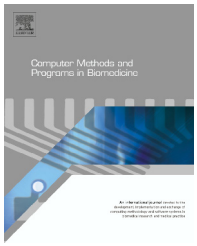
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 112 (2013) 633–639

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



An automatic algorithm for the detection of *Trypanosoma cruzi* parasites in blood sample images

Roger Soberanis-Mukul[a], *Víctor Uc-Cetina*[a], *Carlos Brito-Loeza*[a,*], *Hugo Ruiz-Piña*[b]

[a] *Facultad de Matemáticas, Universidad Autónoma de Yucatán, C.P. 97119, Mérida, Mexico*
[b] *Centro de Investigaciones Regionales Dr. Hideyo Noguchi, Universidad Autónoma de Yucatán, C.P. 97225, Mérida, Mexico*

ARTICLE INFO

Article history:
Received 5 February 2013
Received in revised form
10 July 2013
Accepted 22 July 2013

Keywords:
Pattern recognition
Image detection systems
Medical and biological imaging

ABSTRACT

Chagas disease is a tropical parasitic disease caused by the flagellate protozoan *Trypanosoma cruzi* (*T. cruzi*) and currently affecting large portions of the Americas. One of the standard laboratory methods to determine the presence of the parasite is by direct visualization in blood smears stained with some colorant. This method is time-consuming, requires trained microscopists and is prone to human mistakes. In this article we propose a novel algorithm for the automatic detection of *T. cruzi* parasites, in microscope digital images obtained from peripheral blood smears treated with Wright's stain. Our algorithm achieved a sensitivity of 0.98 and specificity of 0.85 when evaluated against a dataset of 120 test images. Experimental results show the versatility of the method for parasitemia determination.

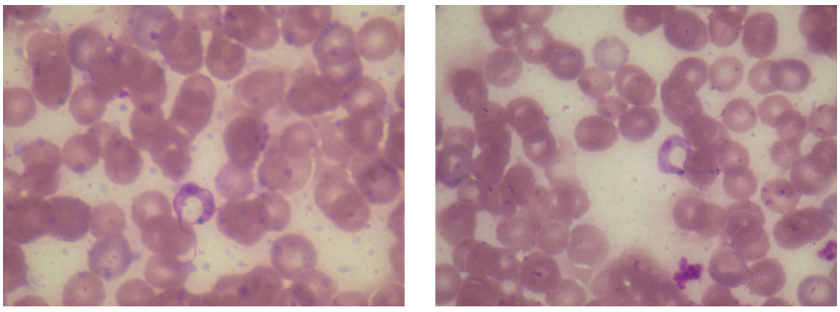
© 2013 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

According to the World Health Organization [1], the American Trypanosomiasis, also known as Chagas disease, is a potentially life-threatening illness caused by the protozoan parasite *T. cruzi*. It is found mainly in Central and South America, where it is mostly transmitted to humans by the faeces of triatomine bugs. More than 25 million people are at risk of the disease and an estimated 10 million people are infected worldwide, mostly in Latin America where Chagas disease is endemic. Approximately 20,000 deaths attributable to Chagas disease occur annually [2].

The Chagas disease presents itself in two phases. The initial, acute phase lasts for about two months after infection. During the acute phase, a high number of parasites circulate in the blood. In most cases, symptoms are absent or mild, but can include fever, headache, enlarged lymph glands, pallor, muscle pain, difficulty in breathing, swelling and abdominal or chest pain. In less than 50% of people bitten by a triatomine bug, characteristic first visible signs can be a skin lesion or a purplish swelling of the lids of one eye. When the Chagas disease is diagnosed early in this phase and a treatment is initiated, the patient can be cured. During the chronic phase, the parasites are hidden mainly in the heart and digestive muscle. Up to 30% of patients suffer from cardiac disorders and up to 10% suffer from digestive (typically enlargement of the oesophagus or colon), neurological or mixed alterations. In later years the infection can lead to sudden death or heart failure caused by progressive destruction of the heart muscle [2].

* *Corresponding author at*: Universidad Autónoma de Yucatán, Mexico.
E-mail address: carlos.brito@uady.mx (C. Brito-Loeza).
0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.07.013



(a) Polluted image (b) Stained blood cell

Fig. 1 – Examples of acquired images at pixel resolution of 2048 × 1536 with 100× zoom.

