DETECTION AND COUNTING OF TB BACILLI ON ZIEHL NEELSEN SPUTUM SMEAR MICROSCOPY IMAGES

Manivas Kandukuri
M.Tech, IIT Guwahati

Contents

1. Introduction

For the development of automated detection techniques of TB bacilli on Ziehl-Neelsen sputum smear microscopic images, a standard sputum smear microscopic image database [1] is developed by Mohammad Imran Shah and the same can be found here. As per the database, a microscopic image consists of 4 different type of objects on each image, i.e., single bacilli, bacilli cluster, unclassified red structures and artifacts.

The overall scope of this project is to identify the potential bacilli object, classify it into either of the 3 classes i.e., single bacillus, bacilli cluster and artifacts, count the number of single bacilli in each bacilli cluster and hence calculate the total number of single bacilli on a given image.

Using the ground truth information available on the microscopic images, annotation is performed on each image and corresponding annotation file is generated for each image.

On the given input image, color space-based segmentation is applied to separate the potential TB objects from the background. Then post processing is performed on the segmented image to remove the smaller size artifacts. From the post processed image, contours are extracted from each object on the image and geometric features (like Relative convex area, Eccentricity, Roughness etc.,) are calculated for each contour. Using the ground truth information in the annotation file, each contour is assigned a class label. Similar procedure is applied for all the images and a final labelled data matrix is created using the afore-mentioned features and class details of all the images. Random forest classifier is developed by training it using the labelled data matrix. For a given test image, each potential TB object is classified into one of the 3 classes (i.e., single bacillus, bacilli cluster and artifacts) using Random forest classifier. Concave points are identified on the bacilli cluster and the same are used to count the number of single bacilli in each bacilli cluster. Hence, an overall count of single bacilli is obtained for a given test image.

2. Materials and Methods

2.1. Image Annotation

Using the ground truth information available on the microscopic image and the online image annotation tool, bounding box-based annotation is performed on each image to label each object into one of 4 classes, i.e., single bacillus, bacilli cluster, unclassified red structures and artifacts.

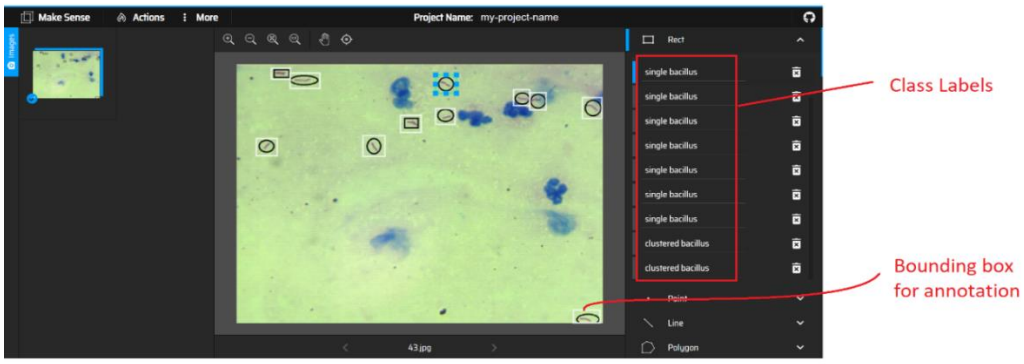


Fig. Screenshot of online image annotation tool

An annotation file (in .csv format) is generated for each image. The annotation file consists of details related to class label, coordinates of top left and bottom right corners of bounding box around each object etc., as shown below.

	A	B	C	D	E	F	G	H
1	single bacillus	43	172	45	40	43.jpg	800	600
2	single bacillus	275	165	48	48	43.jpg	800	600
3	single bacillus	434	24	45	43	43.jpg	800	600
4	single bacillus	434	102	43	37	43.jpg	800	600
5	single bacillus	755	77	43	48	43.jpg	800	600
6	single bacillus	604	58	35	37	43.jpg	800	600
7	single bacillus	641	65	35	40	43.jpg	800	600
8	single bacillus	735	570	58	30	43.jpg	800	600
9	single bacillus	116	20	65	35	43.jpg	800	600
10	clustered bacillus	357	118	45	40	43.jpg	800	600
11	clustered bacillus	75	9	41	26	43.jpg	800	600

Class label of each object

Top left coordinates of bounding boxes

Bottom right coordinates of bounding boxes

Image name

Image Resolution

2.2. Image Segmentation

As mentioned in reference [2], YCbCr and Lab color spaces offer better segmentation results. Specifically, the Cr plane (from the YCbCr image) conveys most of the information related with bacilli and bacilli clusters and rejects most of the artifacts. but it also contains artifacts that result from the varying illumination conditions. On the other side, the Lab color space, and specifically the a-plane, is more robust against illumination artifacts but is also unable to reject other objects, different than bacilli, present in the image. Since both planes, Cr and a, contain information related with the bacilli, but also both differ in the type of artifacts detected, it is possible to make a logical AND between both segmented images, in order to obtain the desired results. Histograms of the Cr-component and a-component are calculated. Threshold (for segmentation) is chosen for Cr-component and a-component based on the first derivative of their histograms as it is robust against varying illumination conditions. If *H[n]* (n = 0, 1, …. 254) is the histogram, then its first derivative is computed as:

