Journal of Theoretical and Applied Information Technology
10th May 2014. Vol. 63 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PRINCIPAL COMPONENT ANALYSIS COMBINED WITH SECOND ORDER STATISTICAL FEATURE METHOD FOR MALARIA PARASITES CLASSIFICATION

[1,3]IIS HAMSIR AYUB WAHAB, [1]ADHI SUSANTO, [1]P. INSAP SANTOSA,
[2]MAESADJI TJOKRONEGORO

[1]Department of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta
[2]Faculty of Medicine, Universitas Gadjah Mada, Yogyakarta, Indonesia
[3]Department of Electrical Engineering, Universitas Khairun, Ternate, Indonesia
E-mail: [1]iishamsiraw_s3_09@mail.ugm.ac.id

ABSTRACT

The main challenge in detecting malaria parasites is how to identify the subset of relevant features. The objective of this study was to identify a subset of features that are most predictive of malaria parasites using second-order statistical features and principal component analysis methods. Relevant features will provide the successful implementation of the overall detection modeling, which will reduce the computational and analytical efforts. The results showed that the combination of the principal components of the feature value the correlation to the ASM, and the contrast to the correlation can separate classes of malaria parasites.

Keywords: *Features, PCA, Identify, Classification, Malaria*

1. INTRODUCTION

Classification is the process of grouping objects into classes according to their respective characteristics. Classifiers' using the feature vector has been extracted in the previous stage. The degree of difficulty in classification depends on the value of diversity characteristics for objects the same class, relative to the difference between the values characteristic for objects in a different class. Diversity characteristic value for objects in the same class can be caused by the level of complexity and can also be caused by noise. Noise can be defined as a trait that is not caused by the pattern of that class but due to randomness or because of the nature of the sensor.

In a broad sense, any method that involves information from samples in designing classifier training is learning by doing. Classifiers can be created by using common models and classifiers using training patterns to learn or estimate the parameters of the model are unknown. Learning refers to some form of algorithm to reduce the error on the training data set. .

The main aim of the research is to analyze the types of malaria parasite by applying the measurement of second–order statistical features from images, and principal component analysis method.

2. PREVIOUS WORK

Automatic detection of malaria parasites has been conducted by several researchers in which the number of malaria parasites in the blood can be calculated based on the results of blood sample acquisition is converted into a digital image [3-19]. Detection systems are generally constructed through several stages of processing, i.e. image acquisition, image pre-processing, image segmentation, feature extraction and classification.

Selena, et al. [3], developed an automated system to calculate the malaria parasite to quickly and accurately based image analysis. There are four stages that can be done to calculate the number of parasetimia, namely: edge detection, edge linking, splitting the clump and parasite detection.

Parasites can be identified by their characteristic properties from the edge of the red blood cells using Sobel operator and the Hough Transform [4][6]. This is also done by Diaz et al. [5], but on a system built using a low pass filter on the image preprocessing, then extract the image pre-processing to obtain information on where blood cells are infected with malaria parasites.

In addition, another way to detect malaria parasites from red blood cells in the image is to make the process of segmentation. Segmentation is intended to be split among red blood cells and the parasite itself. Wahab et al. [7] make the process of segmentation using k-mean method. While Makkapati and Rao [13] perform color segmentation based on HSV color space. The image will be segmented first filtered using a median filter with window size 3 x 3 to eliminate the noise. Another segmentation method that is used to detect malaria parasites in the image are normalized cut method (NCut) [12]. As with other methods, the method NCut is not unsupervised segmentation method based on global criteria. In this method, the color space used is RGB, HSV, YCbCr and NTSC.

Detection of malaria parasites can be improved performance-works when using the strategy machine learning [14]. The system built is a classifier system based on the input characteristics of each type of parasite. Feature of that has been extracted from previous experiments is a histogram feature [5-11, 14, 19], morphological [10-13], texture [15, 18]. The classifier system used is a neural network multilayer perceptron [5], support vector mechine [5, 19], k-nn [14], Bayesian pixel [8], learning vector quantization [7-9], fuzzy learning vector quantization [10,11] and back propagation [18, 19].

