REPORT

Unsupervised identification of malaria parasites using computer vision

Najeed Ahmed Khan*, Hassan Pervaz, Arsalan Latif and Ayesha Musharaff
Department of Computer Science & Software Engineering, NED University of Engineering & Technology, Karachi, Pakistan

Abstract: Malaria in human is a serious and fatal tropical disease. This disease results from Anopheles mosquitoes that are infected by Plasmodium species. The clinical diagnosis of malaria based on the history, symptoms and clinical findings must always be confirmed by laboratory diagnosis. Laboratory diagnosis of malaria involves identification of malaria parasite or its antigen / products in the blood of the patient. Manual diagnosis of malaria parasite by the pathologists has proven to become cumbersome. Therefore, there is a need of automatic, efficient and accurate identification of malaria parasite. In this paper, we proposed a computer vision based approach to identify the malaria parasite from light microscopy images. This research deals with the challenges involved in the automatic detection of malaria parasite tissues. Our proposed method is based on the pixel-based approach. We used K-means clustering (unsupervised approach) for the segmentation to identify malaria parasite tissues.

Keyword: Computer Vision, Malaria parasite detection, unsupervised identification.

INTRODUCTION

Malaria is widespread in tropical and sub-tropical regions (Lee *et al.,* 1993, Pherson *et al.,* 2007), where the humidity is high. According to WHO (World Health Organization) in 2015 it is estimated that 2140 million new cases of malaria were reported and 438000 people were dead due to improper cure and late detection of malaria parasite. (Joint WHO/UNICEF news release 2015). Almost half of the world's population is at risk of malaria. Early diagnosis of malaria by hematological parameters (Pagaro *et al.,* 2013) can prevent the damage caused by it. United Nations (UN) documented Millennium Development Goals (MDG) in September 2000, in which they targeted to early cure and prevention from a malaria parasite to be achieved by the year 2015. Rural and remote areas are predominant in malaria diseases where the early diagnoses and cure is rarely available (Suryakantha, 2010). Traditionally, pathologists have to examine the malaria parasite tissue under a microscope. With the advancements in digital pathology, glass tissue slides can be digitized to generate tissue images. However, the large volume of tissue images that are generated poses a challenge for pathologists to efficiently and accurately perform the diagnosis.

Automatic cell segmentation in pathological images is a challenging task due to the complex structure of the cell tissues in addition to the presences of noise in the images. Manual segmentation for this purpose involves much human intervention (supervision) and is tedious if the dataset consists of numerous images. Thus requiring unsupervised methods that able to perform automatic cell segmentation in an objective and efficient way is necessary. Many automatic segmentation methods have been proposed, Segmentation of white blood cells (WBC) into nucleus (Dorini, 2000), red blood cell classification (Soltanzadeh *et al.* 2010; Mahmood *et al.,* 2013; Aimi *et al*., 2013) extraction and segmentation of sputum cells for lung cancer (Sammouda *et al*., 1993; Taher *et al*., 2010; Taher *et al.,* 2013) however the degree of supervision is involved either in selected restricted region (Somasekar, 2011) or they used predefined size (Aimi *et al.,* 2013) of the cell tissues. We proposed an unsupervised approach to identify the malaria parasite from light microscopy images. We compute low level features and use K-means clustering for the segmentation to identify malaria parasite tissues.

**Corresponding author:* e-mail: najeed@neduet.edu.pk

MATERIALS AND METHODS

Sample preparation

The first step is to acquire the images of malaria samples, shown in fig. 1. In this study, the malaria images of the ring, trophozoite and falciparum gametocyte stages have been captured from the thick and thin blood smears (Dacie *et al.,* 2006) of P. vivax samples.

In a single ring schizonts (mature gametocyte stage) are usually few in numbers with 6-12 large merozoites (USAF-Public Health Information and Resources), which is not addressed in this study.

To generate tissue images the glass tissue slides are digitized using digital microscope. The microscope is one of the key components in the application of imaging for analyzing and viewing different slides. Digital microscope is connected to the computer and images are imported for the processing. The process of digitized the slides carried

out, followed by the chemical processing that includes: permeabilization, fixation, mounting and staining. For more detail, readers are advised to look (Sample preparation, Duke University, Trinity college Dublin).

