Chagas Parasite Movement Characterization based in Optical Flow

Eduardo Patrón-Anchondo

Facultad de Matemáticas, Universidad Autónoma de Yucatán

a21216365@alumnos.uady.mx

Abstract

In the article, we establish a method to determine statistical characteristics for *Trypanosoma Cruzi* in blood samples. Using the Farneback optical flow algorithm, we give a statistical analysis to prove that there exists a movement difference in a region with the presence of *T. Cruzi* and non presence. We built confidence intervals of 95% to establish lower movement bounds for detection purposes.

1 Introduction

Chagas disease (American trypanosomiasis) is a neglected tropical illness caused by the parasite *Trypanosoma cruzi*, which is primarily transmitted through the feces of *triatomine* bugs, commonly known as "kissing bugs". About 6–7 million people worldwide are estimated to be infected [6]. In the Yucatán Peninsula, these insects are present and contribute to the ongoing public health challenge of Chagas disease. The infection typically begins when the parasite enters the human body via contact with contaminated bug feces. However, alternative transmission routes include congenital transmission and the transfusion of infected blood. Once inside the host, *T. Cruzi* disseminates through the circulatory system, initiating an acute phase that is frequently asymptomatic or marked by only mild, nonspecific symptoms. Despite the subtle onset, early diagnosis is crucial since prompt treatment during this phase offers significantly better outcomes.

However, conventional diagnostic methods, which involve analyzing blood samples to detect *T. Cruzi*, can be time-consuming and require extensive expertise. This limitation underscores the need for developing efficient, automated detection algorithms that can reduce reliance on specialized personnel and accelerate the diagnostic process. The present work addresses to characterize the movement of *T.Cruzi* to establish a high confidence lower bound to select possible small regions with the presence of this parasite.

2 Review

Since the acute phase is usually asymptomatic, blood samples are the common way to give a clear diagnosis. A single parasite detected in a blood sample is enough for a positive diagnosis. By capturing the samples in photos or video, it is possible to apply computational techniques such as optical flow and machine learning. For instance, in [2], they used AdaBoost and SVM to analyze the *T. Cruzi* morphology pattern. And, in [4], they used a Residual Convolutional Neural Network to detect semantic patterns as a movement. A more practical training set used was proposed in [1], where the images were taken with a cell phone camera, and then they trained a machine learning algorithm and gave good results even though the lower causality of the phone images.

A simple optical flow algorithm such as Farnebäck [3] can be used to measure video movement. A good hint is that the movement *T. Cruzi* is the easiest characteristic to detect for an expert since they use their flagellum to move trow the blood at a high speed. This characteristic is easily measurable in an optical flow algorithm, and it can be traceable during given lengths of video. In [5], they explore the probability expression of the image velocity conditional on an image sequence; by giving patches of different regions of video, it's possible to compare the velocity of different agents in a single shot. In this work, we aim to make a similar argument with a statistical background for the implementation of simpler software for detection purposes.

3 Proposal

Using the Farnebäck optical flow algorithm [3], we aim to quantify the movement of *T. Cruzi* on blood samples. By transforming the differentials of movement of the Farnebäck algorithm to polar coordinates, we can explore the mean of the magnitude movement for a given region. The hypothesis is that a region with the presence of *T. Cruzi* has a higher mean of movement than the rest of the similar-sized regions of the video. For individual analysis of region, we used patches, which are small regions of video, as seen in figure 1, these were tagged with positive if there is a parasite and negative if not. Also, for positive patches the kind of background affects the estimation of the optical flow; then we reserve a subset of positive patches with a dense background; this will be referred to as interface patches as the background is in the interface between the red blood cells and white blood cells. Therefore, we transform the data to explore these characteristics.

