1977-10-20

AUTOMATED MALARIA PARASITE DETECTION

Per Rowa
Dan Antonsson

INTERNSKRIFT
LiTH-ISY-I-0185
Abstract

A system for malaria parasite detection in thin blood-smears is presented. Sample slides prepared with standard methods are accepted. A low-cost TV-camera mounted on an ordinary microscope with a computer controlled stage is used as a picture sensor. Frames, digitized in windows of 64 x 64 pixels are fed into a special purpose picture processor at normal frame rate (25 frames/sec). In the picture processor measurements are made on the images at high speed. The classification problem is split into different levels each having different characteristics such as different sampling density. Four classes, three of which are different types of malaria parasites, are recognized. As a whole the classification is best labelled as a sequential pattern recognition procedure.

In its preliminary version the system has been run at a speed comparable to that of a human operator, that is 1 500 cells per minute. A test on 80 000 cells gave 25 false negatives out of 283 parasites (9%) and 41 false positives (0.05%).
Problem description and motivation

Malaria is one of the most dangerous and widespread diseases in the tropical and subtropical regions of the world. Once it was believed that the discovery of anti-malaria drugs and chemicals would control and eliminate the disease. Now malaria seems to be advancing again and also new resistant strains are found.

Diagnosis of malaria is presently based on the observation of the patient's symptoms and on microscopic study of blood samples. The latter is quite timeconsuming and puts a serious restriction on the number of diagnosis one can do. This prevents early treatment of the patients, makes it more difficult to localize the infectious mosquitoes and to follow up an eradication program.

The aim of this project was to investigate the feasibility of automating the diagnosis under realistic circumstances. Some previous work on these problems is presented in [3].

First some facts about malaria. See Figure 1. The disease is transmitted to man by mosquitoes of the genus Anopheles.

At a bite of an infected mosquito 'sporozoites' are injected. In the liver these develop to numerous 'merozoites' which are released into the blood where they enter the red blood cells, 'erythrocytes'. First the parasite often looks like a tiny ring, then it grows ('trophozoite' stage) into a 'schizont' which finally breaks up. This is the point where the famous fevers occur. The new released merozoites thus restart a cycle in the blood cells or 'gametocytes' may be picked up by other mosquitoes. These give girth to new sporozoites which closes the outer cycle.

Microscopic diagnosis is based on finding the parasite in the red blood cells (ring, trophozoite, schizont) since the other forms are rare.
The blood smears for diagnosis are stained by the Giemsa or Wright method. Both thick and thin smears can be used. The staining colors the parasite nuclei red and the surrounding cytoplasm blue thus contrasting to unaffected red blood cells.

Automatic interpretation of the complicated structure of the thick smears is not necessary so thin smears were used since they are easier to treat. In fact using thin smears is very advantageous provided you can screen the same volume of blood, see [2].

So the problem given the machine was to count the number of 'ringforms', 'trophozoites' and 'schizonts'. This division is rather arbitrary but indicates the state of the disease.
No attempts were made to distinguish between different main species of malaria such as Plasmodium (P) malariae, P vivax, P ovale, P falciparum, although the characteristic differences in parasite form and position are being measured.

Most of the slides used had blood infected with the malaria type P Brasilianum which resembles P malariae and was considered to be good for training since all stages of development are present at the same time.

**Equipment used**

A medium class microscope with a 100 X objective and immersion oil was used. Mounted on the microscope there is a program controlled moving stage and a low cost camera with a vidicone tube. This gives a lot of shading and a somewhat noisy picture with rather low resolution. The normal appearance of the cells results in very low contrast, say 25% of the full range of the videosignal, although green light was used to improve the contrast.

Lamp intensity was not controlled. Reasonable amounts of dust on the lenses and the vidicone tube was to be accepted. Also, no automatic focusing was used so slightly out-of-focus images were accepted.

