Automated Status Identification of Microscopic Images Obtained from Malaria Thin Blood Smears using Bayes Decision: A study case in Plasmodium Falciparum

Dian Anggraini[1], Anto Satriyo Nugroho[1], Christian Pratama[2], Ismail Ekoprayitno Rozi[2], Vitria Pragesjvara[1], Made Gunawan[1]

[1]*Agency for the Assessment and Application of Technology, Jalan M.H. Thamrin No.8 Jakarta 10340 Indonesia*

[2]*Eijkman Institute for Molecular Biology, Jalan Diponegoro no. 69 Jakarta 10430 Indonesia*

Email: asnugroho@ieee.org

***Abstract*—Diagnosing malaria, as the first step to control the spread of the infectious disease, can be significantly optimized with a Computer Aided Diagnosis system. This study is proposed to develop a novel image processing algorithm to realiably detect the presence of malaria parasites from *Plasmodium falciparum* species in this smears of Giemsa stained peripheral blood sample. The proposed system was built using malaria samples that were specifically prepared by Eijkman Institute for Molecular Biology. Digital microphotographs were acquired using a digital camera connected to a light microscope. Global thresholding and connected component extraction were implemented to identify blood cell components. Two stage classification using separate set of features was built based on Bayes Decision Theory. Infected erythrocytes were identified with sensitivity of 92.59%, specificity of 99.65%, and PPV of 67.56%. The system provided an F_1 measure of 0.78.**

I. INTRODUCTION

Malaria is an infectious disease caused by unicellular protozoan parasite from the genus Plasmodium with 5 species known to infect human by entering bloodstream [1]. Those species are *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae*, *Plasmodium ovale*, and *Plasmodium knowlesi* [2]. It has been reported that malaria caused between 1.5 and 2.7 million deaths every year and, according to WHO, an amount of between US$ 5.0 and US$ 6.2 billion will be required per year for controlling and eliminating malaria globally during 2009 to 2015 periods [3].

In Indonesia, the highest prevalence of malaria is found in remote and forest related areas including Papua, Kalimantan, and Sulawesi [4] where number of medical experts and high quality medical facilities are rarely sufficient.

Furthermore, the key to overcome global health problem caused by malaria is through accurate diagnosis. Currently, there are several methods used to pronounce malaria including light microscopy, fluorescent microscopy, rapid antigen detection method, and polymerase chain reaction. Among these techniques, examining both thick and thin Giemsa stained blood smear under light microscope in order to find infecting parasites is considered to be the most sensitive and specific one [5].

Although the use of light microscope in diagnosing malaria offers many advantages, it also has some drawbacks. The condition of blood smear is highly influenced by time and storage. Moreover, confirming negative status of malaria take considerable time and the diagnosis relies heavily on the expertise and experiences of medical practitioner in the field. These disadvantages become a burden in Mass Blood Screening (MBS) and in controlling the spread of malaria in rural area, not to mention in eastern part of Indonesia. Therefore, an automated image analysis system would improve the performance of microscopy by circumventing its main limitation in term of dependency on the ability of medical practitioner to diagnose blood image accurately, thus providing a milestone for fast and accurate diagnosis of malaria in Indonesia remote area.

There have been several approaches to develop the algorithm of automated malaria diagnosis as addressed in [6]–[10]. These studies focused on either image processing technique to identify red blood cells and infecting parasites [6], morphological method for cell image segmentation and parasites detection [7], automated determination of parasitemia [8], or color segmentation for status identification [9]. Among these studies, some might be effective in cell segmentation, but less effective in selecting the features, while others achieved high specificity with prepared data sets but were not applicable in the field. In addition, only in [6]

was the algorithm built by mimicking the routine used by expert microscopists.

The main objective of this study is to propose a new algorithm for automated malaria status identification based on the standard routine used by medical practitioner performing microscopy diagnosis of malaria. The system is developed to segment the image into parts, i.e. to separate blood cells from the background [11]. The result of this stage is followed by classifying each red blood cell as infected or non-infected in two stage classification scheme by first identifying the components of the parasite. The algorithm is expected to provide a positive and negative diagnosis of malaria with comparable sensitivity and specificity to conventional microscopy. The classification stage becomes the focus of discussion in this paper.

