capacity for accurate quantitation, since

parasite clearance rates are the most commonly used measure of drug efficacy [7]. But quantitation is a time- and labor-intensive measurement requiring the reading of many blood films [8].

A major difficulty with using microscopy in drug efficacy monitoring is the shortage of trained experts in regions where malaria is endemic [9]. Therefore, the development of a computer vision system to aid in malaria diagnosis and quantitation is an appealing research goal, both because of the difficulty of the task and the high potential benefit. In addition, it is an attractive target for application of convolutional neural networks (CNNs), which have shown success in other image classification tasks [10-13]. Before addressing automated malaria diagnosis via computer, we present a brief overview of malaria blood film microscopy.

1.1. Blood film microscopy

Two types of blood films are used to diagnose malaria: thick film and thin film. Here, we will mainly be concerned with thick films because they provide a sufficient volume of blood to enable reliable diagnosis of low parasite density infections [14]. We have also developed a thin film module, which will be presented in a subsequent publication.

The thick film is prepared by placing a drop of blood (about 2 µL) on a slide and using the corner of another slide to spread the drop in a circular pattern to ~1.2 cm diameter. The slide is then dried and stained with a Romanovsky-type stain, e.g. Giemsa [2], then rinsed and dried again. Giemsa results in DNA (e.g. nuclei) staining purple and RNA (e.g. cytoplasm) staining blue.

We confine our discussion here to *P. falciparum*. The most commonly found parasite stages in *P. falciparum* positive blood films are ring forms (immature trophozoites). In Figure 2, a number of examples are shown in finer detail. The small, round, purple disk, found in most of the thumbnails, is the nucleus of the parasite; the wispy blue-gray shape in close proximity is the cytoplasm. Later stage trophozoites (Figure 2, lower-left) do not have a clear, round nucleus and distinct cytoplasm. Note the variety in

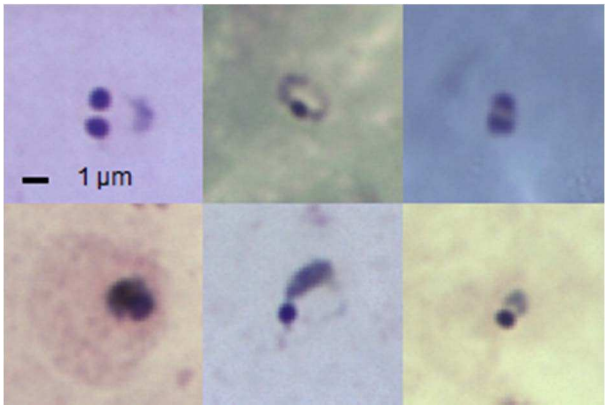

Figure 2: Ring form *P. falciparum* malaria parasites.

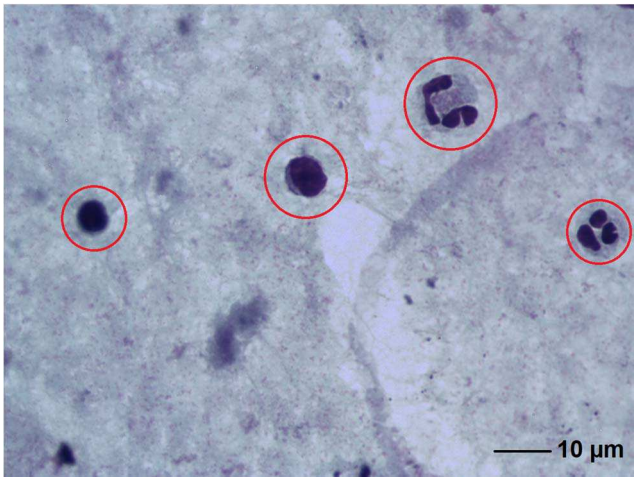

Figure 3. An FoV image of a negative sample, i.e. with no malaria parasites. WBCs are indicated with red circles.

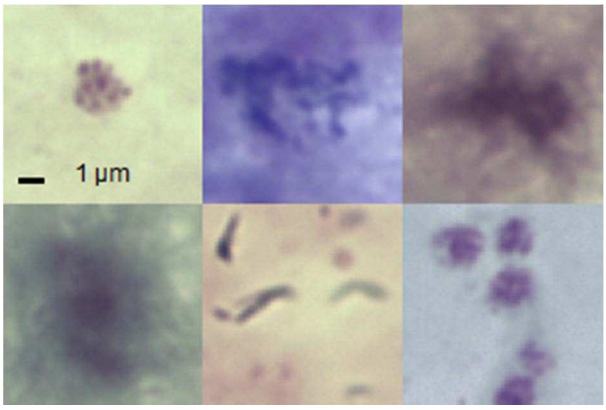

Figure 4. Examples of distractors. The objects in the upper left and lower right corner are platelets.

appearance of trophozoites. All these must be recognized as *P. falciparum* parasites and must be distinguished from non-parasites.