These studies were made using the PICAP system (PICI~ture Array Processor). It was immediately obvious that the strengths of this system should be exploited from the beginning. We did not believe it was possible to first set up a good algorithm without having any special implementation in mind and then afterwards try to find some nice hardware to suit the algorithm.
PICAP is a general purpose programmable picture processor. Digitized and quantized images of 64 by 64 pixels and 16 gray levels (or 16 arbitrary states) are processed.

PICAP uses an operator with 3 x 3 pixels where the center pixel is given a new value depending on its surrounding pixels according to the programmer's specification.

All 4096 pixels are conceptually treated in parallel so that the 3 x 3-operator is applied simultaneously on all pixels. The operator is either arithmetical (weighted sum) or logical (nonlinear). Up to 9 images can be handled and pointwise logical or arithmetical operations between these 9 is also done in parallel.

The picture processor is connected to a minicomputer that handles all programs and conventional processing. The reason for using a hardware picture processor is the increased processing speed, in this case about 100 times faster than performing the same operations in the minicomputer. The PICAP also automatically delivers some otherwise time consuming measurements such as the gray-level histogram, size and coordinate information for objects in the picture and some customer programmable measurements.

A special Television Input Processor (TIP), [5] is a vital part of the system. The 64 x 64 pixel pictures with 16 gray levels would be rather limited unless TIP under program control selected where in the picture plane, with what sampling rate and in what slice of the video signal these pictures are to be fetched.
Knowing that it doesn't take a microscopist more than some 5 minutes to do a routine screening of 100 fields of view (meaning about 2 000 cells/minute) it is obvious that there is no time to do any complicated processing on every cell.

All algorithms claimed to be realistic must seriously take the processing time into account. Coming closer the final implementation you have to worry more and more about the trade offs between processing time and error rate. We wanted to come close to the manual speed even with our present equipment and hence we had to sacrifice better methods for faster methods.

So, the guiding principle when designing the programs for higher speed were first to make extensive use of the nicest facilities of the PICAP system and second, to eliminate non-infected red blood cells from further consideration as soon as possible.

Since the stained parasites are darker than normal cells it is possible first to scan the cells with low resolution and thereby rapidly find preliminary suspected areas. More elaborate measurements with higher resolution are then performed on these areas. For every measurement made a decision is taken whether to continue the investigation and if so what actions to take. For cells passing all tests measurements are put together to a feature set and classification is done with standard methods. Figure 2 shows the resulting complexity and indicates the difficulties for a complete formal optimization of the system.
The sequential procedure in detail

The detailed procedure is best illustrated with some flowcharts and a series of photographs taken while the machine was running. One field of view which happened to contain one of each "class" of parasites is used.

**Figure 3** TIP is adjusted when the illumination condition change
In order to find suspect objects the marked window is first digitized into 64x64 pixels. Time: = 40 ms. Note the uneven illumination of the field.

Since thresholding in this case is difficult or impossible a gradient picture is obtained by first producing two differentiated pictures, one for the horizontal and one for the vertical direction and then taking the maximum absolute magnitude of these two. The differentiation is based on a 5x5 neighborhood.

Thresholding and region growing (gray level restrictions) in the gradient picture. After clustering of points within a cell diameter the center coordinates for each separate object is estimated. Extent and area measurements are performed and coordinates decided not to be originating in malaria objects are not recorded.
Figure 7 A small window for high resolution input is centered on the coordinates found. The coarse scanning eliminates more than 90% of cells.

Figure 8 (TIP is adjusted to cut out the best portion of the video signal amplitude giving a "normalized" histogram.)
Figure 9  A parasite in the last (schizont) stadium. A spatial lowpass filter was also applied. Resolution: 0.2 μ/pixel.

Figure 10  Circular mask applied. This is a crude way to isolate one cell. See Figure 20 ff for backup methods when this is insufficient.