Depending on the phase of the disease some tests can be useful for making a diagnosis. According to [3] the most typical tests used for the diagnosis of the Chagas disease are: blood culture, chest X-ray echocardiogram, electrocardiogram (ECG), enzyme-linked immunoassay (ELISA), and peripheral blood smears. Up to date, one of the most effective ways of detecting the Chagas disease in its initial phase is through the ELISA test. Another commonly used method is the Chagas Stat-Pak rapid immunochromatographic test [4], which provides a performance comparable to that obtained with ELISA.

Screening blood donors for Chagas disease is of much concern in all Latin American countries. Although the World Health Organization (WHO) expert committee and some guidelines recommend a single ELISA test to screen blood donors [5], in some countries, such as Brazil, there is a more restrictive regulation, recommending two simultaneous tests of different techniques [8], performed in parallel. One of the tests that can be performed in parallel is the inspection of peripheral blood smears.

A peripheral blood smear is basically a glass microscope slide coated on one side with a thin layer of venous blood. The slide is stained with a dye, usually Wright's or Giemsa stain, and examined under a microscope. Even though visual detection of the Chagas parasite through microscopic inspection of peripheral blood smears is the most simple used technique for parasitemia determination, it is a time-consuming and laborious process. When the number of blood screenings performed in a laboratory increases, it becomes a problem. To cope with this problem, we introduce an automatic computational method for the detection of Chagas parasites. The proposed method is a first step in the process of building a more complex and efficient machine learning based model to automatically detect a variety of different parasites affecting human beings.

Chagas disease detection using automatic image analysis is, to the best of our knowledge, not yet studied as it is evidenced by the lack of publications on this topic. A first attempt was reported in [6] where Chagas parasite dectection is done through a simple Gaussian discriminant analysis method applied directly to the images in gray scale, without any other preprocessing.

A closely related work is a laser based method for the detection of blood cells infected with Malaria [7]. Although this method may also be used for the Chagas disease, its widespred adoption is limited because of the need to use a UV laser currently only available in large sophisticated instruments. Our current research work focuses in the application of image processing algorithms and the development of a low-cost diagnostic device that can be affordable for Latin American health institutions.

2. Methodology

2.1. *Image acquisition*

A group of mice were infected with an inoculation of 5×10^4 blood trypomastigotes of *Trypanosoma cruzi* via intraperitoneal. The parasitaemia detection started in average between 11 and 15 days afterwards. At this time the blood smears were prepared and stained using Wright stain, which allows the observation of the morphology of different blood cells, as well as parasites such as *T. cruzi*, Leishmania sp., Plasmodium sp., etc. After the staining process the blood smears were placed vertically and were left to dry. Finally, an optical Nikon Eclipse E600 microscope was used to take images. First at 10× and then at 100×, see Fig. 1.

Mathematically we will denote by $I(x, y) : \Omega \subset R^2 \rightarrow R^3$ a color image in the RGB format [9]. Here Ω is the bounded two dimensional domain where the image is defined and $(x, y) \in \Omega$. We also define $\boldsymbol{p}(x, y) = [r, g, b]^T$ to represent a point of $I(x, y)$, also called a pixel, in the RGB space where r, g and b stand for the red, green and blue channel information.

2.2. *Pre-processing*

The captured raw images are large in size and possibly noisy. They also may contain more than one parasite and a number of artifacts such as white blood cells, other living organisms or undesirable stain spots. Using these raw images as input to a detection algorithm may not deliver the best outcome. Therefore a pre-processing step is applied to remove all not required features and to create easier to handle sub-images of smaller size.

It was noted that the blue and green channels provide information that allow us to identify the cellular structures of the parasites. For those pixels belonging to the kinetoplast of the *T. cruzi*, it was observed that, most of the time, the blue channel has intensity values which are greater than those values of the same pixel in the green channel. This property is significantly

(a) Sub-image (b) *BGT* mask (c) Filtered mask

Fig. 2 – Illustration of the different steps of the pre-processing stage.

less frequent in pixels of not stained regions of the blood sample. Therefore, as a first step, we created a new image $BG(x, y)$ by computing the difference between channels blue and green.

Histogram distribution computed over the *BG* image in a number of tests showed that it is easy to identify by a simple thresholding the kinetoplast and nucleus of the parasite. To this end, we construct a binary mask $BGT(x, y) : \Omega \subset R^2 \rightarrow R$ using the following rule over all $(x, y) \in \Omega$:

$$BGT(x, y) = \begin{cases} 1 & \text{if } \mid BG(x, y) \mid > T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where T is a thresholding value that was found by applying the local thresholding method [10] over each one of the 60 images from the training set containing parasites. Each computed value was stored in a vector $[T_1, T_2, \ldots, T_{60}]$ and finally the mean of this vector taken as the best T. By using this methodology, we found $T = 50$ to be the best value.