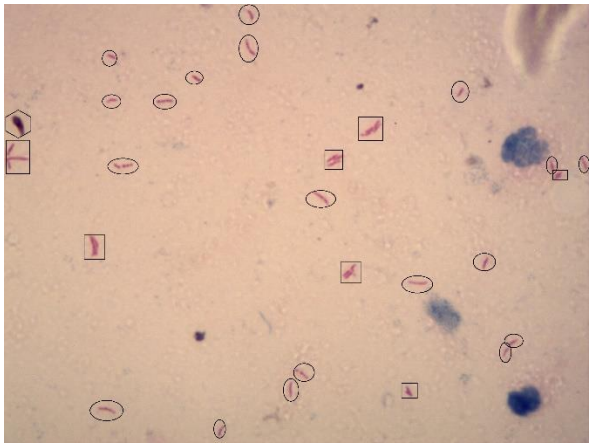
$$\Delta H[n] = H[n+1] - H[n]$$

The threshold is then selected as the maximum level of intensity (n_{max}) chosen between all the possible levels, n, that satisfy the condition:

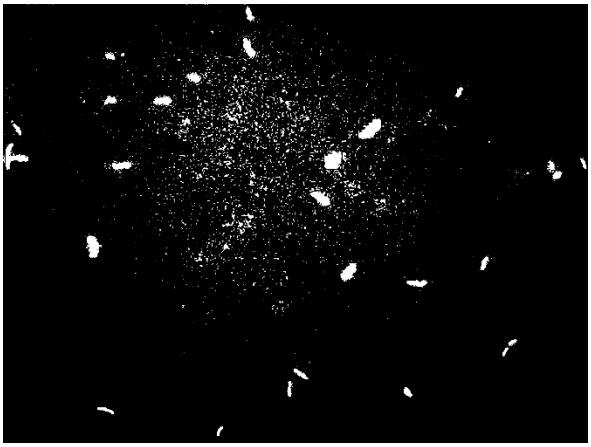
$$\Delta H[n] \leq r$$

Segmentation steps [2]:

A. The given RGB microscopic image is transformed into YCbCr and CIE-Lab spaces.
B. Extracted the Cr-component from the YCbCr image and a-component from the Lab image.
C. Computed threshold levels for the Cr-component and a-component based on their first derivatives.
D. Computed segmented images for Cr-component and a-component based on the thresholds obtained in Step 3.
E. Computed a logical AND between the segmented Cr-component and segmented a-component to get the final segmented image.



(a)Original image



(b)YCbCr segmented image

(c)Lab segmented image

(d)Final segmented image

2.3. Image Post-processing

Along with the true bacilli, the original microscopic image consists of a lot of non-bacillus objects (also called artifacts) having similar color characteristics as that of true bacilli. Due to this, these artifacts are even observed on the segmented image. Also, the count of these artifacts is very high when compared to the count of true bacillus. Removing these artifacts is required to improve the speed of image processing performed in further stages.

The typical size of these artifacts is very small when compared to that of a true bacillus and this fact can be used to remove all these smaller size artifacts.

The segmented image is resized to 600x800. This helps to choose a proper size threshold to reject the small-sized artifacts. After observing many images, it is found that objects with size less than 20 are artifacts. Using morphological operations, objects less than size 20 are removed.

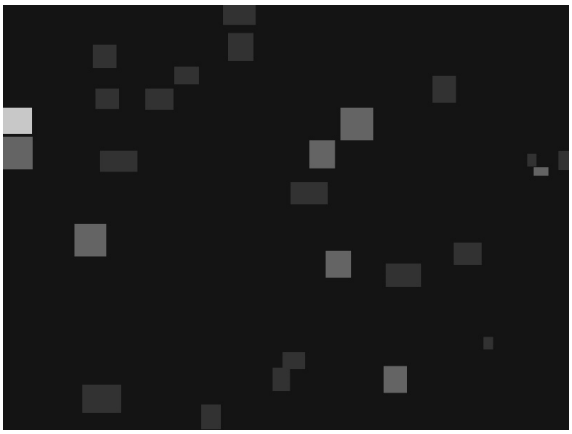
(a)Image after segmentation

(b)Image after post processing

2.4. Extracting the ground truth info and labelling the objects

A gray scale image (with all pixel intensities initialized to zeros) is created. Using the details of bounding boxes available in the annotation file, bounding rectangles are drawn on this image at the location of potential TB objects with different intensities (with each intensity corresponding to a class label) to get the image with ground truth information as shown below.

(a)Postprocessed image



(b)Image with ground truth information

A logical AND is performed between post processed image and the image with ground truth info to get the labelled processed image (image with labelled objects) as shown below.