Uninfected normal human red blood cells (RBCs), which are lysed during staining of the thick film, contain neither DNA nor RNA and do not appear dark blue or purple; thus Giemsa provides good contrast between parasites and background. Nevertheless, interpretation of thick films is challenging because of a noisy and variable background. Example fields-of-view (FoVs) are shown in Figures 1 and 3. Note the white blood cells (WBCs) in Figure 3, whose nuclei are stained similarly to the parasite nuclei in Figure 2. Various non-parasite components of the thick film can also absorb stain, creating artifacts that may mimic parasites, and the stain itself can self-aggregate. Collectively, artifacts and objects that are difficult to distinguish from parasites are termed distractors, examples of which can be seen in Figures 1, 3, and 4.

In this application, images of blood films are acquired with a digital scanning microscope. The nucleus of the trophozoite can be as small as 1 µm in diameter, and other

components used to detect malaria parasites and identify species can have features smaller than 250 nm, which is close to the optical limit of resolution. To resolve features of this size, a high numerical aperture (NA) oil immersion objective, for example $100 \times NA = 1.2$, is required. At these high numerical apertures, the depth of field is on the order of 0.5μm. To detect parasites throughout the depth of the thick film, and to ensure all fields of view are in focus, images must be captured at multiple focal planes spanning the entire depth of the blood film. An FoV refers to a stack of images centered at a particular x, y location in the slide, at one or more focal depths z. To achieve reliable detection at low parasitemia levels, WHO recommends inspecting 100 thick film FoVs before declaring a sample negative [2]. (This pertains to a manual $100 \times$ microscope, whose FoVs tend to be larger than automated microscope FoVs.)

1.2. Related work

Several proposals for computer-automated reading of malaria blood films (thin or thick) have appeared in the literature in the past few years [15-20]. Many of these studies have not presented realistic assessments of the field-effectiveness of their algorithms, due to a lack of emphasis on patient-level metrics. Full reviews of these publications may be found in [21, 22]; here we merely offer a brief overview of a few of these proposals. In Section 3.2, we present a summary analysis of key metrics for all of these systems and a comparison to ours.

An automated system for the diagnosis of malaria from thick blood smears is proposed in [15]. They crop random, overlapping patches from good images (discarding out-of-focus images) representing blood smears of 133 patients. Patches containing a parasite (based on expert annotation) are marked as positive and the remaining patches as negative. Traditional feature engineering and an ensemble decision-tree classifier form the core of their system. The classifier applied to the test set achieves an area-under-the-curve (AUC) metric of 0.97. The authors evaluate this result purely on the object level, reporting that it achieves 20% recall at a precision of 90%. There is no reference to parasitemia level. (The significance of parasitemia in relation to sensitivity and specificity is discussed in Section 2.6). Random assignment of images to the train and test sets implies that these sets were not disjoint at the patient level and therefore, the reported metrics are not predictive of actual field performance (see Section 2.1).

Among the prior automated malaria diagnosis systems considered here, [17] uniquely does not use Romanovsky staining. Rather it uses a cartridge that accepts a sample of blood and automatically creates a film stained with fluorescent Acridine Orange (AO) [23]. Automatic slide creation and AO staining do hold some advantages, but a reluctance to adopt a new, disposable cartridge and reliance on a fluorescence microscope may prove barriers to field acceptance. The authors do provide patient-level metrics and report a limit of detection of tens of parasites per μL of blood. The quantitation results shown, however, are inadequate for use in drug efficacy studies.

An automated malaria diagnosis system that works on thick film microscope images is described in [20]. They train the system on a dataset consisting of 27 *P. falciparum* positive and 36 negative blood samples. The test set consists of 24 *P. falciparum* positive and 20 negative samples. They do provide patient-level metrics and report achieving WHO competence level 1 diagnosis accuracy, albeit at a parasite density of 300 p/μL. They allude to the use of CNNs for feature extraction, but the results they report use traditional feature engineering (morphological, shape, color, texture, and Haar-like features).

Our system is intended for use under field conditions. Thus, the system has the following requirements and characteristics: (1) accepts standard field-prepared, Giemsa slides; (2) is robust to moderate variability in slide quality; (3) scans a sufficient volume of blood, approximately 0.1 μL, ~300 FoVs; (4) scans at multiple focal planes; (5) has high patient-level sensitivity and specificity at low parasitemia—approaching 100 p/μL; (6) has accurate quantitation in the parasitemia range of 200-200,000 p/μL; and (7) has high object-level sensitivity and specificity.

Our system has a resolution of 11.36 pixels/μm and each FoV is 1280×960 pixels. We scan at 9 focus levels, 0.3 μm apart, and thus 300 FoVs amounts to ~2.5 gigapixels. The system is trained on a large and diverse set of images, where the test set is disjoint from the training set at the patient level. We report patient-level diagnosis and quantitation results (as opposed to merely object-level classification), which are the most important metrics for the system's intended use-cases. Our system achieves WHO competence level 1 [24] for *P. falciparum* diagnosis (Section 3.1.1) and *P. falciparum* quantitation accuracy sufficient to be used for drug resistance studies (Section 3.1.2). We now describe the dataset and methods in detail.