Figure 11  The threshold selection is based on the background intensity. The background is taken to have the gray level value closest to the lightest 5% of the image. The threshold is at constant distance from this value.
FEATURE MEASUREMENTS

MEASURE:
- EXTENT IN HORIZONTAL DIRECTION, $X_p$
- EXTENT IN VERTICAL DIRECTION, $Y_p$
- AREA, $A_p$
- PERIMETER, $P_p$
- $P_p^2/A_p$
- GRAYLEVEL DISTRIBUTION FACTOR, $G_p$

AND TEST EACH AGAINST LIMITS; SKIP IF LIMITS EXCEEDED

THRESHOLD FOR CELL AND LP-FILTER

TOUCHING OR OVERLAPPING CELLS?

YES

CUT OUT CONVEX OBJECT SURROUNDING PARASITE

MEASURE $X_c$, $Y_c$, $A_c$, $P_c$
- $P_c^2/A_c$
- $G_c$

TEST AND POSSIBLY SKIP

NO

Figure 12 Subscript P stands for parasite and C for cell
Figure 13 After thresholding for the red bloodcell (also based on background intensity) and filtering, the parasite and cell images are superimposed. Region growing eliminates objects outside the cell.

Figure 14 The histogram for pixels "inside" the parasite is used in one of the features ... (cont in Figure 15).

Figure 15 ... and so is the histogram for pixels in the vicinity of the parasite.
Figure 16  Next coarse scanning window (no parasites)

Figure 17  Thresholded gradient picture. None of the objects will be examined. The upper one is too large and the other one will be treated in the adjacent field of view.

Figure 18  In this window the dark trophozoite will be found and examined
Figure 19 The last window before the automatic stage moves to the next field of view. This will take \(\approx 50\) ms. The two parasites are easily found.

Figure 20 Touching or overlapping cells is a common and difficult problem in cell processing. One way of solving this problem well adapted to parallel processing, is illustrated here in the following Figures 21, 22 and 23. The "parasite" (in this case it was some impurity) is superimposed on the cell which unfortunately is connected to its neighboring cell.

Figure 21 After the normal procedure the program will analyze size and shape and find that this could be a red blood cell touching another one.
Figure 22 Shrinking, superimposing and growing "the parasite" inside the rest of the cell makes it possible to eliminate disconnected parts.

Figure 23 Re-expanding gives a better estimate of the isolated cell. Although it didn't do a perfect job the result is good enough since in the end you have to accept also non-overlapping cells even more corrupted. Two other methods we tried that performed equally well but were slower in our system.

a) Find the typical indentations and cut with a straight line.

b) Start the region growing from the parasite inside the cell and stop if size or gray-level limits are exceeded.
The preceding windows really were free from problems. Usually strange situations are met in at least every 20th field of view. How many parasites can you see??
The final classification

The sequential testing of the parameter values was good for eliminating clearly non-malarial objects one at a time but cannot be used for classification without great risk for errors. The classification is instead based on assuming a multivariate Gaussian distribution in the parameter space. If this assumption is correct and if all errors are considered to have the same "cost" then it is a wellknown result that the optimum classifier computes

\[ d_i(x) = (x-\mu_i)^t \Sigma_i^{-1} (x-\mu_i) + \text{constant}_i \]

where:

- **x**: the parameter vector
- **\( \mu_i \)**: the mean vector for class i
- **\( \Sigma_i^{-1} \)**: the inverse of the covariance matrix for class i
- **\( d_i \)**: the squared Mahalanobis distance

So, \( d_i \) is the weighted distance from the vector to be classified to the mean vector for class i and the classifier selects i for minimum \( d_i \).

There are at least 4 assumptions in the above procedure that do not hold strictly.

1) The features do not have a Gaussian distribution. Although the parameters were intensionally selected to be Gaussian this is an approximation. Some parameters such as \( P^2/A \) are far from Gaussian.

2) The cost function is not symmetrical. For instance, a realistic setting of the costs for misclassification would certainly assign a higher cost for losing a parasite than for confusing two different types.
3) $\Sigma_i^{-1}$ and $\mu_i$ in the formula should be the "true" values but all you can do is to estimate them and thus introduce errors. The same goes for the a priori class probabilities involved in computing the "constant $i$". This is a difficult problem if you have trouble collecting enough samples.

