MP-IDB: The Malaria Parasite Image Database for Image Processing and Analysis

Andrea Loddo[1(✉)], Cecilia Di Ruberto[1], Michel Kocher[2], and Guy Prod'Hom[3]

[1] Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy
{andrea.loddo,dirubert}@unica.it
[2] Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
michel.kocher@heig-vd.ch
[3] Institute of Microbiology, University of Lausanne and University Hospital Center, Lausanne, Switzerland
guy.prodhom@chuv.ch

Abstract. The visual analysis of peripheral blood samples is an important test for blood illnesses diagnosis, like leukaemia or malaria. Malaria is an epidemic health disease and a rapid, accurate diagnosis is necessary for proper intervention. Generally, pathologists visually examine blood stained slides for malaria diagnosis. Nevertheless, this kind of visual inspection is subjective, error-prone and time-consuming. In order to overcome these issues, numerous methods of automatic malaria diagnosis have been proposed so far. Unfortunately, no public image dataset is available to test and compare such algorithms. The aim of this paper is to present the first public dataset of blood samples afflicted by malaria, specifically designed to evaluate and compare algorithms for segmentation and classification of malaria parasite species. Every image is provided with its related ground truth and parasite's classification of type and stage of life. Our purpose is to offer a new comparative test tool to the image processing and pattern matching communities, in order to encourage and improve computer-aided malaria parasites analysis.

Keywords: Malaria · Red blood cells · Medical image analysis · Blood smear images

1 Introduction

Visual examination of peripheral blood samples is an important procedure performed by expert haematologists in order to analyse a pathology or to identify the patient's health condition. The manual analysis of blood smears is tedious, lengthy, repetitive and it suffers from the presence of a non-standard precision

N. Lepore et al. (Eds.): SaMBa 2018, LNCS 11379, pp. 57–65, 2019.
https://doi.org/10.1007/978-3-030-13835-6_7

because it depends on the operators' skill. The use of image processing techniques can help to analyse, count the cells in human blood and, at the same time, to provide useful and precise information about cells morphology. Peripheral blood smears analysis is a common and economical diagnosis technique by which expert pathologists may obtain health information about the patients. Moreover, blood cells images taken from a microscope could vary in their illumination and colouration conditions, as shown in Fig. 1.

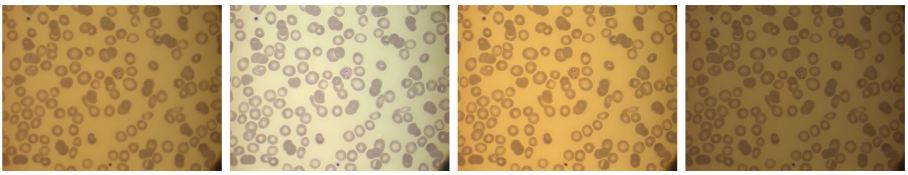

Fig. 1. Different illumination conditions could generate unconventional colour schemes in images. This is due to the absence of a standardized acquisition procedure. From left to right: same smear acquired with four microscope brightness levels. Courtesy of CHUV, Lausanne.

Typically, blood cells images contain three main components of interest: the platelets (or thrombocytes), the red blood cells (RBCs or erythrocytes) and the white blood cells (WBCs or leukocytes). It is worth considering that blood cells exist with different shapes, characteristics and colourations, according to their types. Many tests are designed to determine the number of erythrocytes and leukocytes in the blood (Complete Blood Count or CBC), together with the volume, sedimentation rate, and haemoglobin concentration of the red blood cells (blood count). In addition, certain tests are used to classify blood according to specific red blood cell antigens or blood groups. There are different calculations included in the CBC: number of red blood cells (red blood cell count, RBCC) or white blood cells (white blood cell count, WBCC) in a cubic millimetre (mm^3) of blood, a differential white blood cell count, a haemoglobin assay, a hematocrit, calculations of red cell volume, and a platelet count. Human malaria infection is not strongly related to cell count, but it needs different tests in order to be identified. It can only be caused by parasitic protozoans belonging to the Plasmodium type. The parasites are spread to people through the bites of infected female Anopheles mosquitoes, called "malaria vectors". There are five parasite species that cause malaria in humans and two of these species, Plasmodium Falciparum and Plasmodium Vivax, constitute the greatest threat. Plasmodium Ovale, Plasmodium Malariae and Plasmodium Knowlesi are the three remaining species that are less dangerous in humans [8]. Figure 2 schematically shows the morphologic and shape differences between the four acquired types and the related stages of life.