(c)Image with labelled objects

2.5. Contour finding, feature extraction and formation of labelled data matrix

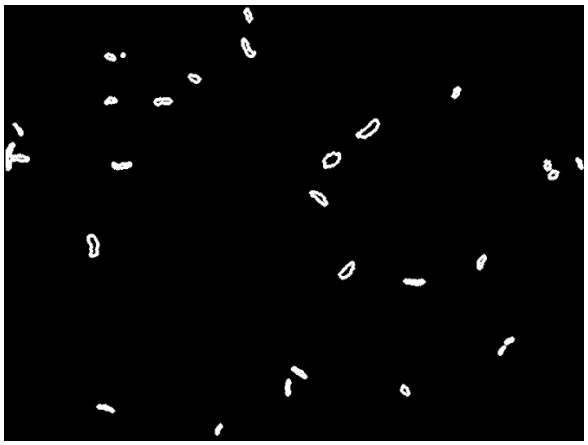
For each object on the labelled post-processed image, corresponding contour is found and contour properties (i.e., Major axis length, minor axis length, contour area, contour perimeter, perimeter of convex hull, area of convex hull) are obtained. From the contour properties, geometric features [3][4][5] like Area, Roughness, Relative Convex area, Circularity, Compactness, Eccentricity are calculated using the below formulae.

(a) Area = Area of contour
(b) Roughness = Perimeter of contour/Perimeter of convex hull of contour
(c) Relative convex area = Area of convex hull of contour/Area of contour
(d) Circularity = 4*pi*area of contour/ (Perimeter of contour) ^2
(e) Compactness = (Perimeter of contour) ^2/ (4*pi*Area of contour)

Class label is assigned to each object based on the intensity of corresponding labelled object on Labelled post processed image. A feature vector along with its class label is created for each object in

this way. The feature vector along with the class label of all the objects on the labelled postprocessed image are stacked horizontally (row-wise) to get the labelled data matrix for the entire image.

(a)Image with labelled objects



(b)Image with contours

	Area	Roughness	Relative convex area	Circularity	Compactness	Eccentricity	Class
0	31.5	1.068810	1.380952	0.369559	2.705929	0.321105	1
1	63.5	1.076540	1.448819	0.285472	3.502966	0.244280	1
2	77.0	1.070369	1.110390	0.648499	1.542023	0.571713	2
3	60.0	1.084323	1.450000	0.314406	3.180600	0.252521	1
4	82.5	1.090662	1.272727	0.359534	2.781379	0.233383	1
5	30.0	1.041201	1.183333	0.476011	2.100792	0.312058	1
6	38.0	1.066908	1.223684	0.497846	2.008652	0.353477	1
7	106.0	1.076760	1.259434	0.330805	3.022926	0.229078	1
8	200.5	1.139657	1.182045	0.483554	2.068020	0.392856	2
9	85.0	1.128010	1.217647	0.416732	2.399621	0.341955	1
10	250.0	1.099904	1.204000	0.501947	1.992244	0.358375	2
11	155.5	1.191284	1.270096	0.358385	2.790297	0.325612	2
12	79.0	1.117282	1.164557	0.591416	1.690857	0.691752	2

(c)Labelled data matrix of single image

The above process is repeated for all the training images to get their corresponding labelled data matrices and all these labelled data matrices are stacked horizontally to obtain the final labelled data matrix for the entire image dataset.

2.6. Building a Machine Learning Model

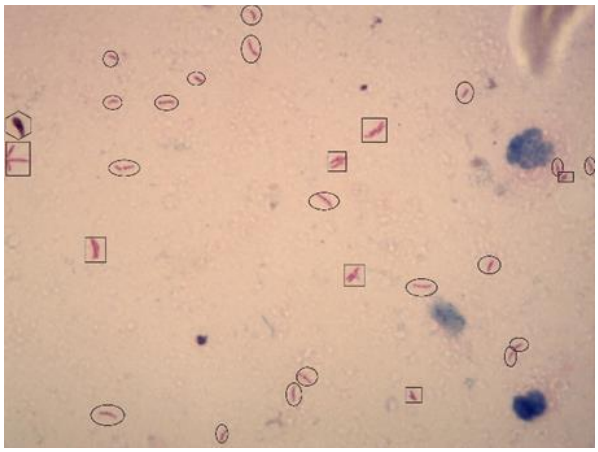
A Random Forest based Machine learning model is trained using the labelled data matrix. This ML model can be used to classify the potential bacillus objects on a given test image into either of 3 classes, i.e., single bacillus, bacilli cluster and artifacts.