2. Method

Our data processing pipeline consists of a number of modules, each designed with the above requirements in mind. The preprocessing module (Section 2.2) implements a new sample-level global white balance method. The candidate object detection module (Section 2.3) processes multiple focal planes for each FoV (image z-stacks) and is based on a novel adaptive nonlinear grayscale intensity image. The feature extraction module (Section 2.4) incorporates CNNs and introduces a new gamma-transform color augmentation scheme. The classification module (Section 2.5) is designed to allow the system to adapt to local variations, e.g. in slide preparation. Finally, the disposition module (Section 2.6) computes patient-level diagnosis and quantification (a multiple-instance learning problem) employing a learning algorithm calibrated on the statistics of the validation set.

2.1. Data

Large numbers and a great variety of images are needed for training the rich deep learning models in our computer vision system. Diversity in the training and testing data contributes to system robustness under heterogeneous field conditions. And because the patient is the atomic unit for diagnosis, samples from a wide variety of patients and labs are essential to validate diagnostic effectiveness. Some relevant statistics of our malaria blood film library are shown in Table 1. The models in our system are trained on image patches of individual objects (parasites and distractors) from a subset of patients in the library; the system is tested against objects from disjoint subset of patients. (See Section 3 for details on numbers of patients in each subset.) Disjoint training and testing sets at the patient level enable realistic estimates of field performance.

Blood samples	1,452
Fields-of-view	5,707,947
Parasite objects	956,531
Countries of origin	12

Table 1. Summary of thick film malaria database.

2.2. Pre-processing

Histologically stained microscope slides typically display color variation within a slide, between slides of different blood specimens, and between different technicians, laboratories, clinics, and regions. Color variation can result from differences in stain pH, age and purity of stain, duration of the staining procedure, and sensor settings; overall slide hue can range from blue to green to pink to golden. Figures 1-4 illustrate a small fraction of the variability in quality, color, and presentation that is typical of field samples. Uncorrected, color variation may degrade system performance.

White balancing techniques may be used to compensate for some, but not all, of the color variation. Traditional white balancing involves the scaling of red, green, and blue (RGB) pixel values based on the mean color of the brightest pixels in each image individually, which can result in color distortion and exaggerated intra-slide color differences. Our white balancing technique pools the pixels from all FoVs and computes a global color balance affine transform for each blood sample.

2.3. Detection

The detection module generates object proposals—potential parasites to be subsequently scored as parasite or distractor by a classifier. To achieve the target limit of detection, some ~300 FoVs need to be processed by the algorithm, due to the Poisson statistics of rare object distributions. To keep the run time within reasonable limits (roughly 20 minutes on a standard quad-core laptop), the computational complexity of the detection algorithm should be as low as possible.

Most generic object detection methods, such as R-CNN [25], YoLo [26], deformable parts model [27], and selective search [28] are either too complex, too insensitive, or too slow for malaria detection on large numbers of FoVs at multiple focal planes. The deformable parts model performs exhaustive search using a support vector machine (SVM) on a histogram of oriented gradients (HOG) feature pyramid. Selective search uses segmentation on multiple color spaces based on a greedy hierarchical grouping of graphs. This leads ~10K detections per image, which would drastically slow down our framework. Processing flow in R-CNN (and its variants Fast R-CNN [29] and Faster R-CNN [30]) consists of region proposals, followed by classification, followed by post-processing to refine the bounding boxes and eliminate duplicate detections. These complex pipelines are slow. While YoLo processes 45 frames per second, it fails to detect small objects and objects appearing in clusters, which negatively impacts quantitation performance. These shortcomings render these methods unsuitable for malaria parasite detection.

Leveraging domain-specific information allows the design of a specialized detection scheme that out-performs more general methods. As mentioned previously, Giemsa-stained microscope images provide good contrast between deep purple nuclei and background. Thus, malaria parasite nuclei, along with white blood cells, are among the darkest objects in the images; a dark threshold applied to a grayscale intensity image may act as a simple and effective initial detector for malaria parasites.

While this simple detector has high sensitivity, its precision is low: many dark distractors are also detected, which degrades low parasitemia performance because of excessive false positive detections. To enhance the object-level specificity of the detector, we introduce two innovations: adaptive grayscale intensity and dynamic local thresholding. The standard grayscale intensity is a linear combination of red, green, and blue pixel values that approximates the human-perceived luminance [31], but does not necessarily provide the best separation between parasites and background. Machine learning techniques may be used to compute a more optimal projection vector.

We make use of the above-noted similarity in color between parasite nuclei and WBC nuclei. The latter are relatively easy to detect and classify at high precision because they are large and contrast strongly with the background. In a first pass through the FoV images, we segment WBC candidates using a dark threshold tied to grayscale intensity statistics. Morphological and clustering operations further filter individual WBC candidates, which are then classified with a Gaussian-kernel SVM [32]. The segmentation of WBCs enables the collection of RGB color statistics for WBC pixels and a random sampling of

background pixels. Machine learning techniques are then used to compute the optimal projection in RGB space that will separate WBC pixels from background pixels.