3. MATERIALS AND METHODS

3.1 Materials

In this research, the data samples used were obtained from blood samples of patients on glass microscope slides. Then preparations in the form of a digital image file format *. Jpg. The files have 24-bit color depth and size of 256x256 pixels. The number of data is 96 with the number of class 4, according to the type of variation in the condition of Plasmodium.

3.2 Second-Order Statistical Feature Extraction

Second-order statistics feature extraction is done by co-occurrence matrix, i.e. a matrix that represents the relationship between the adjacency between pixels in the image at different orientations and directions of spatial distance. One technique for obtaining second-order statistical characteristics is to calculate the probability of adjacency relations between two pixels at a certain distance and angular orientation [20]. This approach works by forming a matrix co-occurrence of image data, followed by determining the characteristics as a function of the matrix between them.

Co-occurrence means joint events, i.e. the number of occurrences of the level of neighboring pixel values with one another level pixel values in the distance (d) and orientation angle (θ) specific. Distance expressed in pixels and orientation is expressed in degrees. Formed in a four-way orientation angle with an angle of 45° intervals, i.e. 0°, 45°, 90°, and 135°. The distance between pixels is usually set at 1 pixel.

Non-normalized frequencies of co-occurrence matrix as functions of distance, d and angle 0°, 45°, 90° and 135° can be represented respectively as

$$\mathrm{H}_{0^{\circ},d}(Im_1,Im_2)=\left|\left\{\begin{matrix}[k,l),(m,n)]\in\boldsymbol{D}:\\ k-m=0,|l-n|=d\\ f(k,l)=Im_1,f(m.n)=Im_2\end{matrix}\right\}\right| \quad (1)$$

$$\mathrm{H}_{45^{\circ},d}(Im_1,Im_2)=\left|\left\{\begin{matrix}[k,l),(m,n)]\in\boldsymbol{D}:\\ (k-m=d,l-n=-d)\lor\\ (k-m=d,l-n=d),\\ f(k,l)=Im_1,f(m.n)=Im_2\end{matrix}\right\}\right| \quad (2)$$

$$\mathrm{H}_{90^{\circ},d}(Im_1,Im_2)=\left|\left\{\begin{matrix}[k,l),(m,n)]\in\boldsymbol{D}:\\ k-m=d,|l-n|=0\\ f(k,l)=Im_1,f(m.n)=Im_2\end{matrix}\right\}\right| \quad (3)$$

$$\mathrm{H}_{90^{\circ},d}(Im_1,Im_2)=\left|\left\{\begin{matrix}[k,l),(m,n)]\in\boldsymbol{D}:\\ (k-m=d,l-n=d)\lor\\ (k-m=-d,l-n=-d),\\ f(k,l)=Im_1,f(m.n)=Im_2\end{matrix}\right\}\right| \quad (4)$$

where $|\{..\}|$ refers to cardinality of set, *f*(*k*, *l*) is intensity at pixel position (*k*, *l*) in the image of order (MxN) and the order of matrix ***D*** is (MxN) x (MxN).

Haralick [21] proposed various types of textural characteristics can be extracted from the co-occurrence matrix. The Angular Second Moment (ASM), Contrast (Cont), Correlation (Cor), Variance (Var), Inverse Difference Moment (IDM), and Entropy are few such measures which are given by:

$$\text{ASM} = \textstyle\sum_{i,j} H(Im_1,Im_2)^2 \quad (5)$$

$$\text{Cont} = \textstyle\sum_{I_{m1},I_{m2}} |Im_1,Im_2|^2 \log H(Im_1,Im_2) \quad (6)$$

$$\text{Corr} = \textstyle\sum_{I_{m1},I_{m2}} \frac{(Im_1-\mu_1)(Im_2-\mu_2)H(Im_1,Im_2)}{\sigma_1\sigma_2} \quad (7)$$

$$\text{Var} = \textstyle\sum_i\sum_j \big(Im_1-\mu_1)(Im_2-\mu_2\big)H(Im_1,Im_2) \quad (8)$$

$$\text{IDM} = \textstyle\sum_i\sum_j \frac{1}{1+(i-j)^2} H\big(Im_1,Im_2\big) \quad (9)$$

$$\text{Entropy} = -\textstyle\sum_{I_{m1},I_{m2}} H(Im_1,Im_2)\log H(Im_1,Im_2) \quad (10)$$

ASM is a feature that measures the smoothness of the image. Contrast is a measure of local level variations which takes high values for image of high contrast. Correlation is a measure of correlation between pixels in two different directions. IDM is a measure that takes high values for low-contrast images. Entropy is a measure of randomness and takes low values for smooth images. Together all these features provide high discriminative power to distinguish two different kind of images.

