






Classification of Plasmodium-Infected Erythrocytes Through Digital Image Processing

Juan Valentín Lorenzo-Ginori¹(✉) , Lyanett China-Valdés¹,
Yanela Izquierdo-Torres¹, Rubén Orozco-Morales¹ ,
Niurka Mollineda-Diogo², Sergio Sifontes-Rodríguez²,
and Alfredo Meneses-Marcel² 

¹ Universidad Central Marta Abreu de Las Villas, 54830 Santa Clara,
Villa Clara, Cuba

juanl@uclv.edu.cu

² Centro de Bioactivos Químicos, 54830 Santa Clara, Villa Clara, Cuba

Abstract. The development of antimalarial drugs requires performing laboratory experiments that include the analysis of blood smears infected with *Plasmodium*. Analyzing visually the resulting microscopy images is usually a slow and tedious task prone to errors due to fatigue and subjectivity of the analysts. These facts have motivated the creation of digital image processing systems to automate this analysis. In this work a computer vision solution to process microscopy images of blood smears containing erythrocytes infected with *Plasmodium* is shown. This system performs tasks like illumination and color correction, image segmentation including splitting of clumped objects and extraction of color features. A set of different classifiers was tested and evaluated to find the best one in terms of indexes of effectiveness. A new feature named pixels fraction was introduced and used together with a number of other color-related features, from which a subset to classify cells into normal or infected was selected. The classifiers evaluated were: support vector machines (SVM), K-nearest neighbors (KNN), J48, Random Forest (RF), Naïve Bayes and linear discriminant analysis (LDA). All of them were evaluated in terms of correct classification rate, sensitivity, specificity, F-measure and area under Receiver Operating Characteristic (ROC) curve (AUC). The effectiveness of the pixels fraction as a new feature was demonstrated by the experimental results. In regard to classifiers, J48 and Random Forest showed the best results.

Keywords: Malaria · Image processing · Computer vision · Feature extraction · Classifiers

1 Introduction

Malaria is an infectious disease showing high degrees of morbidity and mortality, for which the World Health Organization estimated 219 million of infected people and 475000 deaths in 2017 [1]. This serious health problem demands new diagnose tools and anti-malarial drugs. Microscope analysis of large amounts of blood smears in order

to detect the presence of the *Plasmodium* parasite, both to diagnose the disease in humans as well as determining the infection rate in laboratory mice, is an issue of crucial importance during the process of developing anti-malarial drugs. The analysis of blood smears by human experts tends to be slow and tedious, and its results are prone to errors due to tiredness and subjectivity and to the frequently found low rate of positive cases (infected erythrocytes). This situation has motivated a research effort to develop solutions based on digital image processing (DIP) and computer vision (CV) for this process and this has been the objective of the present work.

There are a number of published works dealing with this problem. Diverse image processing methods have been developed to obtain appropriate image features which allow an effective classification of erythrocytes into normal and infected, examples of which can be found in [2–7]. These methods typically include tasks such as image conditioning (non-uniform illumination correction, filtering and color normalization), image segmentation, feature extraction and classification of erythrocytes. It is stressed that testing and selecting effective features and classifier algorithms constitute a key step in the design of the whole system [3]. The use of various classifiers for this purpose, like linear discriminant analysis (LDA), K-nearest neighbors (KNN), support vector machines (SVM), among others, have been reported in the literature [8–13].

The contribution of this research consists in defining new color features having high discriminating capabilities, as well as testing their use together with various classifier algorithms to find the best combination. A complete algorithm was developed, oriented mainly towards the applications in the development of anti-malarial drugs, where the analysis of blood smears from laboratory mice demands a low rate of false positives as a requirement.