The resulting projection, which varies by blood sample, is called the adaptive grayscale intensity. It provides higher precision for parasite detection compared to the standard grayscale intensity. Performance may be further enhanced by adding non-linear terms to the predictor, similar in spirit to polynomial regression. For example, the predictor may be augmented from the linear $\xi = [R, G, B]^T$, to the 2nd order polynomial predictor:

$$\xi = [R, G, B, R^2, G^2, B^2, RG, RB, BG]^T. \qquad (1)$$

More flexible non-linear terms, such as rational functions of the RGB components, may be included, as in the following 12-dimensional non-linear form:

$$\xi = \begin{bmatrix} R, G, B, R^2, B^2, RG, \dots \\ \frac{R}{G+\epsilon}, \frac{R}{B+\epsilon}, \frac{G}{B+\epsilon}, \frac{R+B}{G+\epsilon}, \dots \\ \frac{B-G}{R+G+B+\epsilon}, \frac{G}{R+G+B+\epsilon} \end{bmatrix}^T, \qquad (2)$$

where ϵ is a small constant added to the denominators to prevent overflow. Because of collinearity between the individual components of the predictor, we use regularized regression, such as ridge regression [33], lasso [34], or partial least-squares regression (PLSR) [35]. PLSR with 1 PLS component has performed the best in our experiments. Figure 5 shows a comparison of the detection free-response ROC curves (FROC) [36] using the standard grayscale image vs. the adaptive grayscale image based on the 12-component non-linear predictor of Equation 2. At 98% sensitivity, the adaptive nonlinear grayscale image detects 35% fewer false positives than standard grayscale.

We now address the question of thresholding. Both of the popular thresholding methods [37, 38] assume bimodal intensity histograms. This does not hold when the target objects occupy a negligible fraction of the pixels, as is the case with malaria parasites. Furthermore, a fixed threshold across all FoVs entails compromise between sensitivity and false positive rate. Adaptive thresholding per FoV is better, but still involves compromise because, typically, there are both noisy regions and quiet regions in a single FoV.

Dynamic local (i.e. pixel-wise) thresholding provides the best performance compared to either static or FoV-wise adaptive thresholding. Our thresholding scheme estimates the local noise floor using a large-kernel median filter. WBC pixels (which were detected and classified in the first pass) are replaced with the median FoV pixel value to prevent WBCs from desensitizing the local threshold.

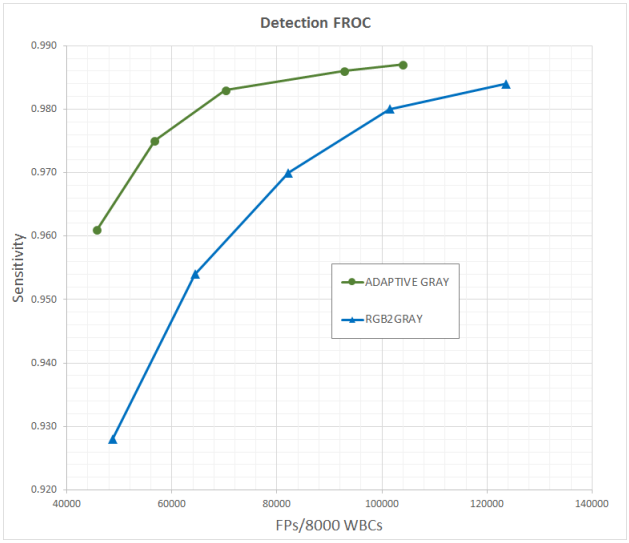

Figure 5. FROC curves for *P. falciparum* detection based on standard grayscale vs. adaptive grayscale images.

The adaptive grayscale intensity image is thresholded—pixel-wise—resulting in a binary image that is processed to detect connected-component blobs. These are treated as candidate objects. Since multiple blobs may be detected in a given z-level for the same object, distance-based clustering is used to associate nearby detected blobs with a single candidate object. In addition, the same object will frequently be detected at multiple z-levels of an FoV. The best-focused version of the object (i.e. z-level image patch with the highest Brenner focus score [39]) is selected. The output of the detector is a list of bounding boxes and thumbnails representing the candidate objects.

Notwithstanding the use of adaptive nonlinear grayscale intensity and dynamic local thresholding, many dark distractors are still detected. For low parasitemia samples, these distractors can overwhelm the number of parasites. To eliminate a large number of obvious distractors, we train a Gaussian-kernel SVM based on low-computational cost geometric, color, gradient, contrast, and texture attributes extracted from thumbnails of the candidate objects, using the database of annotated parasites as ground truth. The classifier achieves AUC of about 0.90 for the both the training and validation sets. The distractor filter threshold is tuned to keep the object-level sensitivity at 95% for training and 90% for validation (and holdout). We have subsequently employed a random forest classifier [40] for the distractor filter with equivalent or better results.

2.4. Feature extraction

We extract features on those candidate objects that survive the distractor filter. The recent widespread adoption of CNNs for feature extraction and classification has led to notable breakthroughs in performance for various computer vision tasks [10-13]. We employ CNNs using the Caffe

Deep Learning Framework [41]. We have experimented with a few CNN architectures, including AlexNet [11], VGG [42], and GoogLeNet [43]. These networks were designed for 1000-category vision problems such as ILSVRC [44], and generally lead to overfitting on our binary classification problem. We therefore use reduced versions, optimizing generalization performance by adjusting numbers of filters and layers in VGG and numbers of filters and inception modules in GoogLeNet.