This paper is organized in 5 sections. In Section 2, the materials and methods to obtain and standardize the input images are explained. Section 3 describes the theoretical and practical aspect of the proposed algorithm. Furthermore, the experimental results are presented and analyzed in Section 4. Finally, Section 5 provides the conclusions.

II. Materials and Methods

A. Malaria Samples

All images were generated from malaria blood slides that are used for teaching purpose prepared by experts from Eijkman Institute for Molecular Biology. For the time being, Plasmodium falciparum was chosen to be the source of parasites sample due to its prevalence in Indonesia as well as being the most dangerous and major contributor to deaths related with this disease [4]. Each slide was stained using standard Giemsa staining protocol [5] and subjected to manual microscopy examination by expert microscopist. Therefore, data on species specific diagnosis and parasitemia level is available.

Since in thin blood smear, there is no loss of parasite during staining; limited artefacts and overlapping cells; parasites are observable in their natural location and shape within the red blood cell since their morphology is conserved; and its extensive used in previous studies, as in [8], [9], [10], thin blood smears were used in this study.

B. Image Acquisition

As for generating the images, slides were examined under light microscope with 10x100 times magnification. The slides were examined under oil immersion in order to adjust the refractory index. Images were captured using a 5-megapixel Nikon digital sight DS 5Mc specifically designed and built-in for the light microscope. However, in this study, the images were generated using only 1.2 megapixel resolution without optical and digital zoom. Images were saved in the JPEG format in 1280 x 960 pixels size. Datasets, specifically dedicated for this study, were then created. In total, 773 microphotographs were acquired from 8 training slides while 60 microphotographs were used in training set and 20 microphotographs were used in testing set.

C. Examining Blood Film forMalaria Parasites

Fig.1 summarizes method used by human operator for analyzing blood smear images and its corresponding image processing algorithm. The design of algorithm will focus on classification of red blood cell as infected or non-infected based on the presence of the parasite from the species Plasmodium falciparum.

In most cases, malaria parasites undergo either trophozoite or gametocyte stages when they are identified inside red blood cells [5]. Since each phase has specific and distinguishable features, identification of infecting parasites will rely heavily on recognizing their key elements.

The evaluation of thin blood smears starts with identification of blood cells including Red Blood Cell (RBC), White Blood Cell (WBC), and platelets (see Fig.2). It proceeds further with recognition of RBC or erythrocytes. Inside red blood cells, development of the parasites starts with trophozoite stage or generally known as ring stage (see Fig.3). With proper Giemsa staining, nucleus of the parasite can be detected as one dark dot (indicated by number1 in Fig.3a) inside the RBC while the ring part will be easily distinguished as ring-shaped structure with more intense color compared to its surrounding (indicated by number 2 in Fig.3a). Based on these characteristics, medical practitioners will try to identify suspected parasite using presence of dark spot inside RBC as an indicator. Absence of black dot points out that the erythrocyte is not infected while erythrocytes with black dots need to be analyzed further.

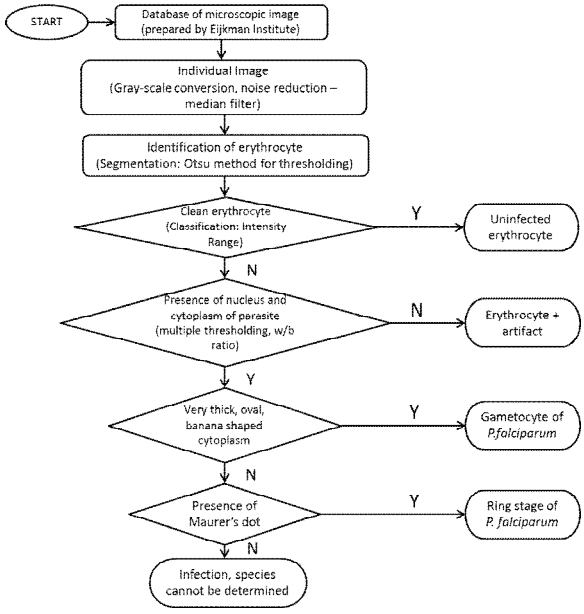


Fig.1 Standard procedure of parasite identification of *P.falciparum* using manual microscopy and its corresponding image processing algorithm

