Image Classification of Unlabeled Malaria Parasites in Red Blood Cells

Zheng Zhang[1], L.L. Sharon Ong[2], Kong Fang[2], Athul Mathew[1], Justin Dauwels[1], Ming Dao[2,3], Harry Asada[2,3]

***Abstract*— This paper presents a method to detect unlabeled malaria parasites in red blood cells.The current "gold standard" for malaria diagnosis is microscopic examination of thick blood smear, a time consuming process requiring extensive training. Our goal is to develop an automate process to identify malaria infected red blood cells. Major issues in automated analysis of microscopy images of unstained blood smears include overlapping cells and oddly shaped cells. Our approach creates robust templates to detect infected and uninfected red cells. Histogram of Oriented Gradients (HOGs) features are extracted from templates and used to train a classifier offline. Next, the Viola-Jones object detection framework is applied to detect infected and uninfected red cells and the image background. Results show our approach out-performs classification approaches with PCA features by 50% and cell detection algorithms applying Hough transforms by 24%.**

Majority of related work are designed to automatically detect stained parasites in blood smears. Although it is more challenging to design algorithms for unstained parasites, our methods will allow analysis of parasite progression under different drug treatments.

I. INTRODUCTION

Malaria is the one of the serious infectious diseases in tropical regions with about 3.2 billion people at risk. according to the World Malaria Report, about 438 thousand deaths were caused by malaria and there were more than 200 million new cases of malaria in 2015 [1]. The definitive diagnosis of malaria infection is done by searching for parasite in blood slides through a microscope. Although newer techniques have been introduced [2], blood smear examination utilizing manual microscopy still remains "the gold standard" when it comes to malaria diagnosis [3]. Diagnosis applying a microscope requires special training, experience and considerable expertise [4]. Several studies have shown that manual microscopy is not a reliable screening method when performed by non-experts especially in the rural areas where malaria is endemic [5].

We present an approach to automatically detect malaria parasites in unstained blood droplets. Majority of automated image analysis algorithms are designed to detect parasites in stained cells [7]. Related image processing algorithms for the automated detection of malaria cells are applied on stained cells. For example, a Giemsa stain can be utilized to stain to cell sample [6]. The malaria infected cell absorbs the stain and typical malaria detection algorithms focus on isolating such stained cells based on intensity of the infected cells.It is more challenging to detect infected cells without any staining due to the lack of contrast between infected and uninfected cells. Furthermore, cells may overlap and there are oddly shaped uninfected cells due to the blood smearing processing which could be detected as infected ones.

This paper applies a machine learning approach to create and match classifiers of infected and uninfected cells to acquired microscopy images. The key to an accurate classifier is the choice of features to represent the cells. When a red blood cells is infected with the parasite, its appearance changes over time.

We build classifiers utilizing Histogram of oriented gradients (HOGs), a technique which calculate the histogram of gradient directions of localized image regions. By dividing an image into small regions and combining calculated histogram of gradient directions of each region, the shape and appearance of the object can be represented. With these classifiers, we apply the Viola Jones object detection framework to detect infected and uninfected cells. We compare our technique to cell detection algorithms utilizing Hough transform. Furthermore, we compare the performance of HOG features against other feature representation such as PCA. Results show that our method outperforms both approaches.

This paper is organized as follows: Section 2 summarises literature related to malaria parasite detection in red blood cells. Section 3 presents our approach. Section 4 and 5 presents our results and conclusion.

II. LITERATURE REVIEW

Most known techniques developed to automatically detect malaria infected red blood cells were applied on images where the parasites were stained. In Purwar et. al [7], the images were pre-processed applying local histogram equalization. Segmentation methods applied include Chan-Vese segmentation method, morphological operations and Hough transform. The infected and uninfected cells were grouped utilizing a probabilistic k-means clustering algorithm. Ruberto et. al [8] applied morphologies methods and thresholding to detect parasites in Giemsa stained blood slides. The size of the red blood cells and the nuclei of parasites were evaluated utilizing granulometry.