2.7. Detection of actual bacilli objects on a test image

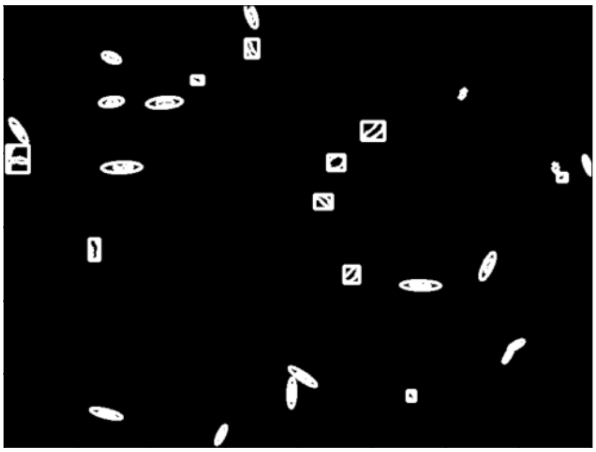
To detect the actual bacilli objects on a test image, the following steps are performed on it to get the feature vectors of all potential bacillus objects.

A. Image segmentation
B. Image post-processing
C. Contour finding and feature extraction

The feature vectors are classified into either of 3 classes, i.e., single bacillus, bacilli cluster and artifacts using the Random forest classifier mentioned above. The classification result for a given test image are shown below. In the classification result shown below, objects with oval shaped bounding boxes represent single bacillus, those with rectangular bounding box represent bacilli cluster and those without any bounding box represent artifact.



(a)Original image



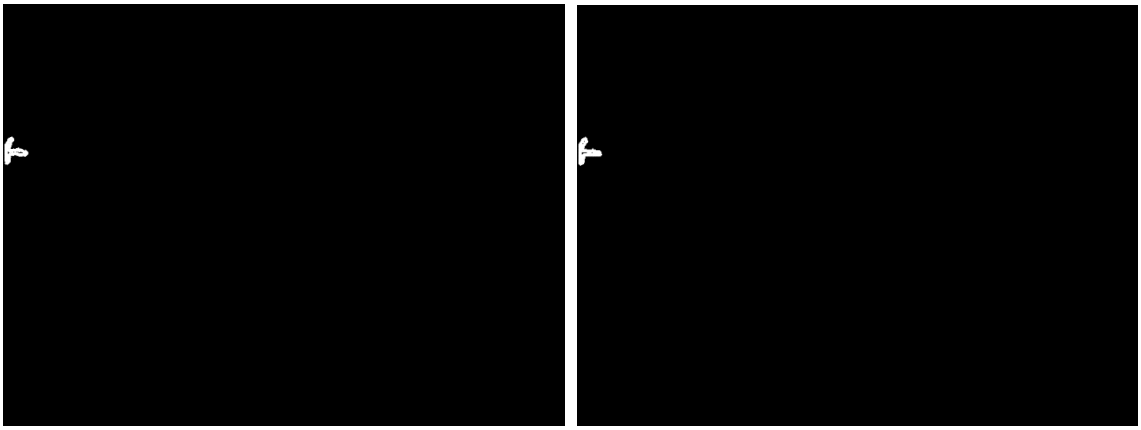
(b)Classification result

2.8. Separation of clustered bacilli objects

Separation of clustered bacilli objects is based on concave points and ellipse fitting [3]. The method includes four parts they are: 1) Polygon Approximation, 2) Concave point extraction, 3) Contour segmentation, and 4) Ellipse processing.

2.8.1. Polygon Approximation

The original contour of clustered bacillus may be rough. So, Polygon approximation (PA) of the contour is necessary to smoothen the irregular rising and falling of overlapping contour. Moreover, Polygon Approximation sufficiently reduces the number of points in the contour boundary and thereby reducing the calculation time in the immediate phases.



(a)Contour of clustered bacillus *(b)PA contour of clustered bacillus*

2.8.2. Concave point extraction

Concave points (or convexity defects) are extracted from the polygon approximated contour which are used for segmenting the contour.

(a)Contour with detected concave points (concave points are marked in red)

2.8.3. Contour segmentation

Concave points identified from the approximated contour are used to divide the approximated contour into number of segments. If '*C*' is a contour then, *C* can be represented as:

$$C=L_1+L_2+.....+L_m$$

where '*m*' is the total number of concave points. Each L_i is a segment on *C* having any two adjacent concave points as end point. Now each of these segments can be used for ellipse fitting.

2.8.4. Ellipse fitting and refinement

Each contour segment is fitted with an ellipse and eccentricity and area are calculated for each of these contour segments. Contour segments with area above a certain threshold value and eccentricity lying between a chosen minimum and maximum threshold values are only considered as valid single bacillus and others are ignored.