4) The outlines scheme seems to assume that all samples coming in for classification really belong to one of the classes $i$. But parameter vectors sufficiently far away from the mean vectors must be originating in some "artifacts" and do not belong to any of the classes. These "everything else" values can hardly be regarded as a new class, at least its distribution function is not very pleasant. (Although seldom discussed, this should be a problem in most pattern recognition problems.)

---

**Figure 25** A symbolic two-dimensional diagram of the situation in feature space. The problem is of course to discriminate the three different malaria types but most important is to discriminate malaria parasites from everything else. This is done by thresholding the three Mahalanobis distances. An optional reject zone is introduced if it is desired to refrain from making a decision in uncertain situations.
Some parasites are lost in the sequential decisions often due to too low contrast or overlapping with other highly contrasting objects (see Figure 24). Adding these gives 9% false negatives. The classifier can be tuned as desired. Halving the false negative rate roughly doubles the false positive rate. Rejecting 10% can reduce the false negative rate with 4%. If the routine is not used for individual diagnosis then sacrificing some recognition rate to gain speed might be a good tradeoff.

The processing time is typically 6 minutes for a standard area of 100 fields of view. With 100 cells/field this means 1 700 cells/minute but the processing time is of course very much dependent on the number of dark objects found. If nothing is found the processing rate is 2 500 cells/minute.

Discussion

The low misclassification rate and the high processing speed indicate that there are good chances to make a really competitive machine, especially if it was decided to construct a dedicated, optimized system for automatic screening of blood smears.
In spite of the 4 objections this classifier was used, mostly because it gives reasonably good results and it is difficult to find a really good alternative. The user is given the opportunity to tune the classifier to his own cost function and his readiness to tolerate rejects to get the missclassification rate down. This is accomplished by adjusting the threshold values for the Mahalanobis distance and adjusting the "constant₁"-values.

The classifier

\[ D_1(x) = (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + k_1; \ k_1 = \text{constant} \]

\[ 1 = 1, 2, 3 \]

ALL \( D_1 > \text{outer limit} \rightarrow \text{artifact} \)

ALL \( D_1 > \text{inner limit} \rightarrow \text{reject} \)

SELECT 1 FOR MIN \( D_1 \)

Figure 26 A reasonable cost matrix probably assigns lower risk/cost to false positives than to false negatives and lowest cost to confusions among different malaria parasite types. The classifier is to be adjusted to yield the lowest overall risk.

The results

Routine prepared slides with blood from apes and man mostly infected with a malaria type called P Brasilianum were directly used. Areas containing about 160 000 red blood cells were scanned. Half of the samples were used for designing the classifier (estimating means and covariance matrices) and the rest for testing. Assuming a slightly higher risk for false negatives, the results for objects reaching the final classification are:
While improving the system, the first thing to do would be to carefully select (and maybe replace) the features used. Since the classes are quite different maybe it would be a good point to have features tailored to every class to avoid the noisy results some features dedicated to one class produced on other classes. A lot of effort would have to be made to better recognize the very small ring-forms since this is a characteristic of the most dangerous type of malaria (P Falciparum).

In a dedicated system a better microscope and TV camera would be a good help. Knowing what to do it is easy to modify the PICAP processor for more suitable picture sizes and for faster processing. (Just replacing the present processor components with modern ones would make it 5 times faster.)

Since a lot of errors are introduced by having a high density of cells maybe it would be possible to slightly change the present slide preparation to get very thin smears. The processing time would not increase thanks to the sequential decision procedure and a fast moving stage.

Maybe the most important improvement would be to use the color information. The parasites are red and blue and many confusing objects are not. Using at least two color components would certainly increase the processing speed and lower the error rate.

Acknowledgement

The authors would like to express their gratitude to Dr Björn Kruse for constructive criticism and contributions throughout this work. The financial support from the Swedish Board of Technical Development is gratefully acknowledged.
References

[1] P.C.C. Garnham: Malaria Parasites and other Haemosporidia