All features are functions of the distance d and the orientation θ. Thus, if an image is rotated, the values of the features will be different. In practice, for each d the resulting values for the four directions are averaged out. This will generate features that will be rotations invariant.

3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a well-known multivariate statistical technique, that converts a larger number of observed variables into a smaller number of variable (called principal components), that will take in account most of the variance in the observed variables[22]. Number of original variables is more than or equal to the number of principal components. In this technique, the first principal component has the highest variance; the second principal component has the second highest variance and so on. But each principal component should be orthogonal to the one before that.

PCA can helps to identify pattern in data and emphasize on the similarities and differences of the dataset. When the data had a large number of variables, patterns in the dataset are difficult to be recognized. So PCA is a useful tool for analyzing large dimensional data[23].

In this study, PCA is used to process the raw data the 2nd-order statistical characteristics of the malaria parasite. Data Normalization aims to see whether the object of plasmodium falciparum, vivax, ovale and malariae may be separate groups. The variable dataset as 6, it is derived from a second order statistical feature. So the data has 6 dimensions. So it is difficult to visualize and classify the data. Applying PCA score plot will provide primary, which will be two-dimensional. Clusters in the Principal scores plot will show spectral similarities and will allow us to classify the type of plasmodium falciparum, vivax, ovale or malariae. Steps of the PCA process are:

Step 1: Getting some data. In this research, four commands (load falciparum.dat, vivax.dat, ovale.dat and malariae.dat) were applied. When using these sorts of matrix techniques in computer vision, representation of image must be well considered. A square, N by N image can be expressed as an. N2 dimensional vector:

$$X = (x_1, x_2, \ldots x_n) \qquad (11)$$

where the rows of pixels in the image are placed one after the other to form a one-dimensional image.

Step 2: Subtracting the raw data with mean of the all data. The equation of the mean data was used as follows:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (12)$$

Step 3: Calculating the covariance matrix Covariance is such a measure and always be measured between 2 dimensions. The formula for covariance is quite similar to the formula for variance. The formula for covariance could also be written like this:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)} \qquad (13)$$

Step 4: Calculating the eigenvectors and eigenvalues of the covariance matrix. Let P has coordinates (x, y) relative to the x, y axes and coordinates (x_1, y_1) relative to the x_1, y_1 axes.

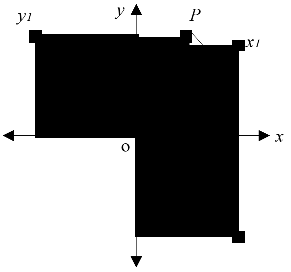


Figure 1: Rotating the axes

$$\begin{aligned} x = OQ &= OP \cos(\phi + \alpha) \\ &= OP(\cos\phi \cos\alpha - \sin\phi \sin\alpha) \\ &= (OP\cos\phi)\cos\alpha - (OP\sin\phi)\sin\alpha \\ &= OR\cos\phi - PR\sin\phi \\ &= x_1\cos\phi - y_1\sin\phi \qquad (14) \end{aligned}$$

Similarly $y = x_1 \sin\phi + y_1 \cos\phi$ (15)