For the results shown in Section 3, we used a 9 layer VGG architecture (6 convolutional + 2 fully-connected + 1 output). When VGG is used as a feature extractor, the output of 2nd fully connected layer (after dropout) is used as the feature vector. This reduced VGG achieves about 93% accuracy on a validation set for *P. falciparum* vs. distractor. The VGG results are markedly better than those achieved with AlexNet. Subsequent experiments with a reduced GoogLeNet architecture performed roughly equivalent to the reduced VGG.

Training the CNN entails augmentation of data to avoid overfitting. Three different kinds of augmentation are employed. The individual object thumbnails are flipped and rotated in 90° increments which gives 8× augmentation. (Smaller angles are avoided to prevent loss of resolution.) Random positional shifts of ± 5 pixels and random augmentation of individual RGB color channels are also performed. Initially, we employed the color augmentation approach described in [11] but found the resulting colors unrealistic. We opted instead for random gamma correction of individual color channels, which gave more realistic blood smear microscopy image colors as well as improved performance. The number of augmentations used depends on the type of object and is anywhere from 16-64×.

The CNN is trained using equal numbers of (augmented) parasites and distractors and the following Caffe settings: batchsize=128, base_lr=0.001, lr_policy="inv", power=1, gamma=10^{-4}, momentum=0.9, and weight_decay=10^{-5}.

2.5. Classification

One approach is to use the CNN as both feature extractor and classifier. Another option is to use the CNN as feature extractor and a different algorithm as external classifier. The first choice has some advantages, including simplicity, speed, and the fact that the CNN is trained with a large (augmented) number of thumbnails. The second option provides more flexibility in responding to new distractor types or sample preparations discovered in the field. Transfer learning [45-47] assures us that a universal CNN feature extractor, trained on a broad set of samples available in-house, can provide discriminative features in most field settings, while an external classifier can be fine-tuned to local conditions. Initially, the CNN and external classifier are trained on the same in-house samples.

We use logistic regression [48] as the external classifier for two reasons. First, logistic regression mimics the CNN's fully-connected + SoftMax output. Second, the software package [49] implements a robust, large-scale learning algorithm for logistic regression based on stochastic gradient descent. We used this architecture for the results of Section 3.

2.6. Patient-level disposition

In object classification tasks (e.g. ILSVRC [44]), a sample is a single image and the endpoint is object classification accuracy. With malaria diagnosis, a sample is a blood film. For each blood film, the system must process hundreds of FoVs and about 10 focal planes per FoV; and it must detect and classify thousands of object thumbnails. The ultimate goal is to diagnose the patient; metrics of success must reflect this goal. Because object identification is only an intermediate goal, good object-level performance is necessary but not sufficient to assure strong performance on patients. Object-level results are relevant only insofar as they affect patient-level accuracy. Thus we develop and emphasize patient-level methods and metrics.

Our system counts the number of detected objects (which include true positives (TP) and false positives (FP)), then diagnoses the patient according to whether this count exceeds some threshold. For patient-level diagnosis, the figure-of-merit (FoM) is the estimated limit-of-detection (LoD) in parasites/μL at fixed specificity (e.g. 95%). This determines whether the system can correctly diagnose low-parasitemia (and healthy) patients.

Consider the following patient-level quantities:

$$\begin{array}{ll} p & \text{actual number of parasites per μL} \\ q & \text{suspected number parasites per μL} \\ t & \text{number of true positives per μL} \\ f & \text{number of false positives per μL} \\ s & \text{object-level sensitivity} \end{array} \quad (3)$$

The following relations hold:

$$t = p \cdot s, \quad (4)$$

$$q = t + f. \quad (5)$$

Substituting (4) into (5), and solving for p, we obtain:

$$p = (q - f)/s\,. \quad (6)$$

Thus, we can estimate parasitemia, p, if we can estimate f and s. We know the ground truth for the validation set, so we can estimate s on the positive samples in the validation set as follows: $\hat{s} = \text{mean}(s)$. We can estimate f on the negative validation set because every suspected parasite is a false positive object. The estimate of f is the threshold $\hat{f}$ on the number of suspected parasites/μL. Let us assume f is Gaussian-distributed at the patient level. If we set the threshold $\hat{f} = \text{mean}(f)$, half of negative patients will be diagnosed as positive. To get 95% patient-level specificity,

we must use a larger threshold:

$$\hat{f} = \text{mean}(f) + 1.65 \cdot \text{std}(f)\,. \qquad (7)$$

The mean and standard deviation by patient are taken over the negative validation set, and sensitivity variation is ignored for ease of calculation. Using this threshold $\hat{f}$, we will obtain 95% sensitivity for positive patients when the parasitemia is greater than the following LoD:

$$\text{LoD} = 3.3 \cdot \text{std}(f)/\hat{s}\,. \qquad (8)$$

The numerator in Equation 8 captures the algorithm's variance in FP rate by patient, while the denominator accounts for the inefficiency of parasite detection. For example, if sensitivity is 50%, and $\text{std}(f)$ = 80 p/µL, then a clean slide with many fewer false positives than usual must contain > 264 p/µL to get a positive-object count that exceeds the threshold. This implies that the critical FoMs for estimating LoD are $\text{mean}(s)$ and $\text{std}(f)$ by patient.