Computer vision techniques for malaria diagnosis and recognition represent a relatively new area for early malaria detection and, in general, for medical

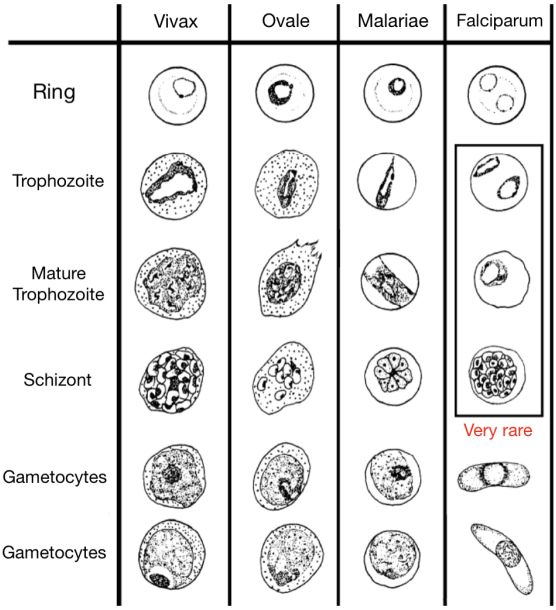

Fig. 2. Morphological scheme of human malaria parasites types and stages of life. Courtesy of Doc. Guy Prod'Hom, CHUV, Lausanne.

imaging, able to overcome the problems related to manual analysis, which is performed by human visual examination of blood smears [5]. The whole process requires an ability to differentiate between non-parasitic stained components/bodies (e.g., red blood cells, white blood cells, platelets, and artefacts) and the malarial parasites using visual information. If the blood sample is diagnosed as positive (i.e., parasites present), an additional capability of differentiating species and life-stages (i.e., identification) is required to specify the infection. Numerous methods of automatic malaria diagnosis have been proposed so far, in order to overcome the issues before mentioned. This kind of diagnosis uses images extracted from blood smears pictures taken by the microscope, after a staining process performed on the smears. Two main factors are generally considered if we refer to staining techniques: the type of colouration, in which Giemsa and Leishman are the most common, and the thickness of blood slide, which may be thick or thin. Typically, thin smears permit the identification of specific parasitic stage and quantification of malaria parasites; on the other hand, thick smears are better if the target is to perform an initial identification of malaria infection using blood pathology. Some examples are shown in Fig. 3. Giemsa stained blood smear is considered in most of the analysed literatures, whereas the Leishman stain is considered in few studies. It is reported that the Leishman stain has more sensitivity for parasite detection than Giemsa [2] and is superior for visualization of red and white blood cell morphology [6]. On the contrary, Giemsa stain highlights both malaria parasites and white blood cells and, therefore, it

is an additional issue to deal with. The Giemsa stain is a more costly and also time-consuming procedure than Leishman. Moreover, magnification of 100X by using an oil immersion objective is used for capturing microscopic images of thin blood smear for identification of specific parasites and their infected stage. Due to the laboratory availability, this dataset has been acquired only from thin blood smears coloured with Giemsa stain. The dataset is public and free available by contacting the main author. This paper is structured as follows. Section 2 offers a deep description about malaria parasites, their morphology and characteristic. Section 3 describes how the dataset has been organized, blood samples analysis, acquisition and staining procedure, the images characteristics and parasites classification. Conclusions are given in Sect. 4.