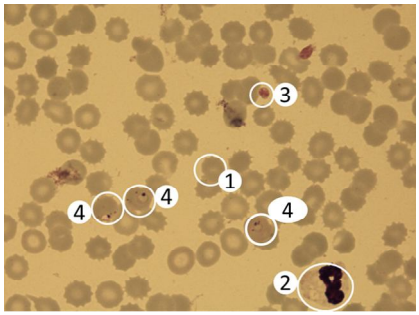
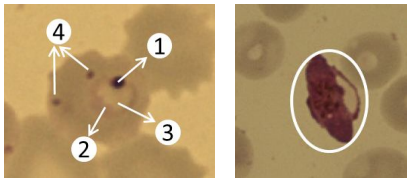


Fig. 2 A sample malaria blood image with blood cell components circled in white and numbered. 1: Red Blood Cell (RBC) 2:White Blood Cell (WBC) 3:platelets 4: Infected RBC



(a) (b)

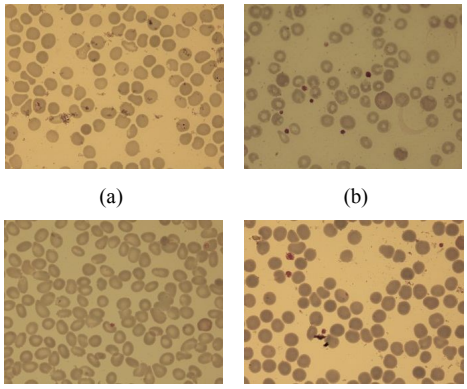
Fig.3 (a) Erythrocyte infected by ring stage parasite. 1: parasite's nucleus 2: parasite's cytoplasm 3: parasite's vacuole 4:Maurer's dots (b) Erythrocyte infected by gamete stage parasite

Furthermore, in order to be identified as infected erythrocyte, the dot must be accompanied by cytoplasm of the parasite that appears as lighter area surrounding the nucleus with ring-shaped structure around it as clearly indicated by number 3 in Fig.3a. The cytoplasm of the parasite can have wide variety of shapes, from a fine ring to an irregular or amoeboid shape [5]. The presence of both nucleus and cytoplasm, as the main constituent of living cell, is the most important prerequisite in validating the infected or non-infected status of a red blood cell. When this requirement is not met, then the black dot is simply an artefact.

There is a possibility that the cytoplasm has very dark or intense color and resembles a banana shaped structure as clearly illustrated in Fig.3b. These two features characterize Plasmodium falciparum in gametocyte stage. If this is not the case, presence of Maurer's dot (see Fig. 3a numbered with 4) will be the main attribute to confirm that the parasite is *P.falciparum* in ring stage.

III. PROPOSED SYSTEM

Substantially, this paper proposes an image processing algorithm to embrace the mechanism discussed in previous section up to the process of classifying each red blood cell as infected or non-infected based on the extracted components of the parasites. The scheme is initiated by implementing pre-processing stages that includes removal of unwanted noise. Blood cells, parasites, and other foreground objects were separated from the background by applying several global thresholding methods and visually comparing the results to qualitatively determine which technique yields the best end. The binary image is then subjected to hole filling process



(a) (b)

(c) (d)

Fig.4 Color variability in acquired images

and applied as marker to label blood cells. Each identified blood cell is fed into the classification model that is built as a two stage classifier using separate sets of features and Bayes Decision Rule [11] in each stage.

A. Pre-processing

The pre-processing stage is aimed to remove unwanted effects from the image and to adjust the image as necessary for further processing [13]. This stage consists of gray-scale conversion and 5x5 median filter. Fig.2 shows a typical image used in this study and may contain several foreground components including RBCs, WBCs, platelets, artefacts, and the parasite itself, which are numbered in ascending order. Since the input images demonstrate some degree of color variability as demonstrated in Fig. 4a–4d, the color image was converted to gray-scale image in order to reduce complexity without compromising the importance of parasite details.