2 Materials and Methods

2.1 Sample Images

In this research digital images obtained from Giemsa-stained blood smears from mice experimentally infected with *Plasmodium berghei* were used, as well as a Zuzi 122/148 tri-ocular microscope, equipped with a Microscopy 319 CU digital camera with 3.2 MP resolution and 8-bit RGB output without compression, producing a matrix having 2048×1536 pixels, with pixel size $3.2 \times 3.2 \mu\text{m}$, signal to noise ratio 43 dB and optical magnification $50\times$. The digital images were saved in.tiff (tagged image file) format. An annotated database was created with the aid of two expert analysts from Centro de Bioactivos Químicos (CBQ). This database was intended to perform all the DIP-CV procedures to obtain the features, training the classifiers and performing tests to evaluate the classification effectiveness. A set of 211 images was obtained, from which 600 images of individual erythrocytes were formed, comprising 400 un-infected and 200 *Plasmodium*-infected cells, as shown in Fig. 1. Notice the reddish-purple spots inside individual erythrocytes that harbor the parasites.

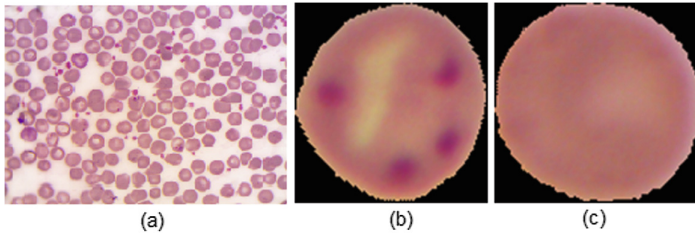


Fig. 1. Microscopy images employed. (a) Image of a blood smear showing multiple erythrocytes, (b) segmented, infected erythrocyte and (c) segmented, normal erythrocyte.

The size of the sample set was calculated from [14] as a minimum required to obtain a reasonable error when evaluating the correct classification rate CCR, which is expected to be above 0.95 for the classifiers tested. These numbers took into account also some class-imbalance. The images size for individual cells depends on their physical (variable) dimensions and can also be affected by the presence of the parasite.

2.2 Image Conditioning

The images used in this research were acquired with a Zuzi 319 CU camera coupled to a microscope. The photos were taken in the RGB color space at 8 bpp/channel. The intensity of the color components was normalized to the interval $[0, 1]$ and converted afterwards to the HSI color space. Other pre-processing steps applied were $[3 \times 3]$ median filtering to the intensity component and a morphological *top hat* with an appropriate structuring element to compensate illumination imbalance. Conversion of the images to the $L^*a^*b^*$ space was made also after segmentation to allow obtaining more features. Information on color spaces is given in [15].

2.3 Segmentation

Segmentation of erythrocytes was performed in two steps. Firstly, a coarse segmentation using the Otsu's algorithm as in [2] was applied adaptively to the intensity component of the image, which was divided into 16 patches that were segmented independently. This coarse segmentation binarized the image into foreground objects (cells, including clumps) and background. Then the cell clumps were detected and segmented employing a version of the algorithm described in [16], using weighted outer distance and marker-controlled watershed transforms, with the regional maxima of the distance transform acting as internal markers. Other components of the blood smears like leukocytes and platelets were suppressed using the procedure described in [3]. Area opening allowed eliminating other artifacts as well, using a threshold derived from the median size of the erythrocytes.

2.4 Color Normalization

In microscopy images, color can be altered due to changes in the illumination source and to the procedure of preparing the samples. This led to the necessity of color normalization by means of DIP techniques. Here the method described in [5] was used for this purpose and the results are illustrated in Fig. 2.

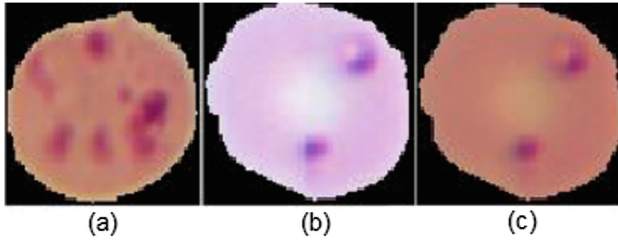


Fig. 2. Color normalization: (a) reference, (b) target and (c) color-corrected images.

2.5 Feature Extraction and Pixels Fraction

In this work, one main goal is achieving a fairly good classification result using only features derived from the color properties of the images. A total of 13 features were obtained for each of the RGB, HSI and $L^*a^*b^*$ color spaces. These were, for each color component: mean, variance and skewness as described in [17], as well as kurtosis and a new feature which is the main contribution of this work, called pixels fraction. For the three color spaces this led to 39 features in total.