3.1 Database

The blood samples came from mice reaching the acute phase, which were infected with the *T. Cruzi* parasite. The samples were extracted in 75mm long capillary tubes with a volume capacity of 30 microliters. The parasite density was 3 million parasites per milliliter. Next, while the parasites remained alive in the tube and within five minutes of sample collection, the blood was centrifuged at 500 rpm for two minutes. The capillary tube was then placed horizontally under the microscope, with the focus set on the interface between the red blood cells and plasma. The microscope employed was a Motic BA300 with a total magnification of 40X. A Canon EOS Rebel 500D camera was utilized to record videos at a resolution of 1920 x 1080 pixels at a 20 frames per second rate. The dataset consists of 72 videos.

From the set of videos, 8 videos were set aside since they contained camera movement or blur. From the 64 videos left, they manually made 154 patches that have one or multiple (nearby) parasites and 105 that don't have any; 232 parasites were captured. Between 2 and 5 positive patches were extracted by each video, so there is no overrepresentation of an individual population. The positive patches have different sizes, between 50x50, 100x100, and 200x200 pixels. The increment in sizes depends on the accumulation of parasites in a zone since our goal is to determine a region where parasites might be. Also, the positive patches contain the same amount of parasites, therefore, the movement has to be equal in every instant. For the negative patches, the size goes from 50x50 to 100x100, and they don't have parasites at any frame. In this stage 29 of the positive patches and 28 negative patches were matched up for interface patches; all of them are of size 50x50. The positive interface patches only have one parasite in each.

4 Methodology

The biggest issue of encapsulating a parasite in regions is that the proportion of parasite in each frame is small, as seen in and it moves over all the space; so the parameter to estimate is related to the maximum value of movement detected in each frame of the mean of the values across the video. Besides, since we would like to implement an algorithm that can detect regions of movement, the parameter of the mean of the highest 20 magnitudes detected covers a bigger region of pixels, and these points can be clustered in only one tag, reducing the number of points to analyze. Additionally, since the maximum is greater than the mean of the 20 highest, we didn't explore the difference between them. The parameters are in table 1.

Parameter	Definition
PosMax	The mean of the maximum of a positive patch.
PosMean	The mean of the highest 20 values of a positive patch.
NegMax	The mean of the maximum of a negative patch.
InterMax	The mean of the maximum of a positive interface patch.
InterNeg	The mean of the maximum of a negative interface patch.
InterMean	The mean of the mean of the highest 20 values of a positive interface patches

Table 1: Parameters to study.