(a)Contour seg No.1 of bacilli cluster

(b)Ellipse fitting on Contour seg No.1

(c)Contour seg No.2 of bacilli cluster

(d)Ellipse fitting on Contour seg No.2

(e)Contour seg No.3 of bacilli cluster

(f) Ellipse fitting on Contour seg No.3

(g)Contour seg No.4 of bacilli cluster

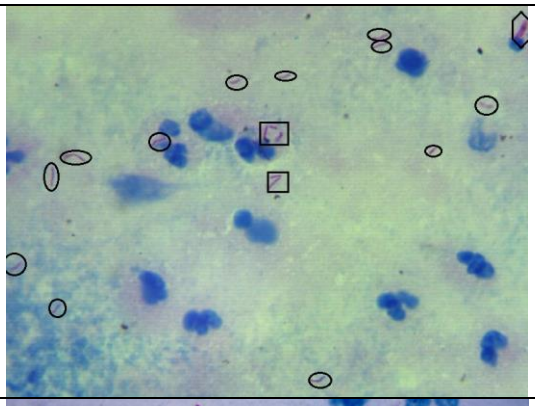
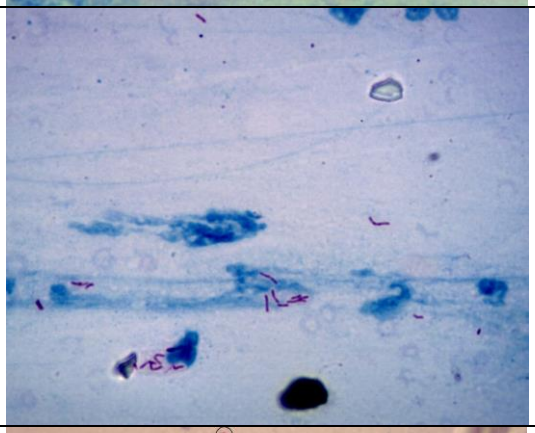
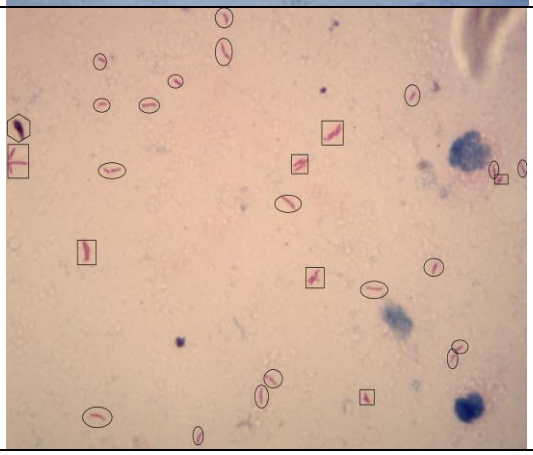
(h) Ellipse fitting on Contour seg No.4

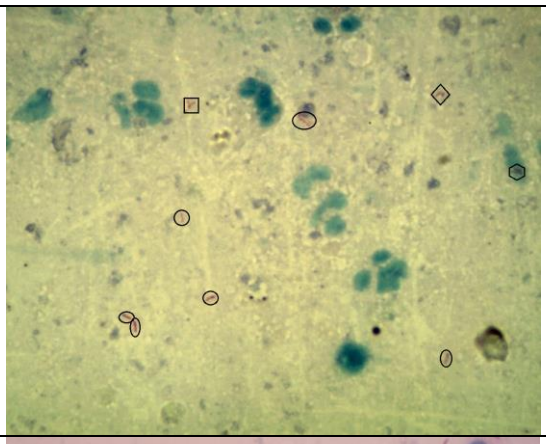
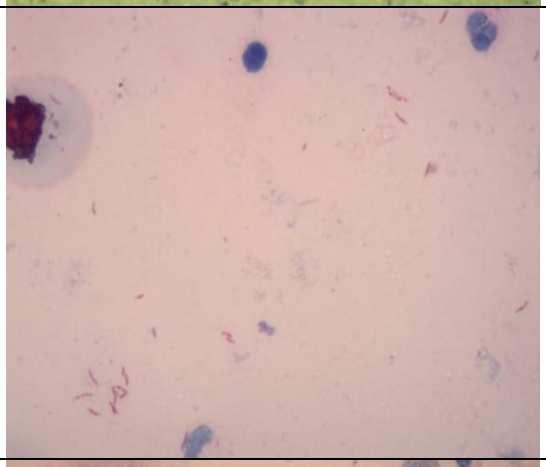
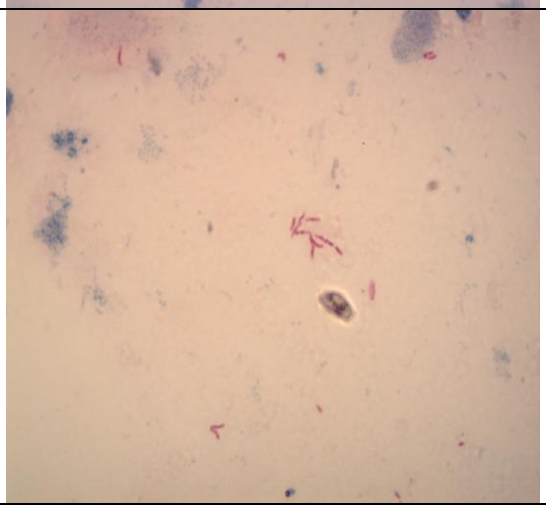
(i)Contour seg No.4 of bacilli cluster