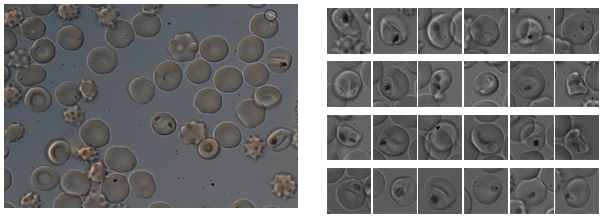
Ritter et. al [9], segmented cells applying thresholding and separated cells that touch and refined cell boundaries with Dijkstras shortest path algorithm. Diaz et. al [10] applied a color pixel classifier to label each pixel as foreground

[1]Z. Zhang, A. Mathew and J. Dauwels are with the School of Electrical and Electronic Engineering, Nangyang Technological University, Singapore. j.dauwels at ntu.edu.sg

[2]L.L.S.Ong, K. Fang, M. Dao and H.H. Asada are with the Singapore-MIT Alliance for Research and Technology, Singapore. sharon.ong, kongfang at smart.mit.edu

[3]M. Dao and H.H. Asada are also with MIT, Cambridge, MA, USA. asada at mit.edu

Preprint submitted to 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received March 14, 2016.



(a) An original image (b) Dataset of infected cells

(c) Dataset of uninfected cells (d) Dataset of non-cell parts

Fig. 1. Example of original images and datasets

or background, followed by template matching to separate clumped cells. Templates were constructed from parasite-stained images utilizing Expected Maximization and utilized to classify infection life stages of each cell.

In summary, the techniques reviewed to classify malaria cells are designed for images of stained parasites. Hence, the parasites can be distinctly detected. However, our approach is designed for images with unstained parasites. Although it is more challenging to detect cells with unstained parasites, our methods will allow further observations of parasite progression over time.

III. Methodology

We present a two-stage approach to detect infected cells. The first stage applies a Viola-Jones object detection framework with trained classifier to detect all red blood cells from a blood droplet image. The second stage classifies each segmented region to an infected or uninfected cell applying morphological features. Figure 1(a) shows a typical image of a blood smear. In order to train a classifier, we create datasets of three object categories; images of infected cells (Figure 1(b)),uninfected cells (Figure 1(c)) and regions without a complete cell (Figure 1(d)).

A. Feature Descriptor Selection For Classifier

In order to build our object detector, we compare the performance of different feature descriptors on our dataset. The flowchart to test different feature descriptors is shown in Figure 2. We compare the following feature descriptors; Histogram of Oriented Gradients (HOGs), Principal Component Analysis (PCA) and GIST [16].

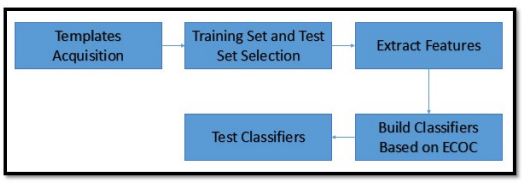
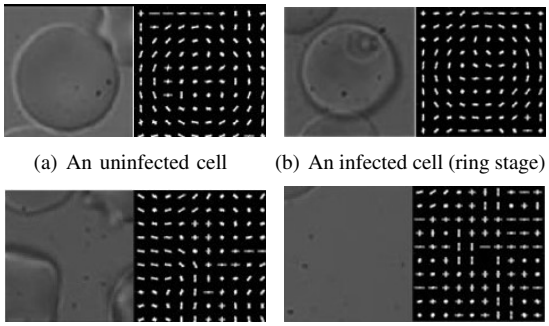


Fig. 2. Flowchart to test the performance of classifiers



(a) An uninfected cell (b) An infected cell (ring stage)

(c) A non-cell part (d) Another non-cell part

Fig. 3. Visualization of the HOG feature descriptors for different templates. Complete images of infected and uninfected cells (a) and (b) have similar visualizations.