As the next step, the labeling algorithm from [11] is used on the *BGT* mask to uniquely label subsets of connected components. Labeling is a method that assigns the same value (label) to all pixels belonging to the same region. This value is unique among different regions. Through such a labeling process we can compute the number of existing regions in *BGT*. The labeling process is helpful because, in addition to the *T. Cruzi*, some unwanted subsets such as blood cells or platelets with similar distribution to that of the kinetoplast and nucleus maybe still present. White blood cells are larger in size than platelets and cellular elements of the parasite hence easily removeed using area discrimination. First, area of each subset is computed by counting the number of its pixels and this value compared against a maximum area threshold. If a region exceeds such a threshold all its (x, y) entries are re-assigned to zero. Using a similarl procedure, very small stain spots can be eliminated using a minimum area threshold. At the end, we obtain a polished *BGT* mask with zero or a small quantity of artifacts. However the rest of the protozoario's body is lost in the process hence this process cannot be called a true parasite segmentation. For the remaining subsets in *BGT*, their centroids are computed to create a sub-image $I_S(x, y)$ of size 256×256 from the color image $I(x, y)$ around each centroid. In Fig. 2 we illustrate the pre-processing stage with an example. Note that as output of this stage for this example, we have the sub-image in Fig. 2(a) and the filtered *BGT* mask in Fig. 2(c).

2.3. *Segmentation*

To this point what we have is a set of sub-images reduced in size and possibly the location (white regions in the filtered *BGT* mask) of the kinetoplast and nucleus in each one of these images. We may use the sub-images set as input to the detection algorithm, however, our tests indicate that a segmented image of the parasite body delivers much better results. For this reason a simple but efficient segmentation algorithm based on the Gaussian classifier method [12,13] was implemented for that purpose.

Segmentation is a procedure employed to divide an image in several regions. This process makes a change in the image representation, allowing to perform an easier analysis [14]. In our algorithm the images were segmented in two regions or phases: background and foreground. The background consists of the elements in the image that are not of interest. Such uninteresting elements are basically those which were not stained. The second region, known as foreground or region of interest, consists of those elements whose pixels did get stained.

On what follows we will briefly describe the binary Gaussian classifier method [15] that discriminates between the two different pixel classes: stained and non-stained. In our algorithm, we decided to model each class as having unimodal Gaussian distribution.

In order this method to perform, the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$ for both classes need to be pre-computed. To this end, two large sets of images were used to build the classifier model. One set of images containing stained pixels (positive training examples) and the other one non-stained pixels (negative training examples). Before any computation is performed, the image is passed through a median filter to reduce noise. Then the mean vectors of both classes are computed as

$$\boldsymbol{\mu}_s = \frac{1}{N_s}\left[\sum r, \sum g, \sum b\right]^T \text{ and } \boldsymbol{\mu}_{ns} = \frac{1}{N_{ns}}\left[\sum r, \sum g, \sum b\right]^T \quad (2)$$

where N_s and N_{ns} are the total number of pixels in each set respectively. And the covariance matrices, also for both classes, at any pixel $\boldsymbol{p}$ are computed as

$$\mathbf{C}_s = E[(\boldsymbol{p} - \boldsymbol{\mu}_s)(\boldsymbol{p} - \boldsymbol{\mu}_s)^T] \quad and \quad \mathbf{C}_{ns} = E[(\boldsymbol{p} - \boldsymbol{\mu}_{ns})(\boldsymbol{p} - \boldsymbol{\mu}_{ns})^T] \quad (3)$$

where E represents the expectation.

Fig. 3 – Final segmentation step: (a) segmented image using the Gaussian classifier still with many artifacts, (b) filtered BGT mask and (c) final segmentation.

Once the parameters, $\boldsymbol{\mu}$ and $\mathbf{C}$ have been obtained, the classification is done using a Bayesian decision rule for minimum cost. Using this decision rule, a pixel $\boldsymbol{p}$ is considered as an stained pixel if

$$\frac{p(\boldsymbol{p}|stained)}{p(\boldsymbol{p}|non\ \ stained)} \geq \tau \tag{4}$$

where $p(\boldsymbol{p}|stained)$ and $p(\boldsymbol{p}|non\ stained)$ are the class-conditional probability distribution functions of stained and nonstained colors and τ is a threshold. It can be shown that the former decision rule is equivalent to

$$[(\boldsymbol{p}-\boldsymbol{\mu}_s)^T \mathbf{C}_s^{-1}(\boldsymbol{p}-\boldsymbol{\mu}_s)] - [(\boldsymbol{p}-\boldsymbol{\mu}_{ns})^T \mathbf{C}_{ns}^{-1}(\boldsymbol{p}-\boldsymbol{\mu}_{ns})] \leq \tau \tag{5}$$

where the value of τ that minimizes the total classification cost depends on the a priori probabilities of stained and nonstained pixels and it was determined empirically. Basically the classifier computes the Mahalanobis distances between $\boldsymbol{p}$ and the stained pixels distribution and between $\boldsymbol{p}$ and the nonstained pixels distribution. If condition (5) is true, then the pixel $\boldsymbol{p}$ is classified as stained (background). In practice, we selected $\tau = 0$.

As a result of this segmentation we obtain regions of interest that contain the parasite. As it usually occurs in this type of classifiers some other and non wanted artifacts may be classified incorrectly as foreground. In our problem, white blood cells and platelets lightly stained may have this problem. In this case, it is difficult to discriminate by area as we did in the pre-processing stage, mainly because several of these regions have similar sizes.

To solve this problem, the final step combines the results from the Gaussian segmentation method and information previously obtained from the pre-processing stage. In particular the location of the kinetoplast and the nucleus of the parasite contained in the BGT mask will be used. The final segmentation setp works as follows: if a region segmented by the Gaussian classifier contains at least one element belonging to the kinetoplast or the nucleus, the region is considered of interest. In other words, the final segmentation is an intersection of the results produced by both basic methods. An example of this process is shown in Fig. 3.

2.4. *Classification*

In the final stage our algorithm decides whether a given subimage belongs to one of two possible classes: parasite or non parasite. Here a simple k-nearest neighbors classifier [16]. The input to this classifier are the segmented images from the segmentation process described in previous section. Examples of these images are presented in Figs. 4(a)–(d). In simple words, we implemented a binary classification method based on the histogram of the segmented images. A diagram of the whole process is presented in Fig. 5.

3. Validation and experimental results

To evaluate the performance of the algorithm we use the well established technique of cross-validation [17] to assess how well our results will generalize to a different and independent data set. We ran 10-fold cross-validation experiments i.e. we randomly partitioned the data set into 10 complementary subsets of images, performed the analysis on nine subsets and validated the analysis on the other subset. To reduce variability, we ran three different rounds of cross-validation using different partitions.

For every admissible partition of each 10-fold cross-validation experiment, also called an iteration, two reference values were calculated: sensitivity and specificity [18]. Sensitivity is the probability that the classifier would consider an example as positive, given that the example is truly positive. Sensitivity is computed using the formula

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{6}$$

where TP represents the number of true positives, which is the number of positive examples classified as positive and FN represents the number of positive examples classified as false.

Specificity is the probability that a negative example will be classified as negative, and it is calculated using

$$\text{specificity} = \frac{TN}{TN + FP} \tag{7}$$

where TN is the number of true negatives and FP is the number of false positives [18].

Table 1 – Sensitivity of the *k*-nearest neighbors binary classifier fed with segmented images. Three runs of cross-validation experiments, each one containing 10 iterations are shown.

Round	Obtained sensitivity for each round and iteration										
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	Mean
#1	0.87	1	1	1	1	1	1	1	1	1	0.99
#2	1	1	1	0.83	1	1	1	1	1	1	0.98
#3	1	0.83	1	1	1	1	1	1	1	1	0.98

In Tables 1 and 2 respectively, we show the obtained sensitivity after three rounds of 10-fold cross-validation. Our algorithm showed to have high sensitivity being the smallest value obtained 0.98. The interpretation of this outcome is that only two percent of images containing a parasite were incorrectly diagnosed as being free of parasite. From the clinical point of view, a sensitivity equal to 1 is the right value since every true positive missed by the algorithm will represent a patient sent home without being given the appropriate medical treatment. On second thought, it should be considered that a blood sample from an infected in the acute phase patient will contain a considerable number of parasites therefore reducing the chances of the algorithm to get an incorrect diagnosis. That is, our algorithm only needs to identify correctly one of many parasites in the blood sample to deliver the right diagnosis. Being that current diagnostic methods are carried out by trained microscopists based on direct visualization of the parasite and due to the large number of samples they need to process in a work day, the chances of them delivering the wrong diagnosis due to tiredness seems very likely.

The second reference value computed to assess our method is specificity which measures how often our algorithm classifies an image free of parasites as an image non free of them. On average, our algorithm obtained 0.85, 0.81 and 0.83 of specificity in each round. The consequences of these not optimal values is that our algorithm will diagnose incorrectly some patients as having parasitemia. However we remark that

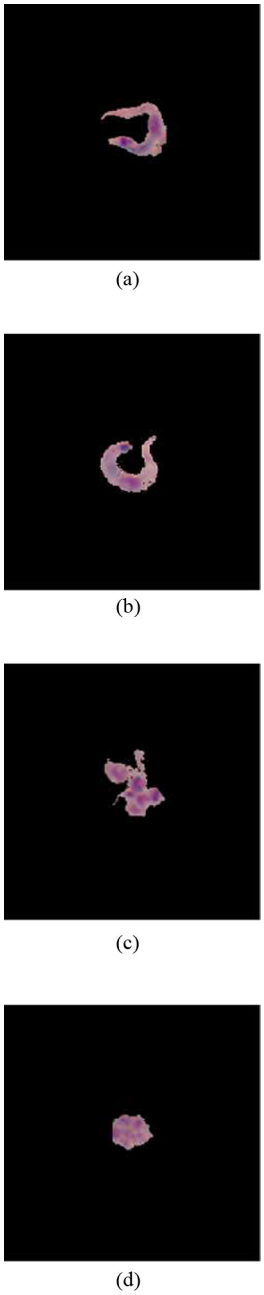


Fig. 4 – Segmented sub-images used for cross-validation.

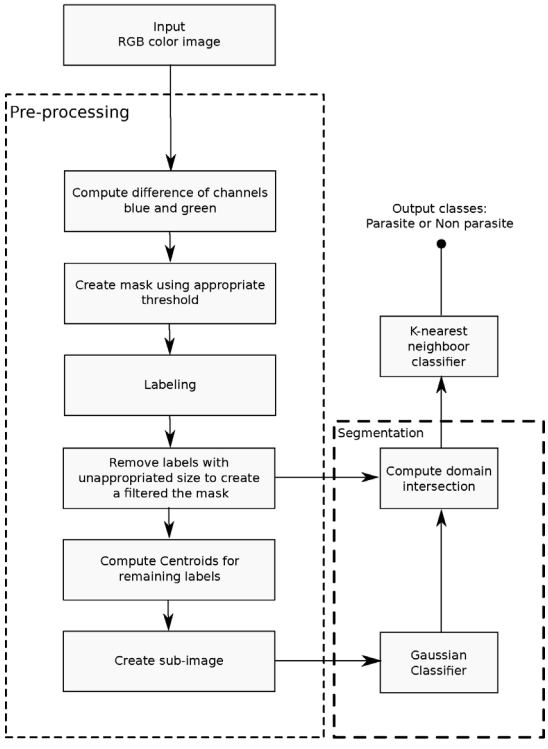


Fig. 5 – A diagram showing the whole process.

Table 2 – Specificity of the *k*-nearest neighbors binary classifier fed with segmented images. Three runs of cross-validation experiments, each one containing 10 iterations are shown.

Round	Obtained specificity for each round and iteration										
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	Mean
#1	0.8	1	1	0.57	0.78	0.75	1	0.71	1	0.88	0.85
#2	0.5	0.67	1	0.83	1	0.75	0.83	0.63	1	0.87	0.81
#3	0.5	1	0.83	0.67	1	0.83	0.83	1	1	0.67	0.83

our method must be seen as a helping tool for the microscopist which will evaluate and confirm the results from our method. From that point of view, our algorithm dismiss more than 80% of false positives highly reducing the number of images a clinician has to analyze.

3.1. *Discussion*

Some adavantages of our proposed method is that the Chagas detection process becomes automatic and a large number of blood samples can be examined in very few time. Therefore, the trained microscopists work is reduced to just validate the outcome from our computer algorithm. Also sophisticated equipment is not required to implement our method in health institutions.

On the other hand, our method in its current state still needs to improve both, the false negative and the true positive rates, which are not perfect yet. We remark that our method is an assisting tool for clinical technicians, therefore the final diagnostic is always under the supervision of expert physicians.

As part of an ongoing project, our detection algorithm will be integrated to an automated microscopic system for the acquisition of images from blood smears. Such a device will be composed of one optical microscope with a controllable electro-mechanical surface specially designed to move the blood smears as needed by the scanning software. The microscope will be equipped with a high-resolution camera in order to take the images. Images will be stored in hard drives to allow their posterior analysis with our Chagas detection algorithm. This device is currently being built as part of one CONACYT research project on Chagas disease led by Dr. Ruiz-Piña in México (project code: salud-2009-01-113848).

In a short future, we have plans to collect more clinical data increasing the size of the training set. As a result, we expect that both, the sensitivity and specificity of our algorithm, will improve. Also as part of future work, more advanced methods of machine learning will be tested aiming to get values of sensitivity closer to 1.

4. Conclusion

In this article we have proposed an algorithm for the automatic detection of *Trypanosoma cruzi* parasites in blood sample images. Evidence presented shows that detection of the parasite is successfully accomplished with our algorithm. The cross-validation experiments show a sensitivity of 0.98 and specificity around 0.80 in the classification task. As a future work, we will test more advanced classifiers such as Adaboost algorithms to improve the classification stage.

REFERENCES

[1] World Health Organization Chagas disease (American trypanosomiasis), Fact sheet number 340, June 2010 http://www.who.int/mediacentre/factsheets/fs340/en/

[2] L.V. Kirchhoff, Chagas Disease (American Trypanosomiasis), eMedicine (2010 May).

[3] L.V. Kirchhoff, Trypanosoma species (American trypanosomiasis, Chagas disease): biology of trypanosomes, in: G.L. Mandell, J.E. Bennett, R. Dolin (Eds.), Principles and Practice of Infectious Diseases., 7th ed., Elsevier Churchill Livingstone, Philadelphia, PA, 2009.

[4] C. Ponce, E. Ponce, E. Vinelli, A. Montoya, V. de Aguilar, A. Gonzalez, B. Zingales, R. Rangel-Aldao, M.J. Levin, J. Esfandiari, E.S. Umezawa, A.O. Luquetti, J.F. da Silveira, Validation of a rapid and reliable test for diagnosis of Chagas disease by detection of *Trypanosoma cruzi*-specific antibodies in blood of donors and patients in Central America, J. Clin. Microbiol. 43 (10) (2005) 5065–5068.

[5] WHO.;1; Control of Chagas disease: second report of WHO expert committee. Book Control of Chagas disease: second report of WHO expert committee (Editor ed. eds.) City, 2002.

[6] Víctor Uc-Cetina, Carlos Brito-Loeza, Hugo Ruiz-Piña., Chagas parasites detection through Gaussian discriminant analysis, Abstr. Appl. 8 (2013).

[7] Benoît Malleret, Carla Claser, Alice Soh Meoy Ong, Rossarin Suwanarusk, Kanlaya Sriprawat, Shanshan Wu Howland, Bruce Russell, Francois Nosten, Laurent Rénia, A rapid and robust tri-color flow cytometry assay for monitoring malaria parasite development", Sci. Rep. 1 (2011), 01/2011;118. DOI:10. 1038/srep00118.

[8] Secretaria de Vigilância em Saúde do Ministério da Saúde Brasil, Brazilian consensus on Chagas disease, Rev. Soc. Bras. Med. Trop., vol. 38, no. 3, pp. 7–29, 2005.

[9] A. Charles, Poynton, "Digital Video and Hdtv: Algorithms and Interfaces", Morgan Kaufmann, San Francisco, CA, USA, 2002.

[10] R. Gonzalez, R. Woods, S. Eddins, "Digital Image Processing using Matlab", Pearson Prentice Hall, Natick, MA, USA, 2002.

[11] Robert M. Haralick, Linda G. Shapiro, Computer and Robot Vision, vol. I, Addison-Wesley, Reading, MA, USA, 1992, pp. 28–48.

[12] J. Yang, A. Waibel, A real-time face tracker, Proc. IEEE Workshop Appl. Comput. Vis. (December) (1996) 142–147, Paris, France.

[13] B. Menser, M. Wien, Segmentation and tracking of facial regions in color image sequences, SPIE Vis. Comm. Image Proc. 4067 (2000) 731–740.

[14] D.A. Forsyth, Ponce, J. Computer Vision: A Modern Approach, Prentice Hall, Upper Saddle River, NJ, USA, 2003.

[15] S.L. Phung, A. Bouzerdoum, D. Chai, Skin segmentation using color pixel classification: analysis and comparison,

IEEE Trans. Pattern Anal. Mach. Intell. 27 (1) (2005 January) 148–154.
[16] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.
[17] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, Encyclopedia Database Systems (2009).
[18] S. Spitalnic, Test properties I: Sensitivity, especificity, and predictive values, Hospital Physician (2004 September).