3. Results

In this section, we report results in two ways, patient-level and object-level. First, we give patient-level diagnosis accuracy on various low parasitemia holdout sets (Figure 6), and quantitation results on holdout sets with a range of parasitemias (Figure 7). Second, we present the object-level metrics that support the patient-level results. We also provide a table that compares our algorithm with various others in the literature. Key metrics include number of patients, and estimated LoD.

3.1. Patient-level results

Our algorithm was trained on a set of 78 positive and 31 negative patients. Hyperparameters for diagnosis and quantitation (such as $\text{mean}(f)$, $\text{std}(f)$, and $\text{mean}(s)$) were calculated on a validation set of 54 positive and 32 negative patients. Each sample consisted of 324 FoVs (~0.1 µL of blood). Target patient-level specificity was set to 95% on the negative validation set.

3.1.1. Diagnosis The algorithm was applied to four holdout sets. Each set contained 20 negative and 10-12 low-parasitemia *P. falciparum* positive slides. Two of the sets were *P. falciparum* diagnosis portions of official WHO55 evaluation sets. The other two sets were ersatz WHO-type sets from different malaria-endemic regions. The low parasitemia samples are important to clinical use-cases. Figure 6 shows diagnosis results by patient and parasitemia. Specificity on the negative slides in each of the holdout sets was ≥ 90%. We present diagnoses by parasitemia to infer empirical LoD because patient-level sensitivity is a meaningful metric only in association with specificity and parasitemia. These results indicate an effective LoD ~100 p/µL at 90% specificity.

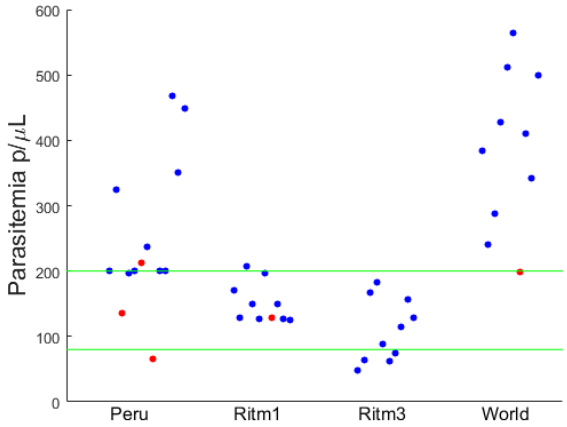

Figure 6. Diagnosis by parasitemia for 4 holdout sets. Each dot represents a positive patient. Blue dot: correct diagnosis, red dot: false negative. Green lines indicates parasitemia range used for WHO evaluation (90% sensitivity @ 90% specificity = competency level 1).

3.1.2. Quantitation Accurate parasite quantitation is important for case management—parasite density can indicate the severity of the infection [2]—and for generating accurate parasite clearance curves [7] in antimalarial efficacy studies [8]. To assess quantitation accuracy, the algorithm was applied to a holdout set of 45 positive *P. falciparum* patients from various regions of the world. Results are shown in Figure 7. The ± 25% error lines represent a range that allows the calculation of the log slope of clearance curves with error ≲10% for antimalarial efficacy studies. The results indicate that quantitation is sufficiently accurate for parasitemia > 1000 p/µL, but that estimates are high for parasitemia < 1000 p/µL.

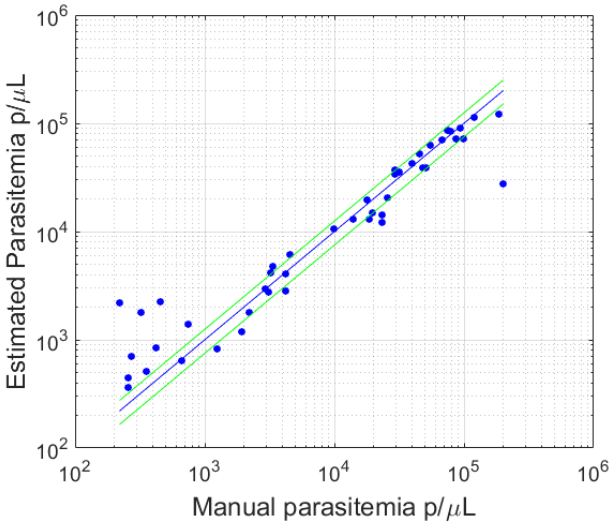

Figure 7. Parasitemia quantitation performance on holdout sets. Green lines indicate ± 25% range.

3.2. Object-level results

We now describe object-level metrics supporting the patient-level results described in Section 3.1.1. Table 2 shows the confusion matrix for the classification of candidate objects from the validation set via the CNN classifier. The confusion matrix was generated by setting the classifier threshold to 0.6. At this setting, the following object-level performance metrics obtain: sensitivity 91.6%, specificity 94.1%, precision 89.7%.

	Positive	Negative
Parasite objects	33,438	3,064
Distractor objects	3,849	61,405

Table 2. Confusion matrix for CNN on validation set.