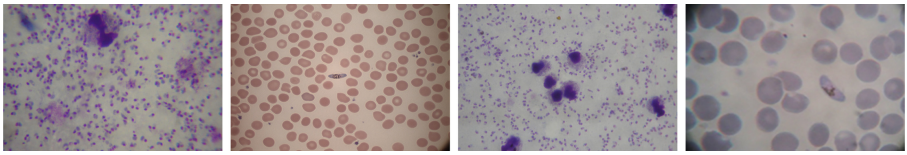

Fig. 3. Malaria infected blood smears types. This image shows a comparison between staining colouration procedures and smears thickness. From left to right: thick smear with Giemsa stain [6], thin smear with Giemsa stain [4], thick smear with Leishman stain [6], thin smear with Leishman stain [4]. Dots in thick smears and rings in thin smears are P. Falciparum ring stages, while elongated erythrocytes (in images on the right) are affected from P. Falciparum in its trophozoite schizont stage. The difference between thick and thin smears is clearly evident by observing cells and parasite shapes. Thin smears typically offer a better shape representation, while thick ones contain smaller and less clear region shapes. Furthermore, Giemsa stain shows a better contrast between cells, parasites and background respect to Leishman stain.

2 Parasite Morphology

A blood smear image, obtained through a microscope, is presented in Fig. 4. It typically contains at least three regions of interest: white blood cells (or leukocytes), red blood cells (or erythrocytes) and platelets (or thrombocytes). Two different categories of leukocytes exist: granulocytes (composed, in turn, of neutrophils, basophils and eosinophils). On the other hand, leukocytes without granules are called agranulocytes (composed of lymphocytes and monocytes). Erythrocytes do not have any subcategory even though malaria parasites (MPs) can infect them, consequently modifying their shape, morphology or colouration conditions. In particular, Fig. 5 shows several examples of malaria parasites in their different life stages and type. Although MPs infect only RBCs, a blood smear image representing both WBCs (particularly granulocytes) and MPs could be very difficult to analyse because of the similarities in colouration and shape between parasites and WBC grains, as shown in Fig. 4.

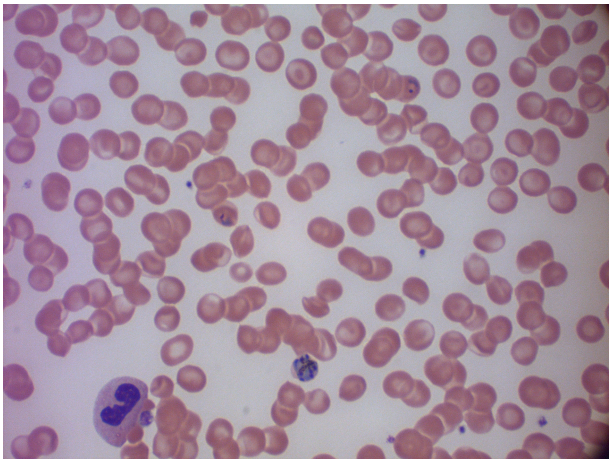

Fig. 4. Example of blood smear image acquired with a good colouration and illumination scheme. The image is characterized by three different regions of interest: a eosinophil granulocyte on bottom left, a schizont Plasmodium Falciparum on bottom centre and all the remaining cells are erythrocytes. Courtesy of CHUV, Lausanne.

MP-IDB collects four malaria parasite species: Plasmodium Falciparum, Ovale, Malariae and Vivax, in four different life-cycle stages: ring, trophozoite, schizont and gametocyte. It must be noted that Plasmodium Falciparum trophozoite and schizont are very rare and they are not present in our data collection. A complete set of examples, extracted from the dataset, are shown in Fig. 6. The life-cycle-stage of the parasite is defined by its morphology, size and the presence or absence of malarial pigment. The species differ in the changes of infected cell's shape, presence of some characteristic dots and the morphology of the parasite in some of the life-cycle-stages [7]. An automated malaria parasites analysis on blood smears usually comprises four different tasks, as follows:

1. Image preprocessing: the images are normalized in colouration, because it can differ a lot from image to image, and the different regions of interest are made the most contrasted possible.
2. Segmentation: red blood cells and/or parasites are separated from the background and white blood cells by using algorithms based on different characteristics of the cells (e.g. shape, colour, texture).
3. Feature extraction: relevant characteristic (e.g. shape, colour, texture) are extracted from the different region of interest in order to train an automatic parasite analyser.
4. Classification: several classification schemes can be performed. Hierarchically, cells are classified in red blood cells and white blood cells. Afterwards, red blood cells are classified in affected from parasite(s) or not. In the end, parasites are classified in their type and life stage. Parasites potentially can also be present outside the cells. In this case, they should need a more specific and dedicated analysis.

3 Dataset Description

Dataset images have been acquired with a Leica DM2000 optical laboratory microscope at Centre Hospitalier Universitaire Vaudois (CHUV) coupled with a built-in camera and software. The entire procedure has been realized under the supervision of expert radiologists, headed by Dr. Guy Prod'Hom. Every image is stored in PNG format with a 2592×1944 resolution and 24 bit colour depth. The images are taken with the same magnification of the microscope: $100\times$. This dataset is composed of 229 images, representing four different kinds of malaria parasite. Plasmodium Falciparum is present in 122 images, Malariae in 37, Ovale in 29 and Vivax in 46. Each image contains, at least, one parasite. Our dataset can be used either for testing segmentation capability of algorithms or classification system methods. It contains about 48000 blood cells, in which malaria parasites have been labelled by expert radiologists. The number of candidate parasites present in the MP-IDB is equal to 840. Specific counting, per parasite type and stage of life, is shown in Table 1. The annotation of the dataset images is described as follows. The image filenames are named with the following

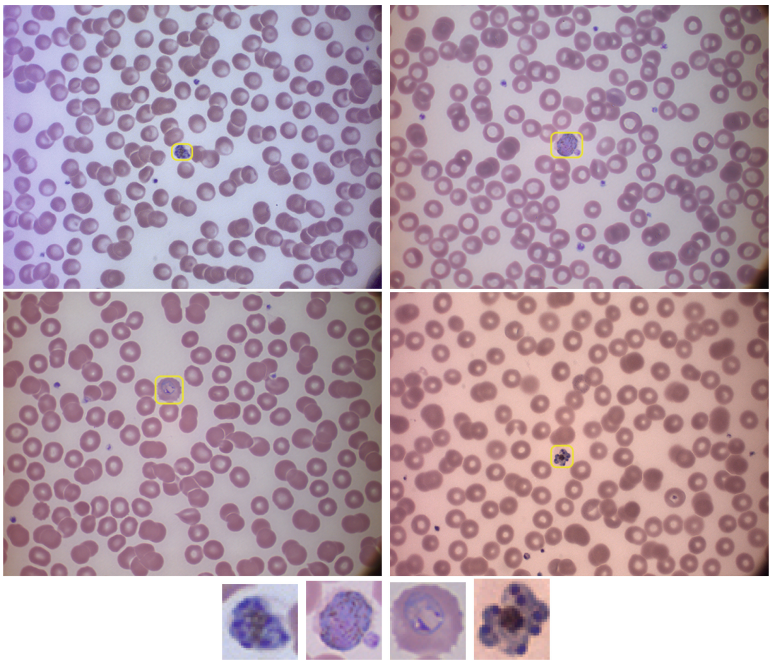

Fig. 5. Types of malaria parasites: from top left, clockwise, *P. Falciparum* in its schizont stage, *P. Vivax* in a gametocytes specimen, *P. Malariae* in its schizont stage, *P. Ovale* in its ring stage. All parasites have been surrounded with a yellow box. Underneath, from left to right: crops of P. Falciparum schizont, P. Vivax gametocyte, P. Ovale ring and P. Malariae schizont, taken from the boxes. Courtesy of CHUV, Lausanne. (Color figure online)

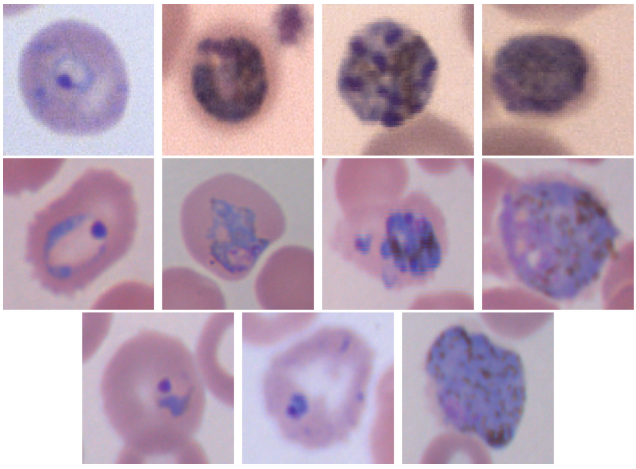

Fig. 6. Examples of malaria parasite stages. From top left: P. falciparum ring, trophozoite, schizont, gametocyte; P. ovale ring, trophozoite, schizont, gametocyte; P. vivax ring, developed trophozoite, gametocyte [4].

Table 1. Composition of dataset's images

Dataset properties		
Parasite (images)	Stage of life	Quantity
P. Falciparum (122)	Ring	695
	Trophozoite	2
	Schizont	20
	Gametocyte	3
P. Vivax (41)	Ring	9
	Trophozoite	27
	Schizont	1
	Gametocyte	10
P. Ovale (29)	Ring	13
	Trophozoite	11
	Schizont	1
	Gametocyte	8
P. Malariae (37)	Ring	1
	Trophozoite	18
	Schizont	10
	Gametocyte	11

notation: "ImXXXPR.png", in which "XXX" identifies a 3-digit integer counter, P represents one of the four parasite species ('F' for P. Falciparum, 'M' for P. Malariae, 'O' for P. Ovale, 'V' for P. Vivax), while the final R stands for the life stage ('R' for ring, 'S' for schizont, 'T' for trophozoite and 'G' for gametocyte stage). Every single image file has a reference text file with the filename notation set to "ImXXXC.xyc", which reports the coordinates of the parasites centroids. In particular, they have manually been estimated by a skilled radiologist at CHUV. These dataset images have been acquired with the same microscope. Unfortunately, lots of them suffer from different issues, like a typical non-uniform background illuminations and overexposed borders, due to the illumination of the microscope lamp (visible in Fig. 1). It causes also that the regions of interest can have different colouration, also due to the age of the analysed smears. It justifies a strong pre-processing step to make the image conditions the most similar possible, in order to realize an automated procedure. In fact, even though the images still remain intelligible, classic segmentation methods, e.g. based on thresholding, can suffer of these issues.

4 Conclusion

In this paper, we have shown how malaria parasite analysis is currently performed and which are the principal issues to deal with. Moreover, we have proposed a public dataset of blood samples, specifically designed to evaluate and compare the performances in segmentation or classification of malaria parasites by computer vision techniques. Our aim in realizing MP-IDB is to offer a strong image processing dataset, specifically designed to help in encourage new studies about malaria image analysis under a fair comparative approach based on a common dataset, like what ALL-IDB [3] has offered for leukaemia detection and white blood cells analysis [1]. We strongly discourage the use of this dataset for different activities than the purpose of this initiative.

References

1. Di Ruberto, C., Loddo, A., Putzu, L.: A leucocytes count system from blood smear images: segmentation and counting of white blood cells based on learning by sampling. Mach. Vis. and Appl. **27**(8), 1151–1160 (2016)
2. Khan, N., Pervaz, H., Latif, A., Musharraf, A., Saniya: Unsupervised identification of malaria parasites using computer vision. In: Proceedings of the 2014 11th International Joint Conference on Computer Science and Software Engineering, pp. 263–267 (2014)
3. Labati, R.D., Piuri, V., Scotti, F.: All-IDB: the acute lymphoblastic leukemia image database for image processing. In: 18th IEEE International Conference on Image Processing, pp. 2045–2048, September 2011. https://doi.org/10.1109/ICIP.2011.6115881
4. Loddo, A., Di Ruberto, C., Kocher, M.: Recent advances of malaria parasites detection systems based on mathematical morphology. Sensors **18**(2), 513 (2018)

5. Rosado, L., da Costa, J.M.C., Elias, D., Cardoso, J.S.: A review of automatic malaria parasites detection and segmentation in microscopic images. Anti-Infect. Agents **14**, 11–22 (2016)
6. Sathpathi, S., et al.: Comparing Leishman and Giemsa staining for the assessment of peripheral blood smear preparations in a malaria-endemic region in India. Malaria J. **13**(1), 512–516 (2014)
7. Somasekar, J.: Computer vision for malaria parasite classification in erythrocytes. Int. J. Comput. Sci. Eng. **3**(6), 2251–2256 (2011)
8. WHO: Malaria fact sheet December 2016. http://www.who.int/mediacentre/factsheets/fs094/en/ (2016). Accessed 06 Mar 2017