To create a HOG descriptor, the image is divided into small regions, where a one-dimensional histogram of gradient directions or edge directions is calculated. These histograms are concatenated to represent the shape and appearance of the object [11]. HOGs can capture very characteristic edges or gradient structures of local shape. When small translations or rotations happen, the local representation has an controllable degree of invariance [11].

PCA converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables by orthogonal transformation. These are called principal components. The goal is to extract the important information from the whole data set and represent it. Mathematically, PCA depends on the eigenvalue of positive semi-definite matrices and the singular value decomposition of rectangular matrices [14]. Figure 4 shows the top two principal components for images with a whole cells (blue) and otherwise (red).

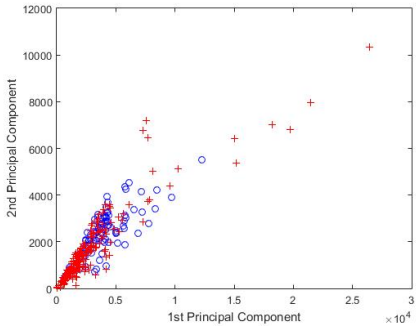


Fig. 4. The top two principal components for images with a whole cells (blue circles) and non whole cells(red crosses). From this visualization, it is difficult to distinguish the different classes.

GIST descriptors can provide a rough description of the scene structure. This feature representation contains statistics of oriented structures in the input image. It can be applied to predict the location, size and presence of the object in a scene [17].

We build a classifier applying multi-class support vector

Preprint submitted to 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received March 14, 2016.

machines (SVM) to determine the best feature descriptor to classify our three object categories. Probabilistic error-correcting output codes (ECOC) is utilized to solve the multi-class SVM [12]. In this approach, the posterior probabilities of the object categories are calculated in addition to the labels from the SVM. The normalized scores of probabilities are ranked and the best label is chosen as the object category.

B. Two stage approach to cell detection and classification

With the feature descriptors selected, we develop a two stage approach to detect infected and uninfected cells. The flowchart of the two stage approach applied to detect infected cells in the raw image is given in Figure 5

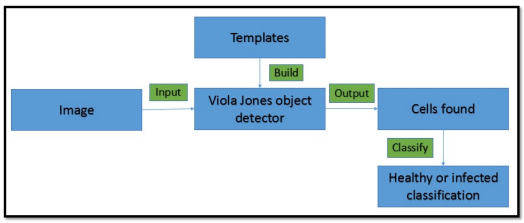


Fig. 5. Flowchart shows the two-stage approach: In the first stage, we detect cells in the raw image. The second stage classifies the cells found into healthy or infected.

In the first stage, we apply ViolaJones object detection framework to build an object detector for detecting cells [13]. This framework has three steps. In the first step, we extract features from the image applying a selected feature descriptor, described in Section IIIA. Next, Adaboost is applied to train a "strong" classifier by combining weighted "weak" classifiers. Finally, a cascade detector is built to detect cells in all sub-windows.

The steps to build a strong classifier with an Adaboost (Adaptive Boosting) algorithm are described in [13]. We assign the same weights all the instances (labelled images). Adaboost then trains a base classifier and increases the weights of the incorrectly classified instances. This algorithm is iterated multiple times. Each time, the algorithm applies a different classifier with the updated weights.

Cascading is a method to concatenate classifiers trained by Adaboost. The whole process of cascading is in the form of a degenerate decision tree [12]. In the first layer, the calculated number of positive samples are utilized and negative samples are generated. In the second layer, all the positive samples and negative samples generated in the first layer are classified. The wrongly classified positive samples are discarded. Positive samples which are classified correctly are kept as positive samples in next layer. The negative samples which are classified as positive are utilized as negative samples in next layer. This procedure is repeated in the following layers.

We compare the Viola Jones object detector for cell detection to Hough transform. The concept of Hough transform method is to extract the features of any shape present in the image [15]. The red blood cells are circular in shape and can be found utilizing a circular Hough transform.