These transformation equations could be combined into a single matrix equation:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\emptyset & -\sin\emptyset \\ \sin\emptyset & \cos\emptyset \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \qquad (16)$$

or

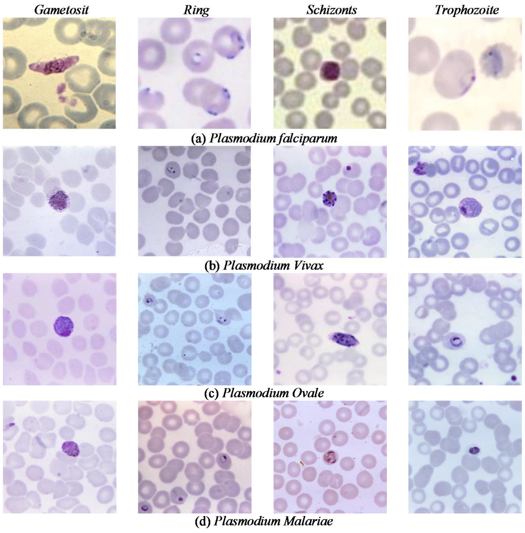
$$X = PY \qquad (17)$$

where X = $\begin{bmatrix} x \\ y \end{bmatrix}$, Y = $\begin{bmatrix} x \\ y \end{bmatrix}$ and P = $\begin{bmatrix} \cos\emptyset & -\sin\emptyset \\ \sin\emptyset & \cos\emptyset \end{bmatrix}$

A matrix of the type P is called a rotation matrix. It will be shown soon that any 2×2 real orthogonal matrixes with determinant equal to 1 is a rotation matrix.

4. RESULTS

The data in this study using two data samples, i.e. samples derived from data bank of parasitic image[2] and immediate retrieval of samples. The process of collecting samples of blood preparations directly assisted by paramedics and laboratory staff of health centers. Figure 2 is an image containing blood samples for malaria parasites and converted to gray scale image. From each image of the parasites obtained values of second-order statistical feature such as Angular Second Moment (ASM), Contrast, Correlation, Variance, Inverse Difference Moment (IDM), and Entropy. Table 1 show of the average values of it.



(a) *Plasmodium falciparum*

(b) *Plasmodium Vivax*

(c) *Plasmodium Ovale*

(d) *Plasmodium Malariae*

Figure 2: Image of Malaria Parasites[1][2]

Tabel 1: The Average Values of Second Order Statistical Feature of Malaria Parasite

Type of Plasmodium	ASM	CON	COR	VAR	IDM	ENT
falciparum:						
a. Gametosit	0.004	13.699	0.986	537.667	0.449	8.832
b. Ring	0.005	11.543	0.984	455.433	0.501	8.971
c. Schizont	0.002	14.680	0.973	263.497	0.367	9.638
d. Tropozoit	0.025	23.103	0.967	428.963	0.462	8.712
vivax:						
a. Gametosit	0.007	8.516	0.981	280.457	0.541	8.450
b. Ring	0.026	10.956	0.983	406.675	0.493	8.758
c. Schizont	0.008	12.752	0.985	501.422	0.453	9.341
d. Tropozoit	0.014	18.629	0.978	589.565	0.415	9.309
ovale:						
a. Gametosit	0.012	35.953	0.945	461.985	0.455	8.577
b. Ring	0.029	31.266	0.911	217.540	0.450	8.394
c. Schizont	0.010	24.827	0.959	318.206	0.482	8.585
d. Tropozoit	0.010	4.151	0.988	164.378	0.614	7.807
malariae:						
a. Gametosit	0.007	22.864	0.973	434.185	0.281	10.25
b. Ring	0.001	12.155	0.987	487.713	0.380	9.887
c. Schizont	0.002	21.664	0.988	946.687	0.328	9.79
d. Tropozoit	0.004	6.027	0.990	329.637	0.490	8.791

Figure 3 show plots of raw data in six dimensions which is measured from second order statistical features. This Figure show the randomized input of four kinds of malaria parasite. So very difficult to classified.

To make PCA works properly, the mean must be subtracted from each of the data dimensions. The mean subtracted is the average across each dimension. Hence, all of the values have x (the mean of the x values of all the data points) subtracted, and all the y values have y subtracted from them. This produces a data set whose mean is zero.

The data have been normalized to zero and then do the calculations to obtain the value of its covariance matrix that would be obtained eigenvalues and eigenvector. Eigenvector calculation results are then used as a multiplier to determine the direction and the axis of the main components of the new data obtained from the 2x2 transformation matrix angle order for the two inputs that produce the characteristic two-dimensional plot. Plots of the PCA result showed in Figure 4.