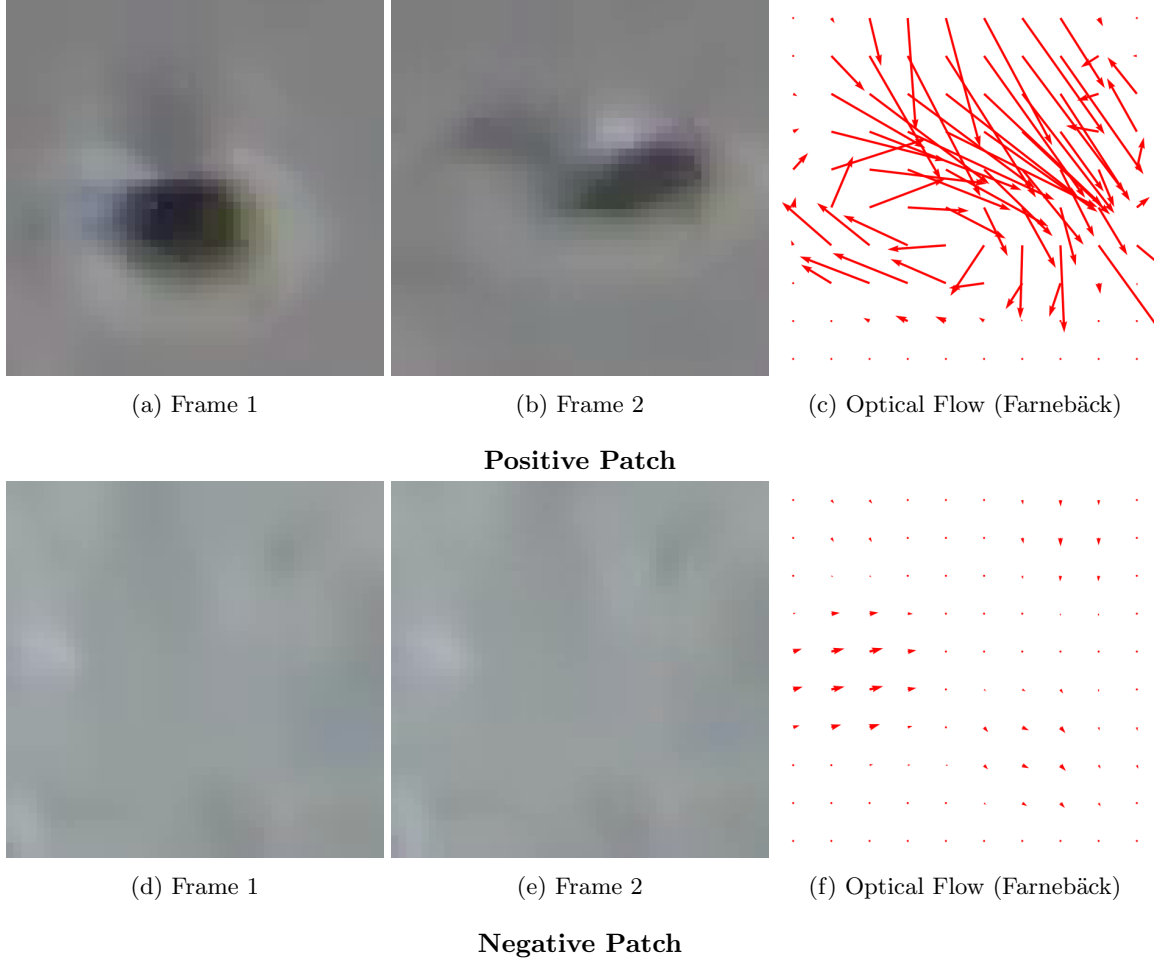These parameters were calculated for each video in the corresponding to their set. Primary, the magnitude

(a) Frame 1   (b) Frame 2   (c) Optical Flow (Farnebäck)

Positive Patch

(d) Frame 1   (e) Frame 2   (f) Optical Flow (Farnebäck)

Negative Patch

Figure 1: **Comparative**. The first row of images corresponds to the positive patch chagas1_1 and the second to the negative patch chagas3_3; frame1 and frame2 are the first frames of the video and the last figure is the flow calculated in that iteration. We see that there exists a considerable difference between them; nevertheless, we see points in the positive patch with not much movement, for this problem those points are considered as noise. Hence, the estimators proposed will try to exclude these values extending the gap of difference.

of the movement was calculated with the Färneback optical flow algorithm implementation in OpenCV. In this implementation, several parameters were configured to optimize motion estimation. The image scale was set to 0.5, allowing the creation of pyramids for multi-scale analysis. A total of three pyramid layers were used to enhance accuracy. The averaging window size was defined as 15, balancing noise reduction and detail preservation. Each pyramid level underwent three iterations, refining motion estimation. The polynomial expansion neighborhood was set to 5, with a standard deviation of 1.2 to control the smoothness of the polynomial expansion. Finally, additional flags were set to 0, keeping the default behavior of the algorithm. Subsequently, for each video, the parameter was calculated through an iteration of frames, and the resulting values were stored in an array. A lognormal distribution was then fitted to this array using scipy.stats.lognorm.fit, as this distribution is well-suited to the problem given that the magnitude is non-negative and its support is also non-negative.