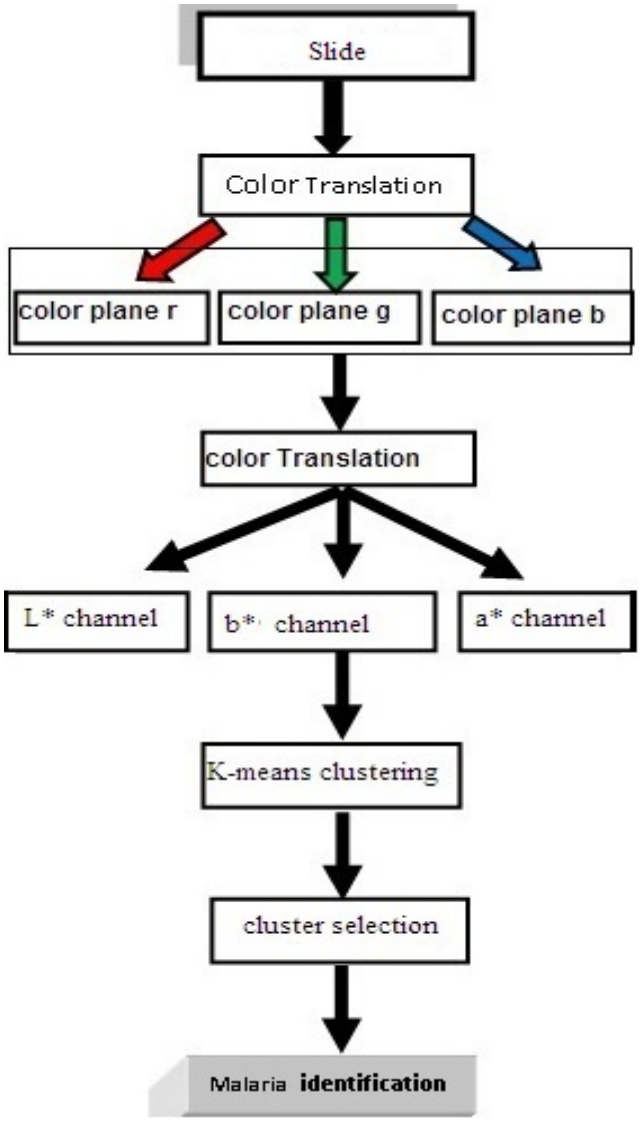

Fig. 1: Block diagram of proposed approach

Permeabilization

It includes treatment of cells with mild surfactant to dissolve cell membranes in order to dye larger molecule inside the cell.

Fixation is used to preserve cell or tissue morphology it involves complex chemical reaction, which creates bonds between proteins to increase their rigidity.

The Mounting process involves mounting the slide on the microscope for observation and analysis. Staining involves to stain cells, components. Slides are usually stained to enhance contrast for easy viewing under microscope. Staining is not limited to biological organism or materials, it has useful application and used in different applications such as to study structure of a material.

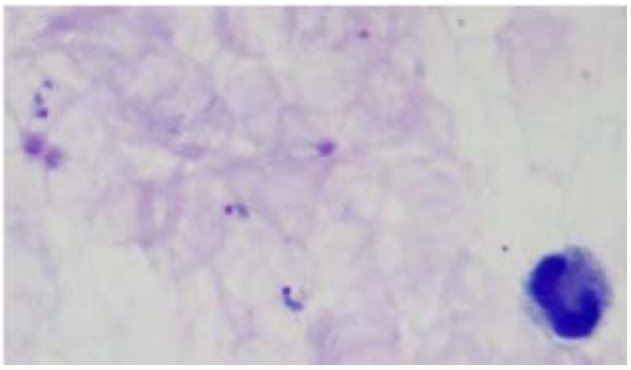

Fig. 2(a): Leishman-stained thick blood film showing late trophozoites of P. vivax