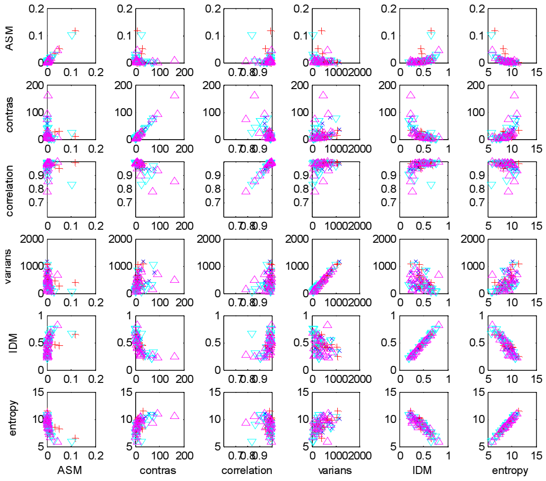
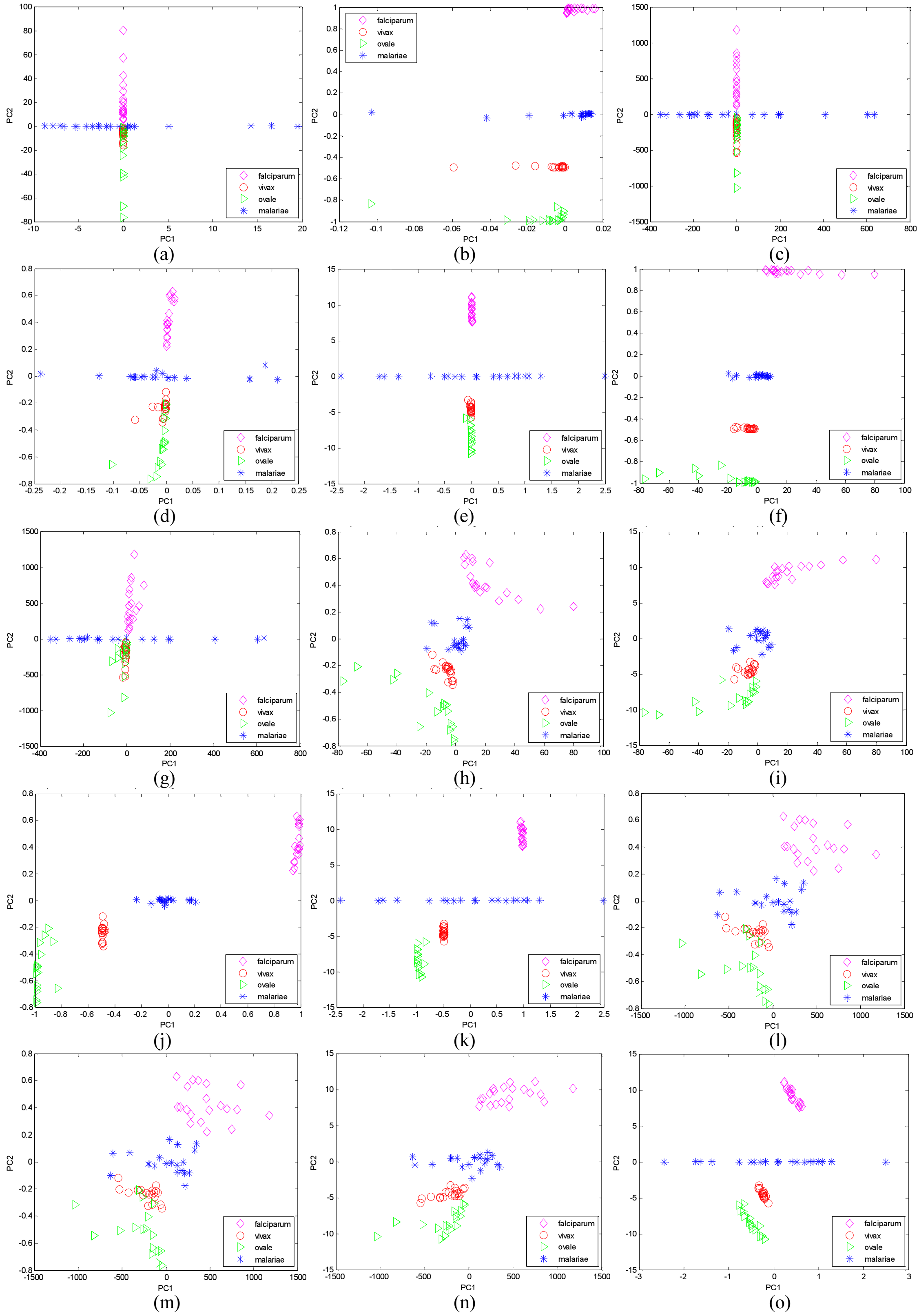


Figure 3: Plots of Raw Data in Six Dimensions

Figure 4 (a) show plot of the results of the principal component (PC) ASM on raw data value of contrast throughout the plasmodium falciparum, plasmodium vivax, plasmodium ovale and plasmodium malaria have not been able to provide significant separation for each class in the malaria parasite. It is also the case for the value of PC of the ASM to the variance (figure 4 (c)), the ASM to the IDM (Figure 4 (d)), the entropy to the ASM (Figure 4 (e)) and the contrast to the variance (Figure 4 (g)). The separations for each class of decoupling results were also not significant.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Figure 4: Plots of The Pca Result

Grouping of class better than the results of the plot happens to characterize the PC value of the contras to the IDM (Figure 4 (h)). This result also was synonymous with the entropy feature to the contras (Figure 4 (i)), the correlation to the entropy (Figure 4 (l)), the variant to the IDM (Figure 4 (m)) and the variance to the entropy (Figure 4 (n)). In this grouping, class plasmodium falciparum, plasmodium malariae already well separated, but the plasmodium vivax and plasmodium class ovale overlap still occurs. So when do further training will possibly lead to errors in classification.

Figure 4 (j), (k) and (o) show the results of the classification are relatively more assertive in classifying the class of the malaria parasite. However, the PC value of the correlation of the variance, the correlation to the IDM, and the IDM to the entropy still allows the occurrence of errors in classification. This is due to plasmodium falciparum class members tending to put a zero in his class at the axis y axis. Thus there will be a cross member of the class.

The best results of classification using PCA is to a major component of the feature of the correlation to the ASM (Figure 4 (b)) and the contras to the correlation (Figure 4 (f)). From the graph it was clear that each class of malaria parasites can be grouped on the x axis. Thus there will be no class member malaria parasite mutually disjoint or overlapping.

5. CONCLUSION

The experimental results show that second order statistical measurement combined with principal component analysis method is promising to distinguish the types of malaria parasite images. The best combination of the principal components of the feature value the correlation to the ASM, and the contrast to the correlation can separate classes of malaria parasites. Further experiments should be coupled with spectral and structural methods for analysis.

REFRENCES:

[1] http://www.rph.wa.gov.au/labs/haem/malaria

[2] http://www.dpd.cdc.gov

[3] Sio, Selena W.S., Weiling Sun, Saravana Kumar, Wong Zeng Bin, Soon Shan Tan, Sim Heng Ong, Haruhisa Kikuchi, Yoshiteru Oshima, Kevin S.W. Tan. “MalariaCount: An image analysis-based program for the accurate determination of parasitemia”. Elsevier. 2006

[4] Minh-Tam Le, Timo R Bretschneider, Claudia Kuss and Peter R Preiser. “A novel semi-automatic image processing approach to determine Plasmodium falciparum parasitemia in Giemsa-stained thin blood smears”. BMC Cell Biology 2008, 9:15 doi:10.1186/1471-2121-9-15. 2008.

[5] Gloria Díaz, Fabio A. González, Eduardo Romero. “A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images”. Journal of Biomedical Informatics 42 (2009) 296–307. 2009.

[6] Charles Ma, Paul Harrison, Lina Wang, Ross L Coppel. “Automated estimation of parasitaemia of Plasmodium yoelii-infected mice by digital image analysis of Giemsa-stained thin blood smears”. Malaria Journal, 9:348. 2010.

[7] Wahab, Iis H A., Adhi Susanto and Litasari. “Classification of Plasmodium falciparum Using Learning vector quantization Neural Network”. Proceeding of International Conference on Biomedical Engineering, pp 17. Surabaya, Indonesia. 2008.

[8] Wahab, Iis H A., Adhi Susanto dan Litasari. “Pengaruh Penggunaan Tapis Bilateral Terhadap Akurasi Identifikasi Plasmodium falciparum Pada Jaringan Syaraf Tiruan Learning vector quantization”. Prosiding pada Seminar Ilmiah Ilmu Komputer Nasional, halaman E21-27. Karawaci, Tanggerang, Indonesia. 2008.

[9] Wahab, Iis H A, Adhi Susanto, P. Insap Santosa, dan Maesadji.T. “Comparison of Cross Entropy Method, Optimize Learning Vector Quantization and Fuzzy Learning Vector Quantization for Classification of Plasmodium Falciparum”. Proceeding SITIA 2011. Institut Teknologi Sepuluh November. Surabaya. 2011.

[10] Subhamoy Mandal, Amit Kumar, J Chatterjee, Manjunatha and Ajoy K. “Segmentation of Blood Smear Images using Normalized Cuts for Detection of Malarial Parasites”. Annual IEEE India Conference (INDICON) 978-1-4244-9073-8/10. 2011.

[11] Vishnu V. Makkapati, Reghuveer M. Rao. “Segmentation of Malaria Parasites in Pheriperal Blood Smear Images”. ICASSP 2009. 978-1-4244-2354-5/09. 2009.

[12] F Boray Tek, Andrew G Dempster and Izzet Kale. “Computer vision for microscopy diagnosis of malaria”. Malaria Journal. 2009

[13] Mohammad Imroze Khan, Bhibhudendra Acharya, Bikesh Kumar Singh & Jigyasa Soni. ‘Content Based Image Retrieval

Approaches for Detection of Malarial Parasite in Blood Images. International" Journal of Biometrics and Bioinformatics (IJBB), Volume (5) : Issue (2) : 2011 97.2011.

[14] J.SOMASEKAR. "Computer Vision for Malaria Parasite Classification in Erythrocytes". International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 Vol. 3 No. 6 June 2011

[15] Dian Anggraini, Anto Satriyo Nugroho, Christian Pratama, Ismail Ekoprayitno Rozi, Aulia Arif Iskandar, Reggio Nurtanio Hartono. "Automated Status Identification of Microscopic Images Obtained from Malaria Thin Blood Smears". International Conference on Electrical Engineering and Informatics. Bandung, Indonesia. 2011.

[16] Maombi Edison, J.B Jeeva and Megha Singh. "Digital Analysis of Changes by Plasmodium Vivax Malaria in Erythrocytes". Indian Journal of Experimental Biology. Vol. 49, January 2011. pp. 11 – 15. 2011.

[17] S. S. Savkare, S. P. Narote. "Automatic Classification of Normal and Infected Blood Cells for Parasitemia Detection". International Journal of Computer Science and Network Security, VOL.11 No.2, February 2011

[18] Jigyasha Soni. "Advanced Image Analysis Based System for Automatic Detection of Malaria Parasite in Blood Images Using SUSAN Approach". International Journal of Engineering Science and Technology. Vol. 3 No. 6 June 2011.

[19] Premaratne, S P., Nadira Dharshani Karunaweera., Shyam Fernando., W Supun R Perera., and R P Asanga S Rajapaksha. "A Neural Network Architecture for Automated Recognition of Intracellular Malaria Parasites in Stained Blood Films". 2006.

[20] Gonzalez, R.C., Richard E. Woods, "Digital Image Processing", Prentice-Hall, Inc., Upper Saddle River, New Jersey, 2008.

[21] Haralick, R. M., Shanmugam, K., Dinstein, I., "Textural features for image classification", IEEE Trans. Syst., Man, Cybern., SMC, 1973.

[22] K. Z. Mao,. "Identifying Critical Variables of Principal Components for Unsupervised Feature Selection". IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 35, No. 2, April 2005

[23] Nurhayati, O.D, Th.Sri W, A.Susanto, Maesadji T., *Principal Component Analysis for Thermal Images Classification*, International Journal of Electronics Engineering Research (IJEER), December, New Delhi, India. 2010.