The pixels fraction is defined here based on the relative coincidence of the pixel values in the three planes corresponding to the color channels respectively, for the color spaces employed. The reasoning behind this idea is that pixels having the color properties associated to the presence of the parasite in an erythrocyte, would be practically absent in the images of healthy erythrocytes. Only a minimum proportion of them could exist eventually due to the presence of artifacts in the images, like for example tiny spots created during the preparation of samples for the microscope. The pixels fraction was thought in this case to allow discriminating the presence of a real parasite from the occurrence of a colored artifact, thus improving the classification indexes of effectiveness when using features derived from the color properties of the images. Following this idea, a small set of regions of interest (ROI) were located inside the reddish-purple colored region characteristic of the parasites in a color-normalized erythrocyte, and taken as a reference, as shown in Fig. 3a. For these regions, the mean and standard deviation of the intensity in each color channel for the color spaces used was calculated. To illustrate this for the RGB color space, consider a cell being analyzed. Then the number of pixels is determined for it, whose three color components imC in which C can take the values R (red), G (green) or B (blue), fall simultaneously within the following intervals:

$$\mu_c = \begin{cases} imC > \mu_c - \sigma_c * 2 \\ imC < \mu_c + \sigma_c * 2 \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

In Eq. 1, μ_c is the mean value and σ_c the standard deviation of the component value inside the ROI and the indexes C indicate the color component to which they correspond. The factor 2 multiplying σ widens the acceptance intervals for a given component and was determined heuristically. The pixels fraction p_f is finally determined for the image of an erythrocyte in a specific color space by dividing the number of pixels n_f satisfying Eq. 1 by the total number N of pixels in the image.

$$p_f = \frac{n_f}{N} \quad (2)$$

The value of p_f was determined analogously in the HSI and L*a*b* color spaces. Figure 3b illustrates the general process followed to calculate the pixels fraction.

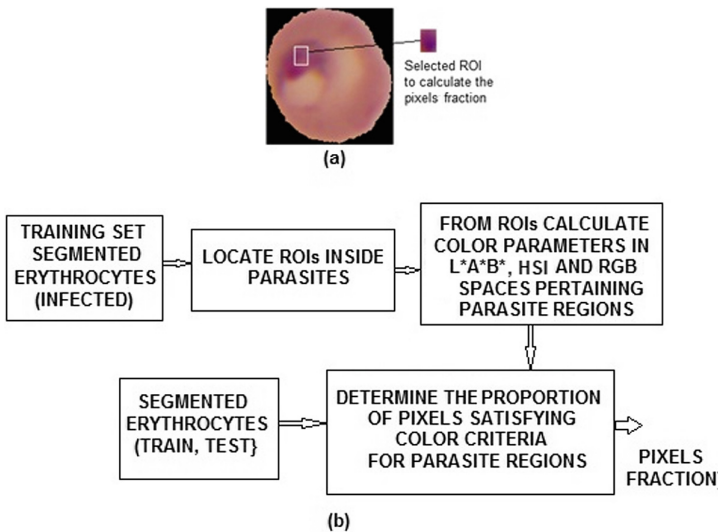


Fig. 3. Illustrating the procedure used to calculate the pixels fraction, (a) ROI selection within an infected erythrocyte, (b) block diagram illustrating steps to calculate the pixels fraction.

2.6 Feature Selection and Classification

Facilities provided by Weka 3.9 [18] were used here. Firstly, filtering (*CfsSubsetEval* with a greedy stepwise search method) was used to make a selection from the erythrocyte features previously described. This resulted in seven features. Three ranking alternatives were tested for classification: (*InfoGainAttributeEva*) using the first 20 ranked features and also the first 7 (to match the number selected through filtering), as well as all the features. Then the effectiveness of these alternatives were compared.

Classification was made using the following algorithms: SVM, KNN, J48, Random Forest (RF), Naïve Bayes (NB) and Linear Discriminant Analysis (LDA). In the case of SVM and KNN, various alternatives in their parameters (polynomial and PUK kernels in SVM, $K = 1, 3, 5, 7$ for KNN) were tested the best ones were used in the comparison to the rest of the classifiers. All the features were normalized previously.