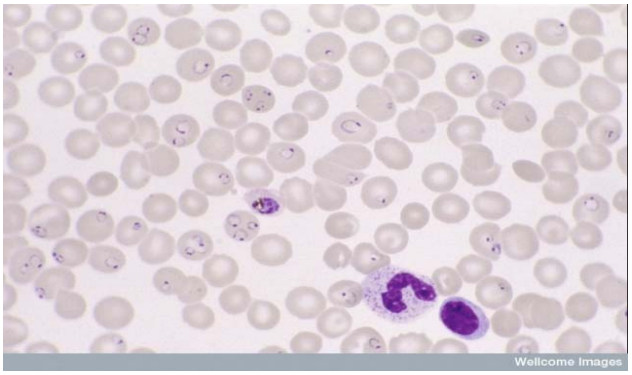

Fig. 2(b): Leishman-stained thin blood film showing P. falciparum gametocyte ring stage can be seen in brilliant violet color. (© Wellcome Images #. W0042199).

For detection of malaria we need high resolution of 100x or more, so we usually use oil emulsion on slides to view. The size of the slides we used for imaging purposes is approximately 1x3 inches in area and 1mm 1.2mm in thickness. For staining, we used Leishman staining (Taher *et al.,* 2013) for blood smears. figs. 2(a) and (b) show the sample of Leishman-stained thick and thin blood film, respectively. It is generally used in blood and bone marrow for identifying leukocytes, malaria parasite and trypanosomes, it provides good quality. Leishman staining has numerous advantages, it gives a good comparison and good sensitivity of malaria parasite and simpler than Giemsa-staining (Taher *et al.,* 2013) in use, therefore it is used extensively.

To assess the proposed segmentation method, around 118 malaria samples with various conditions have been captured. These slides are collected from a reliable laboratory of a local hospital. fig. 3 shows examples of a digitized form of the samples, these samples show that the captured malaria images with uniform distribution of RBCs.

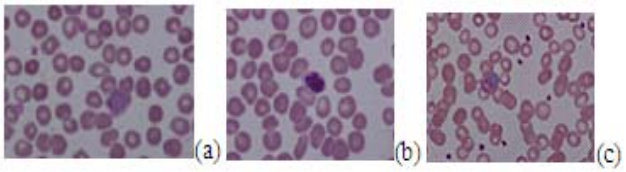

Fig. 3: Samples of malaria infected blood smear images captured at 100x resolution, (a) and (b) show ring stage - cells contain single and multiple chromatin dots, (c) shows the RBC - ring shape malaria parasite

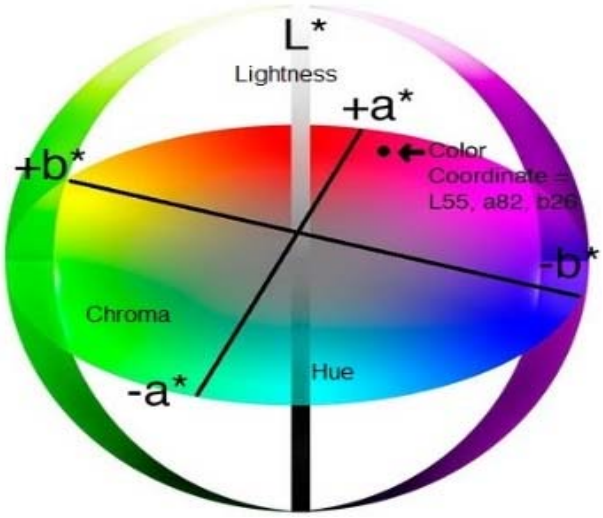

Fig. 4: Example of L*, a*, b* color channels in 3-D color space. [Gordon Pritchard]

The stacking RBCs show the samples of the captured malaria images with the presence of platelet (dark circle) and artifacts. Based on these malaria images, it can be seen that the color of the parasites and normal RBCs regions varies in each slide due to the non-standard preparation of the blood slides.

Feature extraction

Feature selection and computation is the quantitative measurements of images or sequences of images (video) typically used for identifying objects or region of interest and/or analyze the pathology of a structure or tissues in the pathology slides. Once the features have been computed, appropriate selection of a subgroup of the significant and robust features is necessary to improve the classification accuracy and minimizing the overall complexity.

We have used color feature from the malarial images. Perhaps a color is the most significant feature in an image. It is the comprehensively used as pixel level features in computer vision. A color image is a combination of some basic three colors. Each pixel in a color image can be broken down into Blue, Red and Green color channel values. We segment the entire image into a 3-dimentional vector, each one representing color R, G, B color feature respectively.