Gray-scale conversion is followed by 5x5 median filter that substitutes the value of every pixel in the image by the median value of the intensity level in the neighborhood that will result in reduced sharp transitions, which is the indication of random noise, in intensities [13]. 5x5 median filter was observed to perform well in removing unwanted noise without removing parasite's details.

B. Segmentation of Blood Cell Components

The proposed algorithm is pursued with image segmentation to separate foreground objects i.e. erythrocyte, other blood cell components, and potential parasites from the background. Unlike reference [14] and [15] that used edge detection and watershed algorithm for blood cell segmentation, the proposed algorithm relies heavily on thresholding. Pre-processed image was converted into a binary image by selecting a threshold value that maximally divides the image into two classes of intensity: C_1 that contains intensity of background and C_2 that contains intensity of foreground objects.

It was observed that in most pre-processed images the intensity distribution between background and object pixels is not sufficiently distinct, and thus a single global threshold was not suitable. Consequently, the proposed system utilizes Otsu's method [16] that automatically selects threshold level that maximizes the between-class variances of the histogram. In principle, Otsu's method iteratively calculates the separability of the two classes, C_1 and C_2, for different value of threshold and selects the one that give the highest value of between-class variance.

Due to the biconcave nature of erythrocyte, some foreground objects might contain holes, i.e. part of foreground objects that have same intensity as background [12]. Therefore, hole filling process, designed based on extraction of connected component, was included in the proposed system. The idea of the algorithm is presented in Fig.5 and can be described as follow. Fig.6 shows a binary image of blood smear in which some erythrocyte contains hole, i.e. part of foreground object that has the same intensity as the background. Since the principle of CCL is finding foreground point [14], the complement of the image is first made before subjecting it to CCL. Following the CCL stage, the image is reconstructed by setting an intensity of 0 to the labeled objects otherwise the intensity was set to 255. Hole filling process is finalized by making the complement image of the reconstructed image. The result of hole filling process is subjected into another round of connected component extraction to obtain binary marker for individual cell. The whole process of blood cell segmentation is presented in Fig.7.

C. Features extraction

Two sets of feature were developed. The first set is based on the difference in intensity distribution between possibly infected erythrocytes and the healthy one as well as the area of the identified object. In gray-scale image, an infected erythrocyte will have some pixels with intensity value close to 0 due to the presence of parasite's nucleus. The similar condition applies to erythrocyte with artefacts but not with healthy one (see Fig.8). Using this concept, range of intensity or the numerical difference between the lowest and highest intensity value presented in each cell is proposed to be the first set of feature. Along with intensity range, area of segmented object is included in the first set of features to exclude small size objects like artefacts and platelets.

As mentioned before, there are two important components, parasite's nucleus and cytoplasm, that become the prerequisite for status identification of thin blood smear image. Using the knowledge that parasite's cytoplasm appear lighter while parasite's nucleus appear darker than the cytoplasm of the erythrocyte, it is possible to confirm the infected cell by first divide each possibly infected cell into 3 region by means of segmentation using multiple thresholding.