The various classifiers were compared by ten-fold cross-validation and 1/3-2/3 percentage split. The indexes of effectiveness used were the correct classification rate (CCR), sensitivity (Se), specificity (Sp), F-measure and area under the receiver operating characteristic (ROC) curve, AUC. A more realistic experiment considered also the possibility of defining visually dubious cases as a third class, a situation often encountered in practice due to spurious colored pixels. In this case confusion matrices were used to express the results.

The general process of training the various classifiers evaluated and performing the classification process is illustrated in Fig. 4.

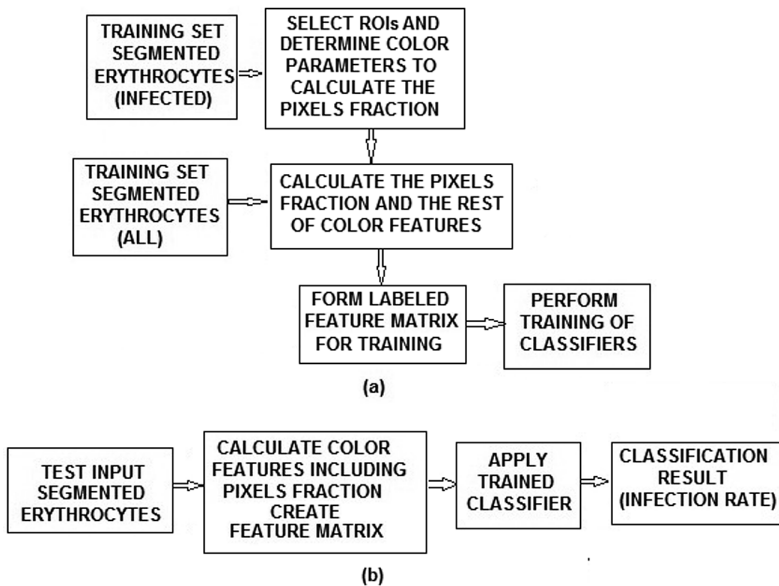


Fig. 4. Diagram of the general classification process. (a) Training the classifiers with the defined features, (b) classifying test erythrocytes.

3 Results and Discussion

The experimental work described in Sect. 2 was performed for the dataset composed of 600 erythrocytes previously mentioned. Special care was taken when building the (600×39) feature matrix corresponding to this dataset.

3.1 Feature Selection

Results from feature selection using the two methods (filter and ranking) are shown in Table 1. Notice that despite in general the seven first features ranked by the InfoGainAttributeEva method differ from those selected by CfsSubsetEval, the pixels fractions in all the color spaces were the first ones in the list, confirming their usefulness.

Table 1. Results of the feature selection process

CfsSubsetEval method		InfoGainAttributeEva, first 20 ranked features			
1	Pixels fraction HSI	1	Pixels fraction HSI	11	Variance, H
2	Pixels fraction L*a*b*	2	Pixels fraction RGB	12	Variance, R
3	Pixels fraction RGB	3	Pixels fraction L*a*b*	13	Skewness, R
4	Variance, G	4	Skewness, G	14	Skewness, L
5	Skewness, R	5	Skewness, a*	15	Skewness, H
6	Skewness, B	6	Skewness, S	16	Kurtosis, H
7	Mean, R	7	Variance, S	17	Kurtosis, R
8		8	Variance, a*	18	Kurtosis, L
9		9	Variance, G	19	Skewness, B
10		10	Variance, L	20	Skewness, b

3.2 Classification

Classification results when using 7 features obtained through ranking and filtering, are shown in Tables 2 and 3, respectively. In this case the performance measures used in a 10-fold cross-validation experiment were CCR, Se, Sp, F- measure and AUC. Classification results using the whole set of features or the first 20 in the ranking list were inferior to those shown in the tables and are not shown due to space limitations. This suggests that there is some degree of noisy behavior in the discarded features whose deletion improved the classification results. Some variants of SVM and KNN were disregarded previously in favor of those included in the tables, which exhibited better behavior. The best performance was obtained by the J48 and Random Forest classifiers, which yielded results close to 100%.