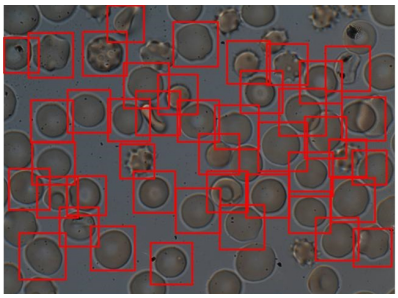


Fig. 6. Example of results given by cascade object detector: Red rectangles represent cells found in the raw image. Cells on the fringes are ignored

In the second stage, we perform additional operations on the cell regions found in the first stage. These crop image regions have different dimensions. Here, we classify the cells into infected and uninfected cells applying four approaches, which are:

1) An additional Viola-Jones object detector is applied as the cell size varies. Here, we utilize samples of infected cells as positive templates and uninfected cells as negative templates to build the detector.
2) In our second approach, we combine the object detector in 1) with circular Hough transforms. In the ring stage of malaria infection, we can see a small circular ring inside the cell. As we know the diameter range of this ring, we can find cells in this stage infection with circular Hough transform.
3) Our third approach combines a cascade objector detector and Hough transform in 2) with intensity thresholding. At the later stages of infection, the parasites appear at dark small regions inside the cell. Hence, we apply intensity thresholding to find dark regions, followed by size and aspect ratio to detect infected cells at this stage.
4) In our final approach, we combine the object detector in 1) and Hough transform in 2) with intensity thresholding of the red channel. From our images, parasites at the later stage appear redder than other regions. Hence, we can apply color thresholding to find regions with a higher red intensity to detect infected cells.

IV. Results

A. Comparing feature descriptors

To train classifiers, templates of cells and background were extracted from phase contrast microscopy images of blood. We acquired 62 templates of infected cells, 200 templates of uninfected cells and 300 templates of non-cell parts. From the total of 562 templates, we selected 300 of them as training set and the remainder is the test set. We cross-validated our results to ensure there are not biases by running our classifier 10 times, randomly selecting the test and training sets. We compare the senstivity and specifivity of the results. We define $Sentivity = True\ Positives/Total\ Positives$ and $Specificity = True\ Negatives/Total\ Negatives$ Table I shows classification accuracy with PCA, HOGs and GIST feature descriptors. We found that HOGs and GIST descriptors outperform PCA. Classifying with GIST feature

Preprint submitted to 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received March 14, 2016.

TABLE I

A COMPARISON OF DIFFERENT DESCRIPTORS TO CLASSIFY IMAGES WITH A WHOLE CELL AND WITHOUT

	Sensitivity	Specificity
PCA	0.4722	0.6232
HOG	0.9912	0.9967
GIST	0.9823	1.0000

TABLE II

RESULTS FROM TRAINING THE CLASSIFIER ON HOG FEATURES

	Classified as infected cell	Classified as uninfected cell	Classified as non-cell parts
Infected cell	0.4501	0.5242	0.0257
Uninfected cell	0.0943	0.9001	0.0051
Non-cell parts	0.0036	0.0083	0.9881

TABLE III

CASCADE OBJECT DETECTOR OUTPERFORMS HOUGH TRANSFORMS FOR CELL DETECTION

	Ratio of Cells found	Ratio of Cells missed
Cascade Object Detector	0.9331	0.0669
Hough Transform	0.6907	0.3093

TABLE IV

COMPARISON OF DIFFERENT APPROACHES TO DETECT INFECTED CELLS FROM EXTRACTED RED BLOOD CELL CANDIDATES

	Cascade Detector(COD)	COD + Hough Transform(HT)	COD+HT +Intensity Thresholding	COD+HT +Colour Thresholding
Sensitivity	0.4286	0.5714	0.8214	0.75
Specificity	0.9401	0.9359	0.8607	0.9109

descriptors results in slightly less true positives in comparison with HOGs. In addition, applying GIST is more computationally expensive. Therefore, we select HOG as the best feature descriptor for our application. If we classify infected and uninfected cells with non-cells with HOG descriptors in a single stage, a large proportion (52%) of infected cells are classified as infected cells (Table II).Therefore, we apply a two-stage approach to detected infected cells.