To assess the goodness of fit, a Kolmogorov-Smirnov test was performed with a 95% confidence level, determining whether the data could reasonably be assumed to follow the fitted distribution. The results were recorded, indicating whether the fit passed the test (True) or failed (False), along with the expected value and median of the distribution. Finally, these values were saved in an Excel sheet corresponding to the parameter

Video	Statitic	P value	Result	Mean (dist)	Median (dist)	Mean (data)
chagas10_1	0.023879022	0.846815305	TRUE	2.865449859	2.593724072	2.862215161
chagas12_1	0.027499032	0.896641584	TRUE	1.072130978	0.760907415	1.05178485
chagas14_6	0.120954548	0.780733357	TRUE	3.895995329	3.44996047	3.887804618
chagas15_1	0.012295226	0.967434737	TRUE	14.27264662	14.01047337	14.27316188
chagas16_1	0.037637078	0.88354847	TRUE	3.615116882	3.362507463	3.613179855

Each parameter has one sheet on an xlsx file. For the result column, we have that for PosMax, 151 of 154 passed the test; for PosMean, 152 of 154; for NegMax, 89 of 105; and for InterMax and InterMean all videos passed the test. This means that more than 92% of the videos preserve a lognormal distribution.

The study of the outliers of each kind of parameter shows that the outliers of PosMax and PosMean are the same, and they are over the mean, this is shown in Figure 2. Moreover, the outliers of the figure are only in the top; the patches on related were exterminated and it was not found any kind of anomalies. For the negative patches, the box plot is shown in figure 3, there are only 2 outliers on the top. The videos from the outliers are very particular. These patches came from video 65 in which a particular flow of cells exists, this movement is from the settlement in the interface. This particular case often happens in lab conditions, nevertheless, this for optical flow detection can be mistaken as a positive region.

For the interface patches, there are no outliers; the box plot that shows this is in figure 4. In the negative interface patches, the videos that were outliers in the general negative patches are not outliers in this set. So, we can infer that the interface patches have more noise from the background, and at the same time, the maximum values are lower, then it is more difficult the detect *T. Cruzi* in the interface. In practice, for the expert is often more difficult to detect this kind, meaning that a bound value would help the reduction of regions, which leads to less time for human diagnosis.
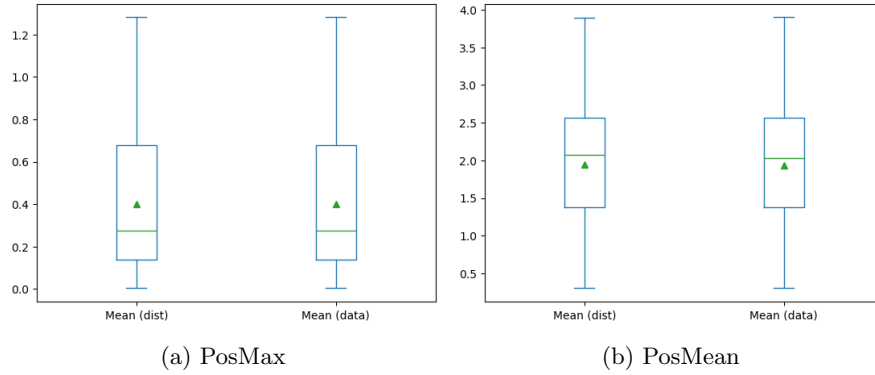

(a) PosMax　　　　　　　　　　　　　(b) PosMean

Figure 2: **Description**. These box plots are from PosMax and PosMean parameters, their is the mean over the data and the distribution represented with a triangle. We see that both a similar since the values are almost the same; hence, we see that the maximum is not an outlier of the samples.
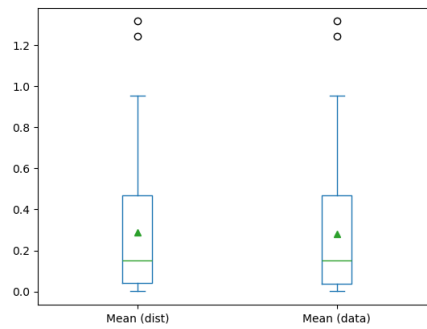