We compare our results with others in the literature in Table 3. Although the most relevant comparison between algorithms is at the patient level, due to the general lack of patient-level information in the publications, we primarily compare object-level results. We also predict patient-level results based on object-level metrics: the final column in Table 3 is a projected LoD at the patient level (estimated via Equation 8 at 95% specificity; when $\mathrm{std}(f)$ is not available, then $\mathrm{mean}(f)$ is a useful indicator of the variance in FP rate). Note that several of the methods were proposed as decision-support rather than standalone systems, and for these methods lower specificity may be tolerable.

4. Conclusions

This paper describes a CNN-based malaria detection algorithm, the first (to our knowledge) that applies CNN models with sufficient training and validation data and patient-level accuracy to meet two key use-cases of the automated malaria problem: clinical diagnosis down to 100 p/μL; and *P. falciparum* quantitation for drug-resistance studies. The system reads thick film blood slides prepared with Giemsa stain according to current field norms, which is a minimum requirement for the above use-cases. Multiple field evaluations to further test the system are currently underway.

Our internal tests indicate that the system has thus far achieved malaria diagnosis accuracy sufficient to attain competence level 1 in the WHO external competency assessment of malaria microscopists for *P. falciparum*, which means that it performs on a par with well-trained microscopists for this species. It is still the case that highly-trained microscopists can out-perform automated systems. While the algorithm shows robustness to wide variation in field-prepared samples, it can fail when confronted with novel slide preparations or artifacts to which it was not exposed. This tendency can be ameliorated as newly encountered material is classified and added to its library via updates.

Our system can also be used for computer-assisted malaria diagnosis, since it outputs an array of thumbnails of the most suspicious (i.e. highest scoring) objects. In this mode, the machine reduces the workload of the user by pre-scanning the slide and presenting the most relevant objects for review. In initial field usage, this mode of operation may allow time for stakeholders to gain confidence in the system's capabilities and robustness. Regardless of usage mode, the thumbnails are always available for confirmation and review in case of unusual findings.

The new computer vision methods we have introduced are relevant to applications in automated medical diagnosis via microscopy, sonography, and radiology, as well as problems dealing with rare-object detection. These applications are important areas of computer vision research.

5. Acknowledgements

The authors gratefully acknowledge the Bill and Melinda Gates Foundation Trust for their sponsorship through Intellectual Ventures' Global Good Fund.

Algorithm	Film type	# Patients training set	μL blood/ patient	mean(f) FP/μL	std(f) FP/μL estimated	mean(s) % estimated	LoD p/μL estimated
WHO level 1 microscopist	thick		0.03-0.07				100
Quinn *et al*. [15]	thick	133 †	0.06	1200 ‡	240	20 ‡	4800
Rosado *et al*. [16]	thick	6	0.04	6470	1294	78	6640
Vink *et al*. [17]	thin AO §	> 22	0.47	< 7	< 2	75	30 ∂
Linder *et al*. [18]	thin	44	0.05	5000	1000	85	9400
Díaz *et al*. [19]	thin	5	0.005	15000	3000	94	12800
Delahunt *et al*. [20]	thick	93	0.1	93	70	20	267
Ours	thick	195	0.1	12	12 *	43 *	112

† Train and validation sets not separated by patient.

‡ Assumes 90% precision, 20% recall per authors' suggestion.

§ Uses non-standard Acridine-Orange staining cartridge and fluorescence microscope.

∂ Results from field trial with 70 positive, 16 negative patients, 84% patient specificity.

* Actual (not estimated).

Table 3. Metrics of manual and automated malaria diagnosis algorithms.

References

[1] World Health Organization. World Malaria Report 2016.

[2] World Health Organization. Basic malaria microscopy – Part I: Learner's guide. Second edition. February 2010.

[3] C. Wongsrichanalai, M.J. Barcus, S. Muth, A. Sutamihardja, and W.H. Wernsdorfer. A review of malaria diagnostic tools: microscopy and rapid diagnostic test (RDT). *The American Journal of Tropical Medicine and Hygiene*, 77(6 Supplement):119-127, 2007.

[4] H. Albert, Y. Manabe, G. Lukyamuzi, P. Ademun, S. Mukkada, B. Nyesiga, M. Joloba, C. N. Paramasivan, M.D. Perkins. Performance of three LED-based fluorescence microscopy systems for detection of tuberculosis in Uganda. *PLOS ONE* 5(12): e15206, 2010.

[5] D.N. Durrheim, P.J. Becker, K. Billinghurst. Diagnostic disagreement–the lessons learnt from malaria diagnosis in Mpumalanga. *South African Medical Journal*, 87:1016, 1997.

[6] World Health Organization. Microscopy for the detection, identification and quantification of malaria parasites on stained thick and thin blood films in research settings, 2015.

[7] N. White. The parasite clearance curve. *Malaria Journal* 10:1–8, 2011.

[8] World Health Organization. Methods for surveillance of antimalarial drug efficacy, 2009.

[9] S. Ashraf, A. Kao, C. Hugo, E.M. Christophel, B. Fatunmbi, J. Luchavez, K. Lilley, and D. Bell. Developing standards for malaria microscopy: external competency assessment for malaria microscopists in the Asia-Pacific, *Malaria Journal*, 11:352, 2012.