B. The two stage approach

We acquired 43 images of malaria culture blood droplets on microscopy slides. We select the first 24 to acquire templates to train classifiers. The remainder as applied as test images. Table III shows the cascade object detector with HOG features achieved 93% accuracy outperforming a circular Hough transform (69%). A circular Hough transform is designed to detect circular objects. This approach failed to detect some red blood cells which shriveled up in the sample preparation process. A second stage classifier was applied to detect infected cells from the red blood cells image regions extracted with the cascade objector. Table IV shows the results from four approaches applied. Results show that the best second stage detector combines a cascade objector detector, circular Hough transforms and intensity thresholding.

V. CONCLUSIONS

Infected red blood cells are automatically detected from microscopy images applying cascade detectors combined with Hough transform and intensity thresholding. A larger data set could be acquired to improve the performance of our classifiers. The performance of different feature descriptors can be explored. If this approach can be applied to build a independent software, it can be utilized to offer convenient diagnosis and expanded to diagnose other diseases.

REFERENCES

[1] WHO, World Malaria Report 2015, Key points, pp.X, World Health Organization 2015

[2] Hanscheid, T. : Current strategies to avoid misdiagnosis of malaria. Clin. Microbiol. Infect. 9, 497-504 2003.

[3] WHO, Basic Malaria Microscopy. Part I. Learners Guide, World Health Organization 1991

[4] Kettelhut, M.M., Chiodini, P.L., Edwards, H., Moody, A. : External quality assessment schemes raise standards: evidence from the UKNEQAS parasitology subschemes. J. Clin. Pathol. 56, 927-932 2003

[5] Bates, I., Bekoe, V., Asamoa-Adu, A. : Improving the accuracy of malaria-related laboratory tests in Ghana. Malar. J. 3, 38 2004

[6] Jager MM, Murk JL, Piqu RD, Hekker TA, Vandenbroucke-Grauls CM. Five-minute Giemsa stain for rapid detectionof malaria parasites in blood smears. Trop Doct. 2011 Jan;41(1):3335.

[7] Purwar, Y., Shah, S.L.,, Clarke, G., Almugairi, A., Muehlenbac, A. : Automated and unsupervised detection of malarial parasites in microscopic images. Malar. J. 10:364 2011

[8] Ruberto, C.D., Dempster, A.G., Khan, S., Jarra, B. : Automatic Thresholding of Infected Blood Images Using Granulometry and Regional Extrema, Proceedings of InternationalConference on Pattern Recognition, pp 3445-3448 2000

[9] Ritter, N., Cooper, J. : Segmentation and Border Identification of Cells in Images of Peripheral Blood Smear Slides, the Thirtieth Australasian Computer Science Conference, Conferences in Research and Practice in Information Technology(CRPIT), Australia, vol. 62 2007

[10] Diaz, G., Gonzalez, F.A., Romero, E. :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

[11] Dalal, N., Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

[12] Wang, Z., Xu, W., Hu, J., Guo, J. (2010, August). A Multiclass SVM Method via Probabilistic Error-Correcting Output Codes. In Internet Technology and Applications, 2010 International Conference on (pp. 1-4). IEEE.

[13] Viola, P., Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-511). IEEE.

[14] Abdi, H., Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459.

[15] J. Illingworth and J. Kittler, The Adaptive Hough Transform, PAMI-9 , Issue: 5, 1987, pp 690-698

[16] Oliva, A., Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3), 145-175.

[17] Torralba, A. (2003). Contextual priming for object detection. International journal of computer vision, 53(2), 169-191.

Preprint submitted to 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Received March 14, 2016.