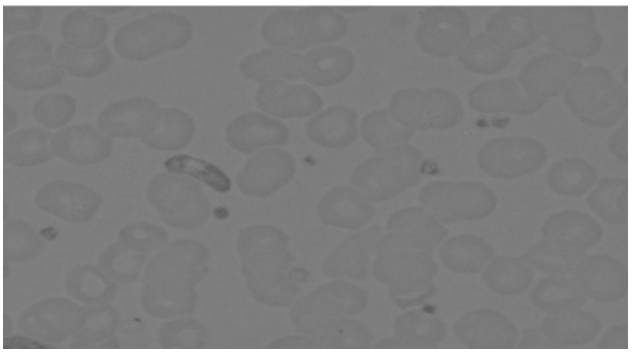

Fig. 5: Extracted b-channel from the image.

To use color as a feature, the selection of appropriate color space is important. A color image can be studied in color spaces such as RGB, HSV, CIE L*u*v*, CMYK and CIE L*a*b*. The, RGB color channels are perceptually inconsistent, as the two colors with greater distance could be visually more alike than any other two colors with smaller distance. In other words, the color distance in RGB channels does not represent perceptual color distance (Somasekar *et al.,* 2011). We choose perceptually consistent color space CIE L*a*b* for a malaria parasite identification. In an L*a*b* color space, a* and b* represent the color proportions for chrominance where L indicates dimension for luminance. The example of distribution of L*, a* and b* is shown in fig. 4. The values of a* are from Green to Red where the values of b*are from Blue to Yellow based on a non-linearly-compressed CIE XYZ color space. The nonlinear relationship for L*, a*, b* are intended to mimic the logarithmic response of the camera. The transformation from CIE XYZ to CIE Lab can be defined mathematically as follows:

$$C^t = MA^t \quad (1)$$

where C = [X,Y,Z] tri-stimulus values, A = [R,G,B] and M is a 3x3 matrix. The values of [R, G, B] □ [0-255] .

Since the color range of channel b* is from Blue to Yellow and is non-linearly compressed CIE -XYZ color space, from our initial experiments, it proved more prominent color channel to segment the malarial parasites (in blue color).

Let *I* represents a digitized image of the blood smear slide of size (m x n) pixels, considered by the color b* at location (i, j), i.e. $b = I(i, j)$. The set of pixels characterized by the color b* in the image *I* is defined as:

$$F = \{bk \mid k = 1,2....(m \times n)\} \quad (2)$$

Clustering to segment the malarial parasites

Clustering is the next step after the acquisition of training data to obtain malarial parasites, fig. 2, we cluster the matrix of feature vector *F*. We cluster training set into two clusters. The obtained set of clusters of input training set is represented as a partition of the pixels characterized by the color **b*** in the image *I*:

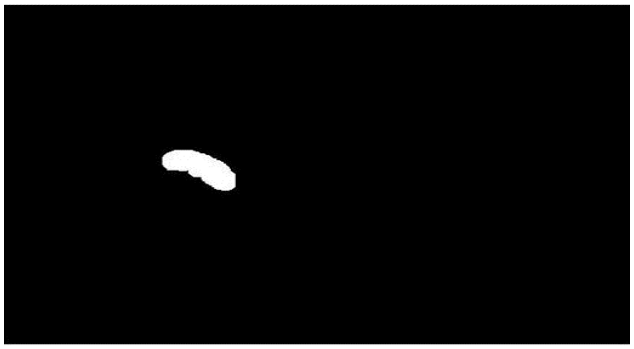

Fig. 6: Segmented malaria parasite

$$\theta = \{\theta 1, \theta 2, ...\theta N\} \text{ for } N < (m \times n) \quad (3)$$

Where, □ can be given as,

$$\theta = \{bi \,|\, bi \in f\} \quad (4)$$

The clustering feature vector *F* provides class labels for each pixel bk of the training set. Finally, we group the pixels of each class label to distinguish the malarial parasites from the other cells in the image.