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation* 4:541-551, 1989.

[11] A. Krizhevsky. I. Sutskever, G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25:1097-1105, 2012.

[12] D.C. Ciresan, A. Giusti, L.M. Gambardella, J. Schmidhuber. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 411-418. Springer Berlin Heidelberg, 2013.

[13] K. He X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[14] D.C. Warhurst, J.E. Williams. Laboratory diagnosis of malaria. *Journal of Clinical Pathology* 49:533-538, 1996.

[15] J.A. Quinn, A. Andama, I. Munabi, F.N. Kiwanuka. Automated Blood Smear Analysis for Mobile Malaria Diagnosis. In Mobile Point-of-Care Monitors and Diagnostic Device Design, Chapter: Automated Blood Smear Analysis for Mobile Malaria Diagnosis, Editor: W. Karle, pp. 115–132, 2014.

[16] L. Rosado, J.M. Correia da Costa, D. Elias, J.S. Cardoso. Automated detection of malaria parasites on thick blood smears via mobile devices. *Procedia Computer Science*, 90:138–144, 2016.

[17] J.P. Vink, M. Laubscher, R. Vlutters, K. Silamut, R. J. Maude, M. U. Hasan, and G. Haan. An automatic vision-based malaria diagnosis system. *Journal of Microscopy*, 250:3:166-178, 2013.

[18] N. Linder, R. Turkki, M. Walliander, A. Mårtensson, V. Diwan, E. Rahtu, M. Pietikäinen, M. Lundin, J. Lundin. A Malaria Diagnostic Tool Based on Computer Vision Screening and Visualization of *Plasmodium falciparum* Candidate Areas in Digitized Blood Smears, *PLoS One* 9, e104855, 2014.

[19] G. Díaz, F.A. González, E. Romero. A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images. *Journal of Biomedical Informatics*, 42:296–307, 2009.

[20] C.B. Delahunt, C. Mehanian, L. Hu, S. K. McGuire, C.R. Champlin, M.P. Horning, B.K. Wilson, C.T. Thompson. Automated Microscopy and Machine Learning for Expert-Level Malaria Field Diagnosis. *2015 IEEE Global Humanitarian Technology Conference (GHTC)*. Seattle, WA, 2015, pp. 393-399.

[21] L. Rosado, J.M. Correia da Costa, D. Elias, J.S. Cardoso. A review of automatic malaria parasites detection and segmentation in microscopic images. *Anti-Infective Agents,* 14(1):11-22, 2016.

[22] D.K. Das, R. Mukherjee, C. Chakraborty. Computational microscopic imaging for malaria parasite detection: a systematic review. *Journal of Microscopy*, 260(1):1-19, 2015.

[23] S. Strugger. Fluorescence microscope examination of bacteria in soil. *Canadian Journal of Research*, 26c(2): pp. 188-193, 1948.

[24] World Health Organization. Malaria microscopy quality assurance manual, 2016.

[25] R.B. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580-587, 2014.

[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *The IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp. 779-788, 2016.

[27] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Raman. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9):1627-1645, 2010.

[28] J.R.R. Uijlings, K.E.A. Van De Sande, T. Gevers, A.W.M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154-171, 2013.

[29] R.B. Girshick. Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), pp. 1440-1448, 2015.

[30] S. Ren, K. He, R.B. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*, pp. 91-99, 2015.

[31] International Telecommunication Union. Recommendation ITU-R BT601-7, 2015.

[32] B.E. Boser, I.M. Guyon, V. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, ACM, Pittsburgh, 1992.

[33] A.Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proc. ICML*, 2004.
[34] R. Tibshirani. Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–88, 1996.
[35] R. Rosipal, N. Kramer. Overview and Recent Advances in Partial Least Squares. In *Subspace, Latent Structure and Feature Selection Techniques*, pp. 34–51, 2006.
[36] C. E. Metz. Evaluation of digital mammography by ROC analysis. In *Proc. International Workshop on Digital Mammography*, 61-68, 1996.
[37] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.
[38] J. Kittler, J. Illingworth. Minimum error thresholding, *Pattern Recognition*, 19(1):41-47, 1986.
[39] J.F. Brenner, B.S. Dew, J.B. Horton, T. King, P.W. Neurath, W.D. Selles. An automated microscope for cytologic research: a preliminary evaluation. *Journal of Histochemistry and Cytochemistry*, 24(1): pp.100-111, 1976.
[40] L. Breiman. Random Forests. Machine Learning, 45(1):5–32, 2001.
[41] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding. *arXiv*:1408.5093v1 [cs.CV], 2014.
[42] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV], 2015.
[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, 2015.
[44] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. *ImageNet Large Scale Visual Recognition Competition (ILSVRC)*, 2012.
[45] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. arXiv:1403.6382v3 [cs.CV], 2014.
[46] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. arXiv:1310.1531v1 [cs.CV], 2013.
[47] J. Yosinski, J. Clune, Y. Bengio, H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
[48] D.R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society B*, 20:215–242, 1958.
[49] R.E. Fan; K.W. Chang; C.J. Hsieh; X.R. Wang; C.J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.