Figure 3: **Description**. This box plot is from the NegMax parameter, which is the mean over the data and the distribution represented with a triangle. The two outliers are over 1, which corresponds with negative patches in video 65; there exists a particular flow of cells.
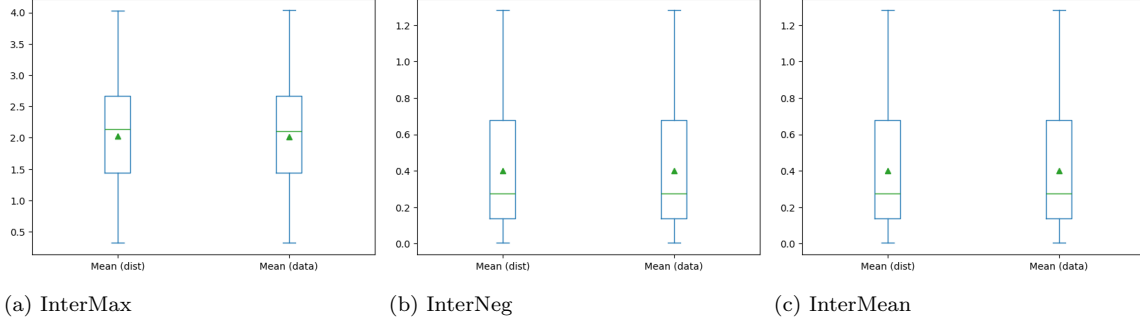
4

(a) InterMax   (b) InterNeg   (c) InterMean

Figure 4: **Description**. These are the box plots of the interface parameters, there are no outliers in these sets.

4.1 Statistical test

For the results chapter, we used two statistical proofs to compare the mean and variance of different parameters. Firstly, for the variance, we used Levene's test to determine if the sets have similar variance. The hypotheses for the test are

$$H_0 : \frac{\sigma_A}{\sigma_B} = 1 \quad vs \quad H_1 : \frac{\sigma_A}{\sigma_B} \neq 1$$

The null hypothesis is related to homoscedasticity, this means that de variances are the same. This characteristic is important for t-test and ANOVA. Nevertheless, their exists equivalently statistical proof without homoscedasticity as the Welch's test for means. We used Welch's test with the hypothesis

$$H_0 : \mu_A \leq \mu_B \quad vs \quad H_1 : \mu_A > \mu_B$$

where null hypothesis establishes that the mean set A is equal or lower than the mean of the set B and the alternative hypothesis indicates that the mean of the set A is greater than mean of B. We used this structure of hypothesis for all the tests in the results chapter. We conventionally established that the population with the mean we want to verify as smaller takes the role of set B so that if the null hypothesis is rejected with a high confidence, we can suppose the desired result.

5 Results

Our main goal in this study is to show that we can determine with movement detected by an optical flow algorithm that a region with a parasite has more movement than a region without. Then, a mean difference test was performed between the PosMax and NegMax parameters, which are independent since the data came from different patches. We will use as an estimator the mean of the data because is similar to the mean of the distribution and it doesn't depend on the result being true for the fit of the lognormal distribution. The means of these parameters are in 2.

Parameter	Outliers	Non Outliers
PosMax	3.1694	2.6944
NegMax	0.2819	0.2625

Table 2: Means of PosMax and NegMax

At first sight, the difference in both cases is visually obvious, nevertheless, we support it with a statistical argument. The next table shows Levene's test for the difference in the variance and Welch's test for the difference of means, which depends on the fact of the difference in variance. We considered de p-values for the samples with outliers and without. The results of these test are table 3.