MATERIALS AND METHODS

We have used 330 malarial parasite data (images), captured at 100x resolution after the Leishman staining processes. The size of each image is set 352 x 288 pixels. The noise in the captured images is not uncommon due to factors such as, digital artifacts, screen-door effects, distortion, silkscreen effects, chromatic aberration and color banding. These artifacts are corrected by applying different filters and proper maintenance of the microscope.

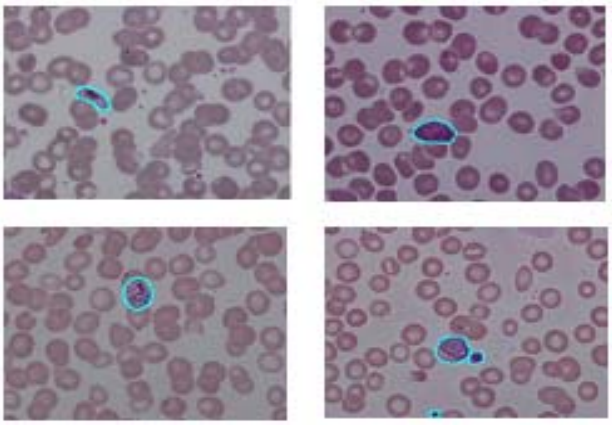

Fig. 7: Examples of identifying malarial parasites, bottom right image show two false positive.

from an example image of the dataset. In this fig. the malaria parasite is evident where the artifacts and other cell tissues are visible with low intensities in the output image. These artifacts and other the cell tissues merge with background part of the image due to low intensity shown in fig. 6.To organize the similar tissues in the segmented image region, standard K-means clustering is used to cluster the feature vector ***F*** defined in Equation (2). The purpose of K-means clustering is that the clusters of items with the same target category are identified. The predictions for new data items are made by assuming that they are of the same type and nearest to the cluster center. Finally we apply morphological operations on each cluster objects to identify them as malignant or non-malignant tissues. We cluster them into two clusters, the motivation of selecting two clusters is that, we interested to segment only malarial parasite tissues from the rest of the tissues in blood sample. Varying number of clusters approach (K. Najeed, DC Hogg, 2014) can also be used, if the utmost groups of blood tissues are required to be identified. A little supervision is involved with setting k=2. Due to the variability of random clustering produced by k-means, we run k-means multiple times and take the average of their output. A distance function (default function) is used with the k-means to separate the clusters. Finally to reconstruct the morphological shape (Z. Dongming, G.D. David, 1991) of the parasite in each cluster 8-pixel's neighbors connected component algorithm is used to represent the parasites in the original image.

RESULTS

A crude analysis of the segmented results reveals that the proposed approach of unsupervised identification of malaria parasite can clearly distinguish the malaria parasites from other segments.

Our proposed approach is simple and robust for the identification of malarial parasites from the stained slides blood smears.

The examples of qualitative results are shown in fig. 7. Fig. 7 shows identified malarial parasites. The quantitative results of the proposed approach are described by the "Precision" and "Recall". The Precision of a measurement system is the degree to which repeated measurements show the same results. "Recall" is the measures the proportion of actual positives in the data set, which are correctly identified. Sensitivity is the proportion of people that are known to have the disease who test positive for it. This is also written in Table 1, which shows the Precision and Recall results obtained through a dataset consisting of 118 images.

DISCUSSION

The pathological images of human blood smear are used to automatic segment malarial tissues. These the images of malaria samples acquired from the blood slides after digitizing using digital microscope at 100x resolution. From these digital images a consistent color space CIE L*a*b* is chosen for a malaria parasite identification. Color channel b* is non-linearly compressed proved more prominent color channel to segment the malarial parasites. Unsupervised K-means clustering and image processing

Table 1: Precision and Recall results, In table, column three represents True Positive (T.P), column four represents False Positive (F.P), column six represents True Negative (T.N) and column seven represents False Negative (F.N).

Dataset	No of samples	T.P	F.P	F.N	Precision $\frac{TP}{TP+FP}$	Recall $\frac{TP}{TP+FN}$
Human Blood Samples	118	82	4	36	0.95	0.69

techniques (8-neighbors connected component) are used to classify and identify the malarial parasites in the blood sample. The results (Fig. 7) show that the malaria parasites are evident in the images after applying proposed technique, where the artifacts and other cell tissues are visible with low intensities.