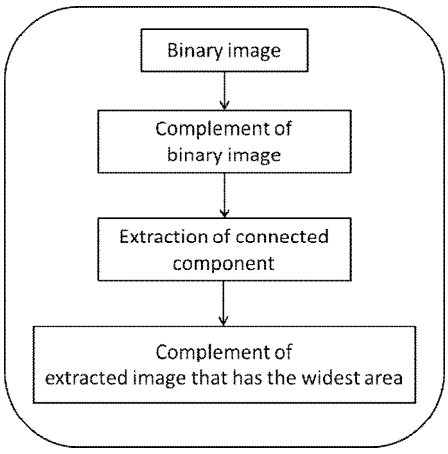


Fig.5 Hole filling process

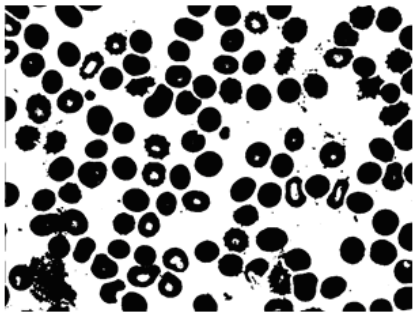


Fig.6 Binary image obtained with Otsu's method

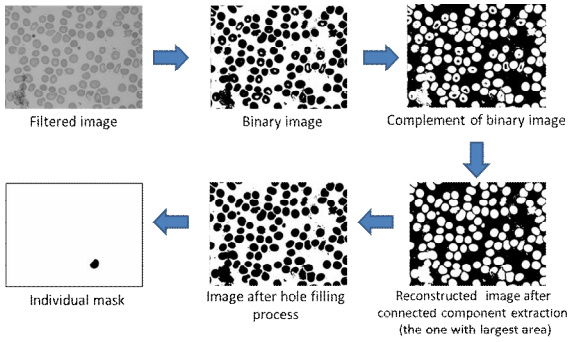


Fig.7 Segmentation process

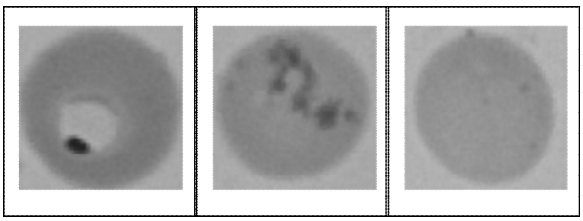


Fig.8 Comparison between (from left to right) infected RBC, RBC with artefact, and healthy RBC

It was evident that infected erythrocytes were successfully separated into 3 regions as shown in Fig.9. However, some erythrocytes with artefacts were also separated into 3 regions. Therefore, the algorithm was enhanced to obtained an area surrounding the possibly parasite's nucleus that has height that was two times the height of the nucleus and width that was two times the width of it. Within this area, the number of white and

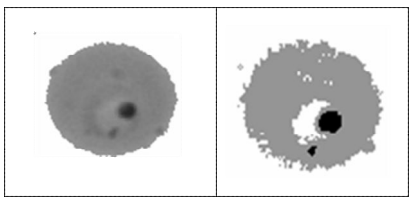


Fig.9 Infected erythrocyte and result after multiple threshold

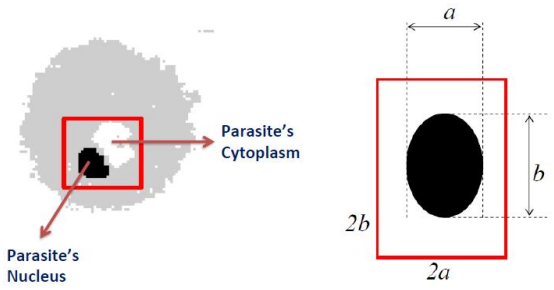


Fig.10 Obtaining white to black ratio

black pixels was calculated, thus the ratio between white and black pixels in each cell could be generated as shown in Fig.10. This ratio was proposed to be used in classifying positively infected erythrocyte.

D. Classification

The classification of an erythrocyte as infected or not fell into two-stage classifier, as demonstrated in Fig.11, with an erythrocyte classified as possibly infected by parasite at the first node and the confirmation of the positive status at the second node. Taking into account that, at each node, two features could actually give enough information for classification, Bayes classifier is implemented in each node.

The first classification stage can be considered as two-class case in which possibly infected RBCs were separated from normal one on the basis of range of intensity. This means that the segmented objects belong to either to class *N* or negative or to class *Pp* or possibly infected. For each cell, the range of intensity was measured and this feature was assigned as *r*. The probability density function [17] of the range of intensity measurement was calculated based on the determination of distribution of range of intensity for both classes using 60 training images.

A segmented blood cell component would be assigned to class *Pp* or possibly infected by parasite if and only if:

$$P(Pp|r) > P(N|r) \quad (1)$$

with $P(Pp|r)$ denotes the posterior probability that a cell belongs to possibly infected class based on range of intensity while $P(N|r)$ denotes the posterior probability that a cell belongs to non-infected class based on range of intensity.

In the second stage of classification, similar approach was utilized. The previously classified cells would be further identified as negative and belong to class *N* or positive and belong to class *P* based on the ratio of white area to black area alone.

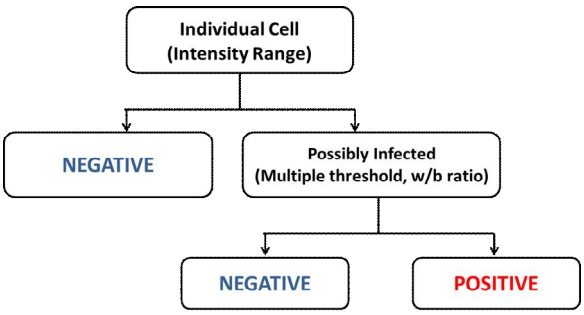


Fig.11 Diagram of final classification model

IV. RESULTS

The performance of this classification model was evaluated using 20 microphotographs obtained from different blood smears, to ensure the reproducibility of the proposed classifier. From the 20 thin blood smear images that contained 27 erythrocytes infected by ring-stage parasites, 25 infected RBCs were successfully identified, 2 infected erythrocytes were not identified (contributes to false negative), and 12

healthy erythrocytes or artefacts were identified as infected erythrocytes. Table 1 provides the detail of classification performance.

Four measures of algorithm performance and accuracy are used: sensitivity, specificity, positive predictive value (*PPV*) and F_1 score. These values are expressed in terms of true positives (*TP*), false positives (*FP*) and false negatives (*FN*):

$$Sensitivity = \frac{TP}{TP + FN} = \frac{25}{25 + 2} = 0.9259 \quad (2)$$

$$Specivicity = \frac{TN}{TN + FP} = \frac{3447}{3447 + 12} = 0.9965 \quad (3)$$

$$PPV = \frac{TP}{TP + FP} = \frac{25}{25 + 12} = 0.6756 \quad (4)$$

$$F_1 = \frac{2 \times PPV \times Sensitivity}{PPV + Sensitivity} = 0.7812 \quad (5)$$

The final classification model provided acceptable sensitivity. However, the improvement in sensitivity brought significant consequence regarding the low *PPV* and *F1* score. One of the underlying causes might be the second stage classifier that identified possibly infected erythrocytes that have low w/b ratio as positively infected erythrocytes. This condition leaves more rooms for improvement in future work.

All in all, the performance of the final classification model both for sample classification and individual infected cell identification are reasonable. However, it is important to recall that this evaluation was given for a limited number of samples. In some cases, the proposed algorithm failed to identify infected erythrocyte when the parasite was in the very early trophozoite stage since the cytoplasm of the parasite was not quite visible.

TABLE I
EXPERIMENTAL RESULTS USING PROPOSED METHOD

	Positive	Negative
Num. of cells	27	3459
Correctly classified cells	25	3447
Incorrectly classified cells	2	12

V. Conclusion

An automated image processing and classification method should be able to assist expert microscopists in detecting infected erythrocytes from malaria thin blood smears. This study provides a good basis for those who are aiming to investigate the automated blood film analysis because, unlike prior algorithm, the method proposed in this paper was developed based on the routine manual microscopy.

The main goal of this research was accomplished by solving two main problems by mean of image segmentation and developing a two stage classifier using Bayes Decision Rule in each stage. At first, the algorithm employed automatic threshold selection and successfully segmented blood cell components from background. It was further pursued by novel adaptation of CCL and connected component extraction to perform hole filling process. Two set of features based on image characteristics such as range of intensity and area of RBCs as well as features that mimic parasite‘s characteristics observed by human operator when diagnosing malaria were generated to first identify possibly infected erythrocytes and later used to confirm the infection.

The classification model identified infected erythrocytes with sensitivity of 92.59%, specificity of 99.65%, a *PPV* of 67.56%, and F_1 score of 0.7812. Several problems were identified when the proposed system was used to analyzed images that have significant amount of clumped erythrocytes. In addition, the w/b ratio feature failed to identify infected erythrocyte when the parasites were in the very early trophozoite stage in which the cytoplasm of parasite was hardly apparent.

As further work, it is planned to classify the phase of the parasite based on the shape and size of the parasites‘ cytoplasm as well as identification of Maurer‘s dot. Larger scale of experiment will also be necessary. Moreover, limitation of this study will be its focus on the species of *Plasmodium falciparum* and thus might not be optimally applicable to other species. Therefore, further work will also involve some improvements to accommodate the diagnosis of other species of parasites.

Acknowledgment

This research has been supported by the Ministry of Research and Technology, Government of Indonesia under Grant Insentif Riset RT 2011 – 2311. All supports are, therefore, hereby acknowledged. Distinctive gratitude is due to Eijkman Institute of Molecular Biology Indonesia for preparing the materials for this research and granting exceptional access to the datasets. Equally important, the authors received help from many and specially would like to deliver the best gratitude to Andree Ang Surya and Teresa Vania Tjahja from Swiss German University, Indonesia for the supports and scholarly inputs.

References

[1] Murray, R. Patrick, *Medical Microbiology*, 6th ed., USA: Mosby, 2008.

[2] W.E. Collins, and J.W. Barnwell, “*Plasmodium knowlesi*: Finally being recognized”, Journal of Infectious Disease, vol. 199, pp. 1107-1108, 2009.

[3] World Health Organization, *World Malaria Report*, Geneva: WHO, 2009

[4] World Health Organization, *Malaria in Southeast Asian Region*, WHO, 2010
[5] World Health Organization, *Basic Malaria Microscopy. Part1. Learner's Guide*, 2nd ed., Geneva: WHO, 2010
[6] N.E. Ross, et al, "Automated image processing method for the diagnosis and classification of malatia on thin blood smeras", Medical and Biological Engineering and Computing 44, pp. 427-436, 2006.
[7] C.D. Ruberto, et al, "Morphological image processing for evaluating malaria disease", IWVF4, pp. 739-748, 2001.
[8] S. Sio, et al, "Malaria Count: an image analysis-based program for accurate determination of parasitemia", Journal of Microbiological Methods, 2006. Available: http://www.sciencedirect.com
[9] G. Diaz, F. Gonzales, and E. Romero, "Infected cell identification in thin blood images based on color pixel classification: comparison and analysis", Springer Berlin, pp. 812-821, 2007.
[10] F.Tek, A.Dempster, and I.Kale, "Computer vision for microscopy diagnosis of malaria", Malaria Joournal 8
[11] D. Anggraini, A.S. Nugroho, C. Pratama, I. Ekoprayitno, A.A. Iskandar, R.N. Hartono: Automatic Status Identification of Microscopic Images Obtained from Malaria Thin Blood Smears, Proc. of 3rd International Conference on Electrical Engineering and Informatics (ICEEI 2011), CDROM A3-2, Institut Teknologi Bandung, Bandung, Indonesia, July 17-19, 2011
[12] Q. Wu Q., F. Merchant F., K. Castleman, Microscope Image Processing. 2008. New Jersey: Academic Press
[13] R.C. Gonzales, R.E. Woods,and S.L. Eddin, Digital Image Processing, 3rd ed., USA: Prentice Hall, 2008
[14] N. Ritter, and J. Copper, "Segmentation and border identification of cells in image of peripheral blood smear slides", Proceeding of Thirtieth Autralasian Computer Science Conference, 2007.
[15] F.Tek, A.Dempster, and I.Kale, "Blood cell segmentation using minimum area watershed and circle radon transformation", Mathematical Morphology: 40 Years On, pp. 441-454, 2005
[16] N. Otsu, "A threshold selection method from grey level histogram, *IEEE Transactions on Man, System, and Cybernetic,* vol.9, no.1 pp. 62-66, 1979
[17] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. New York: Willey-Interscience, 2001