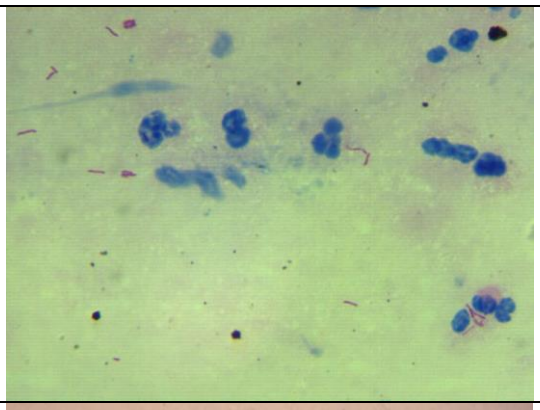
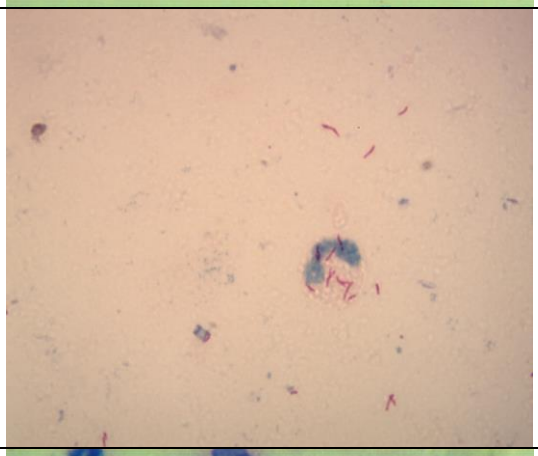
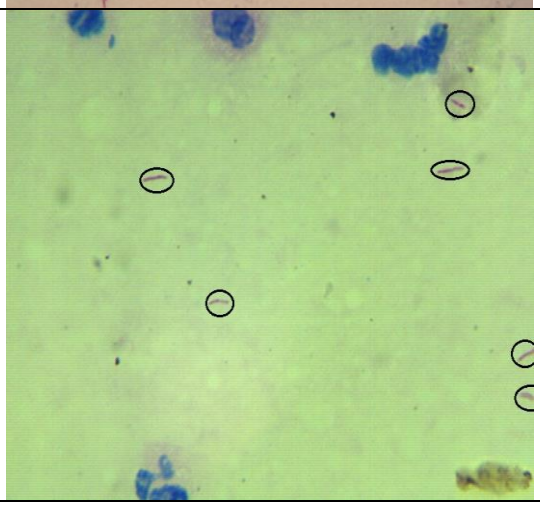
(j) Ellipse fitting on Contour seg No.4

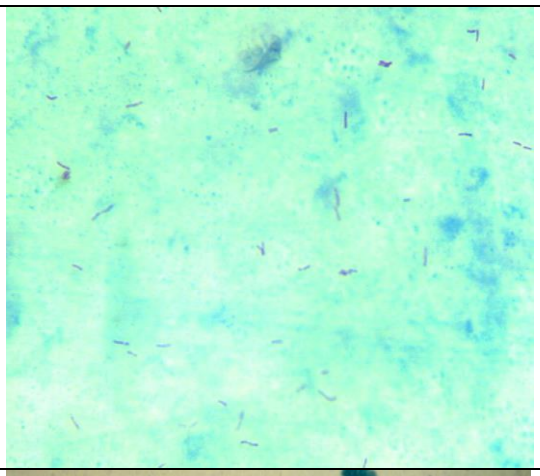
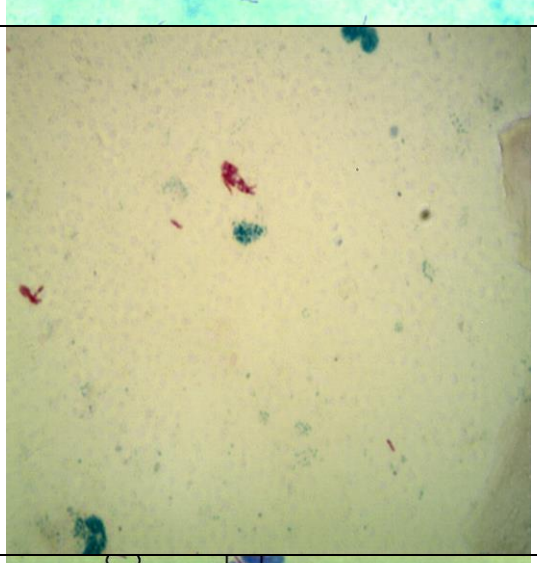
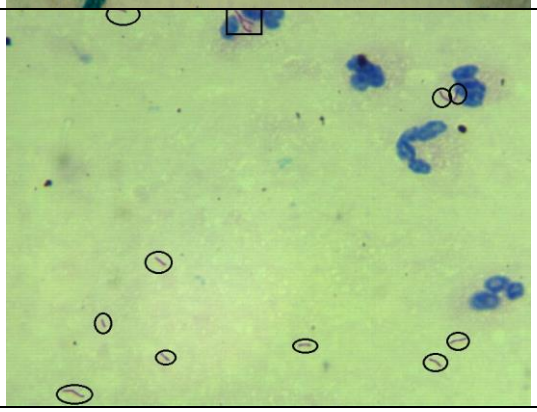
3. Results:

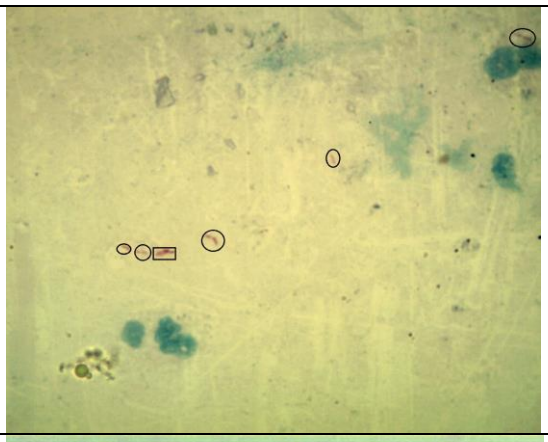
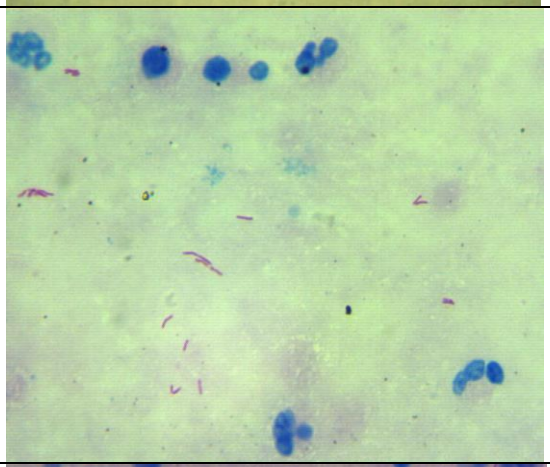
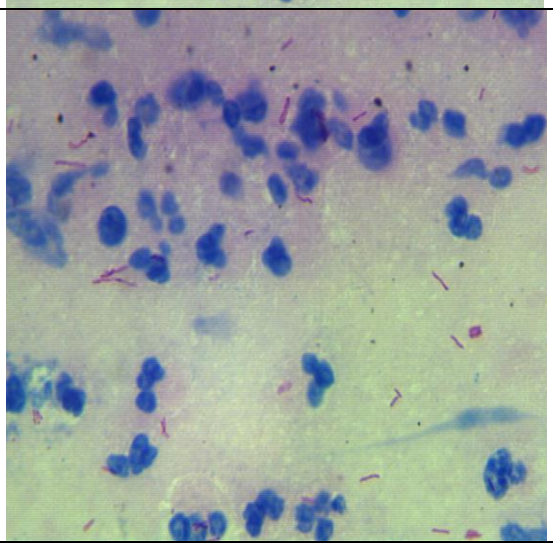
The performance of the system can be found from the below test results by checking the closeness of values of actual bacilli count with estimated bacilli count.

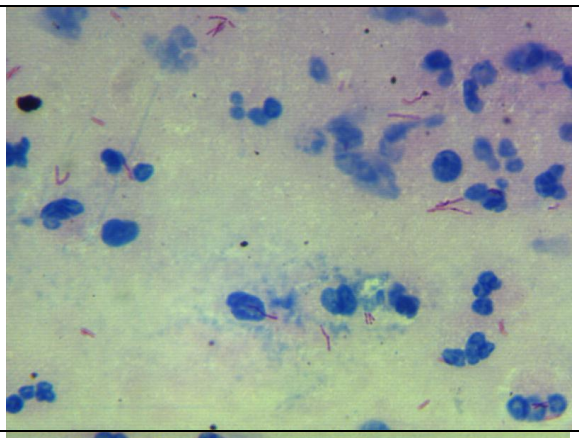
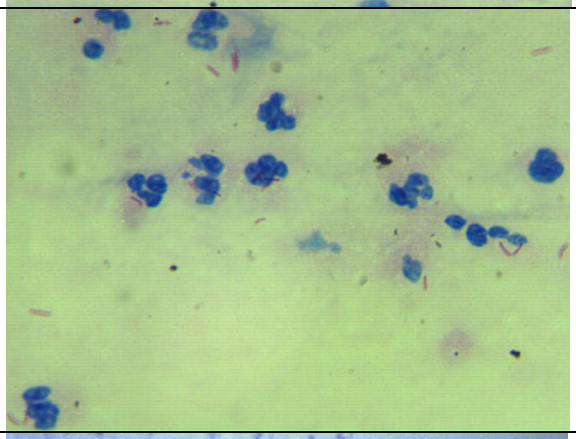
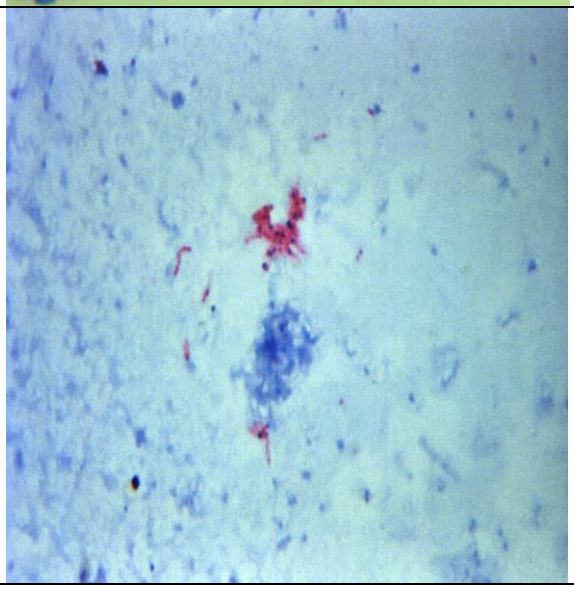
S.No.	Test image	Actual bacilli count	Bacilli count estimated by system
1.		19	25
2.		18	23
3.		34	31