From the results, it is reveal that the malaria parasite can identify from the images of human blood slides using CIE L*a*b* color channel space.

CONCLUSION

The proposed framework in this paper incorporates low level features such as the color channels. The experimental results show that the feature set segmented b*-color channel from the CIE L*a*b* color space provides better clustering results. With this suitable color channel segmentation technique, malarial parasites can be identified appropriately in pathology slides' images. Our proposed framework for malaria parasite identification may not be suitable to identify other kinds of diseased parasite in pathology slide images, therefore, other low level features such as texture, color histogram and SIFT descriptors (D. G. Lowe, 1999) features may be combined with the single b*-color channel to improve the automatic identifying diseased system.

REFERENCES

Aimi S Yusoff M Zeehaida M (2013). Color Image Segmentation Approach for Detection of Malaria Parasites Using Various Color Models and k-Means Clustering, Wseas Transactions n Biology and Biomedicine., **10:** 41-55

Dacie SJ and Lewis SM (2006). Reference ranges and normal values, Practical hematology. 10th Ed, Philadelphia: UK, Churchill Livingstone Publication., 14-7.

Dongming Z and David GD (1991). Morphological hit or miss transformation for shape recognition. *Journal of Visual Communication and Image Representation*., **2:** 230-243.

Dorini LB Minetto R and Leite NJ (2007). White blood cell segmentation using morphological operators and scale-space analysis, *Proceedings of the 20th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*., *294–301*.

Introduction to sample preparation, Light microscopy Core Facility, Duke University and Duke University Medical Centre, access on February 2016, http://microscopy.duke.edu/sampleprep.

Lee GR Bithell TC Foerster J Athens JW and Lukens JN (1993). Eds. *Wintrobe's Clinical Hematology*, Ninth Edition. Philadelphia: Lea & Febiger.,158-194.

Lowe DG (1999). Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision*., **2:** 1150-1157.

Mahmood NH Lim PC Mazalan SM Azhar M and Razak A (2013). Blood Cell Extraction Using Color Based Extraction Technique. *International Journal of Life Sciences Biotechnology and Pharma Research., **2:*** 233-240

Najeed K A and Hogg DC (2014). Unsupervised learning of object detectors for everyday scene, International Journal of u- and e- Service, Science and Technology., **7:** 159-176.

Pagaro PM Jadhav P (2013). Hematological aspects in malaria. *Med. J. DY. Patil. Univ.,* **6:** 175-178.

Pherson Mc and Pincus MR (2007). Blood and Tissue Protozoa, Henry's Clinical Diagnosis and Management by Laboratory Methods, New Delhi: Elsevier, Division of Reed Elsevier, 21st ed, **67:**.1127-34.

Sammouda R Niki N Nishitani H Nakamura S and Mori S (1998). Segmentation of sputum color image for lung cancer diagnosis based on Neural Network. *IEICE Transactions on Information and Systems.,* **8:** 862-870.

Sample preparation, (2015). School of Biochemistry and Immunology, Microscopy and Imaging Facility, Trinity College Dublin.

Soltanzadeh R Rabbani H (2010). Classification of three types of red blood cells in peripheral blood smear based on morphology, IEEE 10th International conference on signal processing proceedings, Beijing., 707-710.

Somasekar J. (2011). Computer vision for Malaria parasite. Classification in erythrocytes. *International Journal on Computer Science and Engineering*., **3:** 2251-2256

Suryakantha AH (2010). Vector borne diseases. *In*: Community Medicine with recent advances. Bengaluru: *Jaypee*., 399-436.

Taher F and Sammouda R (2010). Morphology Analysis of Sputum Color Images for Early Lung Cancer Diagnosis, Proceeding of 10th International Conference

on Information Science, Signal Processing and their Applications, Kuala Lumpur, Malaysia., 296-299.

Taher F N Werghi Al-Ahmad H and Donner C (2013). Extraction and Segmentation of Sputum Cells for Lung Cancer Early Diagnosis. *Algorithms*., 512-531.

USAF-Public Health Information and Resources. Access on December 2015, http://www.Phsource.us/ PH/PARA/Chapter_9.htm.

WHO/UNICEF report (2015), *Achieving the malaria MDG target.*