The use of Welch's test supposed that the variances are different, which is supported by Levene's test, which has p-values lower than 0.01 in both cases. Then, since the p-values for Welch's test are lower than 0.01, we

Sample Statistics	Test	Outliers		Non Outliers	
		Stat	p-value	Stat	p-value
Variance	Levene's test	68.2624	7.65×10^{-15}	126.4355	6.07×10^{-24}
Mean	Welch's test	14.0950	3.9×10^{-30}	16.8987	3.02×10^{-37}

Table 3: PosMax vs NegMax

can be 99% confident that the mean of maximum movement for positive patches is higher than the maximum movement for negative patches. This means that by a considerable margin, we could determine a value so we can determine with high confidence the movement comes from a *T. Cruzi*.

In this sense, the parameter PosMean by construction is thought to get a bound to accept more values in a near neighborhood, so the detection won't depend on a single point. This reduces the possibilities of possible outliers of non-parasite regions and the classification of a cluster of points for a single individual. In a similar statistical argument as before we compare the NegMax and PosMean, which are independent, to see if their means are different. The means of these parameters are in table 4.

Parameter	Outliers	Non Outliers
PosMean	3.0342	2.599
NegMax	0.2819	0.2625

Table 4: Means of PosMean and NegMax

And for the results of the mean difference, we used the same argument as before, and they are explicit in table 5.

Sample Statistics	Test	Outliers		Non Outliers	
		Stat	p-value	Stat	p-value
Variance	Levene's test	67.3060	1.12×10^{-14}	124.9715	9.89×10^{-24}
Mean	Welch's test	13.9013	1.27×10^{-29}	16.6889	9.37×10^{-37}

Table 5: PosMean vs NegMax

Therefore, we know with 99% certainty that the mean of the mean of the 20 highest values is bigger than the mean maximum of the negative patches. Then, we could determine a bound with high confidence that the movement comes from a region with *T. Cruzi*. These lower bounds can be given with a confidence interval.

5.1   Lower bound for ROI detection

We proceed to construct a 95% confidence interval for the mean of PosMean; we work for two types, the symmetric and the left side confidence intervals, since the right one covers the lowest values which are related to the NegMax values. These intervals are

- The interval $(2.6506, 34179)$ is a symmetric confidence interval of 95%.

- The interval $(2.7123, \infty)$ is a left side confidence interval of 95%

So we get the lower bounds 2.6506 and 2.7123 for the mean of movement of a neighborhood with the presence of *T. Cruzi*. These bounds are for finding in all the images a cluster of points that could detect the parasite; from this cluster, we could take a point and take an ROI of the region to be looked at by an expert.

Nevertheless, by analyzing the data numerically and visually, it was observed that in interface regions, where eye detection was difficult, the PosMean values were lower than the rest; this set is going to be a statistical analysis.

5.2 Interface patches analysis

Similar to previous cases, we made a statistical analysis to determine if there exists a mean difference between positive patches with non-interface and negative interface patches. For this section, we refer to the PosMax and PosMean parameters that do not contain the interface, this is to ensure the independence between positive and interface parameters. Furthermore, since the interface parameters do not have any outliers, the test has to be without outliers in both sets. The means of these parameters are in table 6

Parameter	Means
PosMax	2.0168
InterNegMax	0.4019
InterMax	2.0168

Table 6: Means of PosMax, InterNegMax and InterMax

The results of the test of difference between InterMax and InterNegMax is in table 7

Sample Statistics	Test	Non Outliers	
		Stat	p-value
Variance	Levene's test	22.6708	1.44×10^{-5}
Mean	Welch's test	8.9297	4.42×10^{-11}

Table 7: InterMax vs InterNeg

This result implies that with a 99% confidence, there exists a significant difference in maximum movement between interface negative patches and interface positive patches. The next step is to look at the difference between InterMax and PosMax parameters; the results in table 8 exhibits a significant difference between the means, so for detection purposes the bound value has to change. Furthermore, this restriction will require the insistence of the expert to look at if the region has an interface.