Table 2. Results of classification with features selected through InfoGainAttributeEva ranker, using the 7 best ranked features and ten-fold cross-validation

Classifier	CCR %	Sp	Se	F-measure	AUC
SVM	95.13	0.85	0.99	0.97	0.93
RF	99.95	1.00	1.00	1.00	1.00
J48	100.00	1.00	1.00	1.00	1.00
LDA	94.27	0.84	0.99	0.96	0.96
KNN, K = 1	96.20	0.91	0.97	0.97	0.95
KNN, K = 3	96.18	0.90	0.98	0.97	0.96
NB	98.50	0.95	0.98	0.99	0.99

The pixels fraction should be theoretically zero for a normal erythrocyte. However, the classification of a cell is not a trivial task: some spurious colored pixels could appear in practice and provoke an erroneous classification. Here a larger set of color-based features was employed to improve the classification results in this situation and a third “dubious” class was introduced. Table 4 shows the confusion matrix obtained in this case. This is especially important when determining the infection rate in laboratory mice through microscopy analysis, where dubious cells are usually disregarded by human analysts. In this case only four features (pixels fraction among them) were chosen by the filter selector. When using the J48 and RF classifiers, almost all dubious cases were correctly classified, all normal cells were still classified as normal and a small proportion of infected erythrocytes were classified as dubious. The success of these classifiers is consistent with the experience about their good behavior in numerous applications [19].

Table 3. Results of classification, features selected by CfsSubsetEval (Greedy Stepwise), 10-fold cross-validation

Classifier	CCR	Sp	Se	F-measure	AUC
SMO, con Puk	94.78	0.84	1.00	0.96	0.92
Random forest	99.93	1.00	1.00	1.00	1.00
J48	100.00	1.00	1.00	1.00	1.00
LDA	91.45	0.74	1.00	0.94	0.97
KNN, con K = 1	93.80	0.87	0.97	0.95	0.92
KNN, con K = 3	94.82	0.86	0.99	0.96	0.95
Naive bayes	98.67	0.96	1.00	0.99	0.99

Table 4. Confusion matrices from the classification results, considering a third class (dubious cases), J48 and RF classifiers

J48, %CCR = 98.667				Random Forest, %CCR = 98,5				
Classified as →	a	b	c		a	b	c	
Normal	a	400	0	0	a	400	0	0
Infected	b	0	159	7	b	0	159	7
Dubious	c	0	1	33	c	0	2	32

4 Conclusion

Automated classification of erythrocytes to detect the presence of *Plasmodium* parasites in blood smears is currently an open area of research and this work presents two contributions to it. The first one has been an improvement of the use of color information in the classification process through a new feature, called the *pixels fraction*, whose effectiveness was proved by two facts. Firstly, the values obtained for it in the three color spaces involved (RGB, HSI and L*a*b*) were selected among the most

important features by both the filter and the ranker feature selectors used. Secondly, the classification results using the pixels fraction were remarkable, showing that it is worth to enhance the use of color information when defining the features for classification and this is just what was achieved through this new feature. Several classifier algorithms were tested, among which J48 and RF exhibited the best results in terms of the several evaluated measures of performance. The second contribution was linking a set of image processing steps with the classifiers, to complete a computationally efficient system to classify erythrocytes in malaria studies.

Acknowledgments. The authors express their gratitude to Dr. José Antonio Escario García Trevijano from the Faculty of Pharmacy, Universidad Complutense de Madrid, who kindly donated the Giemsa-stained blood smears from mice experimentally infected with *Plasmodium berghei* used in the reported experimental work.

Conflicts of Interest. The authors declare that they have no conflict of interest.

References

1. WHO World Malaria Report (2018)
2. Arco, J.E., Górriz, J.M., Ramírez, J., et al.: Digital image analysis for automatic enumeration of malaria parasites using morphological operations. *Expert Syst. Appl.* **42**, 3041–3047 (2015). <https://doi.org/10.1016/j.eswa.2014.11.037>
3. Abdul-Nasir, A.S., Mashor, M.Y., Mohamed, Z.: Colour image segmentation approach for detection of malaria parasites using various colour models and k-means clustering. *WSEAS Trans. Biol. Biomed.* **10**(1), 41–55 (2013)
4. Preedanant, W., Phothisonothai, M., Senavongse, W., Tantisirapong, S.: Automated detection of plasmodium falciparum from giemsa-stained thin blood films. In: 2016 8th International Conference on Knowledge Smart Technology, KST, pp. 215–218 (2016). <https://doi.org/10.1109/KST.2016.7440501>
5. Tek, F.B., Dempster, A.G., Kale, I.: Parasite detection and identification for automated thin blood film malaria diagnosis. *Comput. Vis. Image Underst.* **114**, 21–32 (2010). <https://doi.org/10.1016/j.cviu.2009.08.003>
6. Das, D.K., Maiti, A.K., Chakraborty, C.: Automated system for characterization and classification of malaria-infected stages using light microscopic images of thin blood smears. *J. Microsc.* **257**, 238–252 (2015). <https://doi.org/10.1111/jmi.12206>
7. Xiong, W., Ong, S.H., Lim, J.H.: Malaria infection detection in color blood cell images using local regional features. In: Proceedings of the Second APSIPA Annual Summit and Conference, 14–17 December 2010, Biopolis, Singapore, pp. 753–756 (2010)
8. Di Ruberto, C., Dempster, A., Khan, S., Jarra, B.: Analysis of infected blood cell images using morphological operators. *Image Vis. Comput.* **20**, 133–146 (2002). [https://doi.org/10.1016/S0262-8856\(01\)00092-0](https://doi.org/10.1016/S0262-8856(01)00092-0)
9. Ajala, F.: Comparative analysis of different types of malaria diseases using first order features. *Int. J. Appl.* **8**, 20–26 (2015). <https://doi.org/10.5120/ijais15-451297>

10. Pinkaew, A., Limpiti, T., Trirat, A.: Automated classification of malaria parasite species on thick blood film using support vector machine. In: 2015 8th Biomedical Engineering International Conference BMEiCON, pp. 1–5 (2015). <https://doi.org/10.1109/BMEiCON.2015.7399524>
11. Loddo, A., Di Ruberto, C., Kocher, M.: Recent advances of malaria parasites detection systems based on mathematical morphology. *Sensors* **18**, 513 (2018). <https://doi.org/10.3390/s18020513>
12. Muralidharan, V., Dong, Y., Pan, W.D.: A comparison of feature selection methods for machine learning based automatic malarial cell recognition in whole slide images. In: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 216–219. IEEE (2016)
13. Chavan, S., Nagmode, M.: Malaria disease identification and analysis using image processing. *Int. J. Latest Trends Eng. Technol. (IJLTET)*. **3**(3), 263–269 (2016)
14. Walpole, R.E., Myers, R.H., Myers, S.L., Keying, E.Y.: *Probability and Statistics for Engineers and Scientists*. Pearson New International Edition, New York (2013). Pearson Higher Ed
15. Gonzalez, R.C.: *Digital Image Processing*. Prentice Hall, Upper Saddle River (2017)
16. Jierong, C., Rajapakse, J.C.: Segmentation of clustered nuclei with shape markers and marking function. *IEEE Trans. Biomed. Eng.* **56**, 741–748 (2009). <https://doi.org/10.1109/TBME.2008.2008635>
17. Saikrishna, T.V., Yesubabu, A., Anandarao, A., Rani, T.S.: A novel image retrieval method using segmentation and color moments. *Adv. Comput.* **3**, 75 (2012). <https://doi.org/10.5121/acij.2012.3106.75>
18. Bouckaert, R., Frank, E., Hall, M., et al.: *WEKA Manual for Version 3-6-10*. CreateSpace Independent Publishing Platform (2015)
19. Biau, G., Scornet, E.: A random forest guided tour. *TEST* **25**(2), 197–227 (2016). <https://doi.org/10.1007/s11749-016-0481-7>