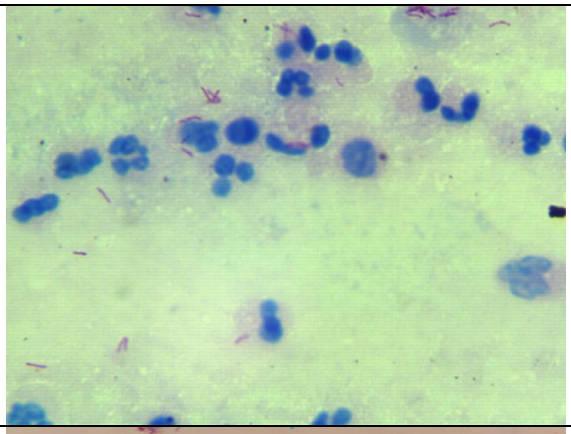
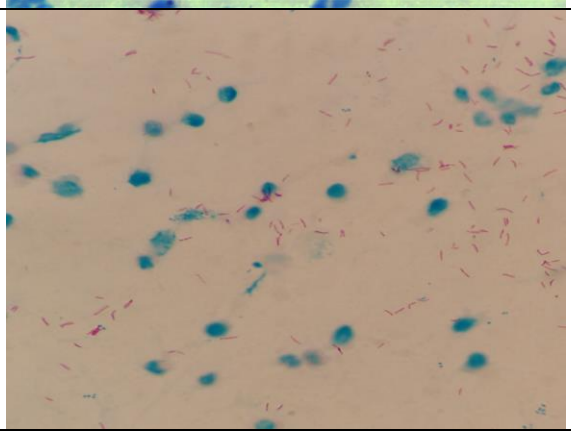
4.		8	6
5.		22	23
6.		16	17

7.		17	32
8.		16	15
9.		6	9

10.		34	30
11.		14	11
12.		13	12

13.		7	6
14.		17	26
15.		25	21

16.		22	20
17.		13	31
18.		6	5

19.		50	43
20.		23	36
21.		90	80

From the above results, the estimated bacilli count is almost closely matching with that of actual ones for most of the images. For the images with significant amount of background stains and artifacts, the difference in bacilli count is huge. The main reasons for mismatch in the bacilli count are the following:

(a)some of the stains/artifacts are getting classified as true bacillus and vice versa by the ML classifier and (b)due to error in counting of bacilli in bacilli cluster.

4. Conclusion

In this project, a method is formulated to classify the bacillus as single bacillus or clustered bacillus or artifact using a Random Forest classifier. The predicted clustered bacilli are separated by segmenting based on concave points and filtering the contour segments through ellipse fitting. And finally, the total number of single bacillus for a given image are estimated.

5. References

[1] Mohammad Imran Shah, Smriti Mishra, Vinod Kumar Yadav, Arun Chauhan, Malay Sarkar, Sudarshan K. Sharma, Chittaranjan Rout, "Ziehl–Neelsen sputum smear microscopy image database: a resource to facilitate automated bacilli detection for tuberculosis diagnosis", J. Med. Imag. 4(2), 027503 (2017), doi: 10.1117/1.JMI.4.2.027503.

[2] Sotaquir´a, M., Rueda, L. and Narvaez, R. "Detection and quantification of bacilli and clusters present in sputum smear samples: a novel algorithm for pulmonary tuberculosis diagnosis". International Conference on Digital Image Processing, 2009.

[3] Reshma S R, Rehannara Beegum T. "Microscope Image Processing for TB Diagnosis Using Shape Features and Ellipse Fitting". IEEE SPICES 2017 1570365912.

[4] Ebenezer Priya a, Subramanian Srinivasan. "Separation of overlapping bacilli in microscopic digital TB images". biocybernetics and biomedical engineering 35 (2015) 87 – 99, ScienceDirect.

[5] Shape Analysis and Measurement by Michael A. Wirth, University of Guelph Computing and Information Science Image Processing Group.