Sample Statistics	Test	Non Outliers	
		Stat	p-value
Variance	Levene's test	12.8491	4.62×10^{-4}
Mean	Welch's test	3.7178	1.74×10^{-4}

Table 8: PosMax vs InterMax

Nevertheless, the parameters have different means with a 99% confidence. This difference can be attributed to a more congested region, which could lead to lower precision in the detection of ROI's. However, we need to see if the difference between the InterMean and InterNeg to determinate a lower bound of movement that reduce noise, the results are in table 9.

Sample Statistics	Test	Non Outliers	
		Stat	p-value
Variance	Levene's test	21.4201	2.29×10^{-5}
Mean	Welch's test	8.7272	6.92×10^{-11}

Table 9: InterMean vs InterNeg

Then we can state with 99% confidence that there mean of InterMean is greater than InterNeg: hence we can determinate a lower bound for the InterMean whit high confidence that the movement comes from a region with *T. Cruzi*. For the construction of confidence intervals, since there are only 29 values, we used the *t*-student bounds.
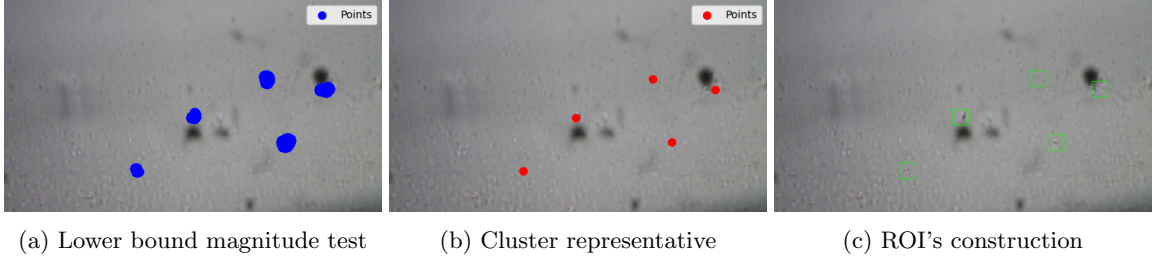
- The interval (1.60, 2.27) is a symmetric confidence interval of 95%.

- The interval $(1.66, \infty)$ is a left side confidence interval of 95%.

Then, we can consider 1.6 and 1.66 lower bounds for the mean of movement of *T. Cruzi* in interface regions. These values are lower than for all the image values, meaning that they are less precise. Nevertheless, it's possible to restrict the area to interface regions using interface filter detection and thresholding.
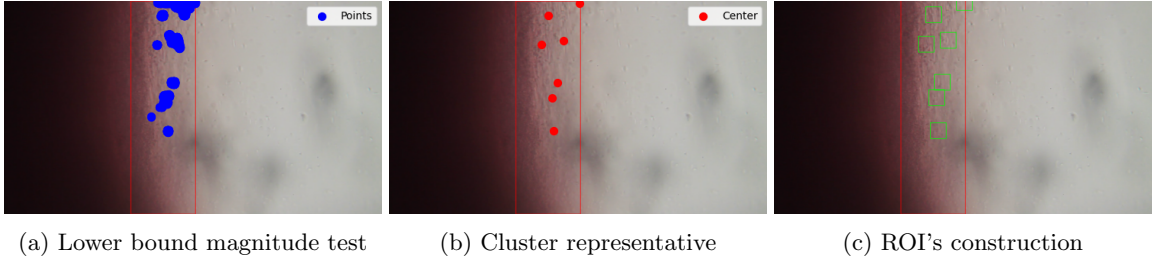
6 Application

With the lower bounds established in the previous section, we build a minor procedure to determine regions with the presence of *T. Cruzi*. For this task, we applied the Färneback optical flow algorithm with the parameters used for processing the data in section 4. To illustrate this process, we use as an example video chagas_56 where the interface does not appear, so we used the lower bound condition 2.7.
First, we calculate the Färneback optical flow of the first iteration. Then, we transform the result matrix to obtain the magnitude related to each coordinate. So, we can apply the lower bound condition to extract the coordinates where the magnitude is greater than 2.7. The coordinates that pass this test in the example are shown in figure 5a as blue points. Next, we apply a DBSCAN clustering algorithm, which clusters nearby points given a radius and minimal amount of points to determinate a cluster. Since the PosMean parameter is defined as the mean of the 20 highest values on a rectangle of radius 25, we used these values for the DBSCAN algorithm; minimal amount of points 20 and radius 25 (Chebyshev metric). We choose as representative of a cluster the middle values of the array; in the example, the representative is marked with a red point in figure 5b. Finally, we build a square of radius 40 over each cluster representative to highlight a possible region with *T. Cruzi* as shown in figure 5c.

| (a) Lower bound magnitude test | (b) Cluster representative | (c) ROI's construction |

Description. The phases of the procedure to highlight possible *T. Cruzi* regions based in the lower bound magnitude criterion.

In a similar procedure, we proceed for an interface region. Nevertheless, we have to extract this region; for this task, we used a Sobel filter and selected the biggest contour region. For video chagas_47, the biggest counter is delimited by the red lines in the figures 6a, 6b, and 6c. The process is repeated with the lower bound value of 1.6 given in the interface patch analysis.

| (a) Lower bound magnitude test | (b) Cluster representative | (c) ROI's construction |

Description. The phases of the procedure to highlight possible *T. Cruzi* regions based on the lower bound magnitude criterion in an interface region.

7 Conclusions

By the study of the main characteristic of *T. Cruzi*, we establish by statistical analysis that the mean of movement differs from the parasite and the surroundings even in dense matter regions. By a simple and innovative procedure, this approach differs from previous work since the use of the characterization is based on the usual statistical test and could state with a high confidence.

By applying the lower bound criterion for the different regions, we could structure a procedure to detect parasites and highlight these regions for an expert examination. It is important to note that in practice, a single detection is enough to confirm positive parasitemia. Nevertheless, the methodology followed has the advantage that it does not depend on some background procedure; it can be replaced as a form to characterize having control of the data and possible lousy information.

References

[1] M. Cafundó de Morais, D. Silva, M. Milagre, M. Oliveira, T. Pereira, J. Silva, L. da Costa, P. Minoprio, R. Junior, R. Gazzinelli, M. Lana, and H. Nakaya. "Automatic detection of the parasite *Trypanosoma cruzi* in blood smears using a machine learning approach applied to mobile phone images". *PeerJ*, vol. 10, e13470, May 2022. DOI: 10.7717/peerj.13470.

[2] V. Uc-Cetina, C. Brito-Loeza, and H. Ruiz-Piña. "Chagas Parasite Detection in Blood Images Using AdaBoost". *Computational and Mathematical Methods in Medicine*, 2015, Article ID 139681. DOI: 10.1155/2015/139681.

[3] G. Farnebäck. "Two-Frame Motion Estimation Based on Polynomial Expansion". In: *Lecture Notes in Computer Science*, 2003. DOI: 10.1007/3-540-45103-X_50.

[4] A. Ojeda-Pat, et al. "Effective residual convolutional neural network for Chagas disease parasite segmentation". *Medical & Biological Engineering & Computing*, Mar. 2022. DOI: 10.1007/s11517-022-02537-9.

[5] E. Simoncelli, E. Adelson, and D. Heeger. "Probability distributions of optical flow". In: *Proceedings of the 1991 IEEE Conference on Computer Vision and Pattern Recognition*, 1991, pp. 310–315. DOI: 10.1109/CVPR.1991.139707.

[6] World Health Organization. *Chagas Disease (American Trypanosomiasis) Fact Sheet*. Accessed: 2025-03-09. URL: https://www.who.int/en/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis).