Principal Component Analysis combined with First Order Statistical Method for Breast Thermal Images Classification

[1]Oky Dwi Nurhayati, [2]Dr. Adhi Susanto, [3]Dr. Thomas Sri Widodo, [4]Dr. Maesadji Tjokronagoro
[1]Dept. of Computer Eng., Faculty of Eng., Diponegoro University, Indonesia.
[1]Dept. of Electrical Engineering and Information Technology, Faculty of Engineering, Gadjah Mada University, Yogyakarta, Indonesia
[2]Dept. of Electrical Eng. and Information Technology, Faculty of Eng., Gadjah Mada University, Yogyakarta. Indonesia
[3]Dept. of Electrical Engineering and Information Technology, Faculty of Engineering, Gadjah Mada University, Yogyakarta. Indonesia
[4]Dept. of Medical, Faculty of Medicine, Gadjah Mada University, Yogyakarta. Indonesia

Abstract
This paper is aimed to report the experiment in revealing the classification of randomized thermograms tabulated by the first order statistics method including the mean values, skewness values, entropy values, kurtosis values, and variance values with the thermal camera of Fluke as a tool for capturing images, after the mathematical method of measurement. Five statistical features combined with principal component analysis (PCA) have been applied in this research to classify the types of thermograms after the image preprocessing. The results show that the method is quite promising to distinguish the thermal images.

Keywords
thermal images, image preprocessing, first order statistics method, principal component analysis.

I. Introduction

PCA is a useful statistical technique that has found an application in fields such as face recognition and image compression, and also as a common technique for finding patterns in data of high dimension. PCA is used abundantly in all forms of analysis - from neuroscience to computer graphics - because it is a simple, non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA could provide a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it. Principal component analysis is a pre-processing transformation that creates new images from the uncorrelated values of different images. This is then accomplished by a linear transformation of variables corresponding to a rotation and translation of the original coordinate system. It refers to a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data could be difficult to find in data of high dimension, where the luxury of graphical representation is not available, PCA in return is becoming a powerful tool for analyzing the data.

The first principal component (the eigenvector with the largest eigenvalue) corresponds to a line that passes through the mean and minimizes sum squared error with those points. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted out from the points. Each eigenvalue indicates the portion of the variance that is correlated with each eigenvector. Thus, the sum of all the eigenvalues is equal to the sum squared distance of the points with their mean divided by the number of dimensions. PCA essentially rotates the set of points around their mean in order to align with the first few principal components. This moves as much of the variance as possible (using a linear transformation) into the first few dimensions. The values in the remaining dimensions, therefore, tend to be highly correlated and may be dropped with minimal loss of information. PCA is often used in this manner for dimensionality reduction. PCA has the distinction of being the optimal linear transformation for keeping the subspace that has largest variance. This advantage, however, comes at the price of greater computational requirement if compared, for example, to the discrete cosine transform. Nonlinear dimensionality reduction techniques tend to be more computationally demanding than PCA.

Another main advantage of PCA is that once the patterns in the data have been found and the data and examples have been compressed, by reducing the number of dimensions, the information in it would not be much in lost. Five features namely mean value, entropy value, skewness value, kurtosis value, and variance value were applied in this research and been plotted in a two-dimension graph to classify the types of the thermogram images.

II. The Underlying Theory

To probe further thermal images, some standard image processes such as resizing images into 256x256 pixels and converting true color into grayscale thermograms were applied. The statistical method was used to extract the texture feature of an image such as mean value, entropy value, skewness value, kurtosis value, and variance value. Image characteristics such as the arranged pixel intensity and statistical texture feature were counted from the image intensity.

In this case, the first order statistics were applied after image preprocessing of the thermograms to obtain the matrix data. Then the process of classification was carried out to separate normal, chemotherapy, and advanced thermograms from random thermal images.

Hence the objective of the research is to analyze the types of thermograms by applying the measurement of mean value, entropy value, skewness value, kurtosis value, and variance value from images, and principal component analysis method.

III. Materials and Methods

The present research was performed at Dr.Sarjito Hospital Yogyakarta in which 150 women were examined. Digital thermal camera Fluke was used for thermogram acquisition. Three groups were then assigned including Healthy Group consisted of 50 images, Chemotherapy Group with 50 images, and Advanced Group with 50 images.

Several methods in the image processing included resizing image and converting true color image to grayscale image.

A. Resizing Image

To resize an image, the imresize function in matlab was used. The images then were converted to be 256x256 pixels in obtaining a square matrix from images and mathematical accounting could be much easier by application of PCA technique.

B. Grayscale Image

The objective in converting true color to grayscale is to enhance the computation with the program and to make it easier in accounting statistical feature extraction, namely mean value and entropy value from gray level histogram.

C. Statistical Feature Extraction

Feature extraction is the process of defining a set of features, or image characteristics, which will most efficiently or meaningfully represent the information that is important for analysis and classification. Much of the information in the data set may be of little value for discrimination. Indeed, pattern recognition using the original measurements is frequently inefficient and may even obscure interpretation. First order statistics or moments of the gray level histogram are the nth moment of the (normalized) gray level histogram is given by:

$$\mu_n = \sum_{i=1}^{L} (k_i - mean)^n \, p(k_i)$$

where:
k_i = gray value of the ith pixel
mean = mean gray value of the pixel set
L = the number of distinct gray levels
$p(k_i)$ = normalized histogram (probability density function of the pixel set).
Note that the mean is given by:

$$mean = \sum_{i=1}^{L} k_i \, p(k_i)$$

Thus: $\mu_0 = 1$; $\mu_1 = 0$; $\mu_2 = s^2$ = variance
Variance is a square of standard deviation. The variance is given by:

$$\mu_2 = s^2 = \frac{1}{N} \sum_{i=1}^{L} [k_i - mean]^2$$

where N is total number of pixel in an image.
Skewness is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero.
The skewness is given by:

$$\mu_3 = \frac{1}{s^3} \sum_{i=1}^{L} (k_i - mean)^3 \, p(k_i)$$

Kurtosis is a measure of how outlier-prone a distribution is. The kurtosis of the normal distribution is 3. Distributions that are more outlier-prone than the normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3.
The kurtosis is given by:

$$\mu_4 = \frac{1}{s^4} \sum_{i=1}^{L} (k_i - mean)^4 \, p(k_i) - 3$$

Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image. Entropy is defined as:

$$\text{Entropy} = -\sum_{i1} \sum_{i2} p(i_1, i_2) \log p(i_1, i_2)$$

D. Step of PCA

Step 1: Getting some data.
In this research, three commands (load normal.dat, chemotherapy. dat, and advanced.dat) were applied. When using these sort of matrix techniques in computer vision, representation of image must be well considered. A square, N . by . N image can be expressed as an . N_2 dimensional vector:
$X = (x_1, x_2, \ldots x_n)$
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:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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:

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

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.

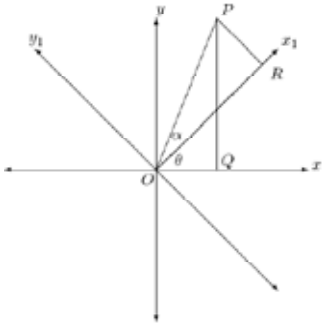

Fig. 1: Rotating the axes

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

Similarly $y = x_1 \sin\theta + y_1 \cos\theta$

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

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$ or X = PY , where: $X = \begin{bmatrix} x \\ y \end{bmatrix}$,

$$Y = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \text{and} \quad P = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

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

IV. Results and Discussion

Thermography reveals the infrared heat emittion distributions differently temperature. A proper image processing is to be developed to enhance the readability of the thermograms to ease the technician's diagnosis. Fig. 2 presents example of thermal image of healthy breast or normal thermogram. Fig. 3 presents example of thermal image of advanced breast cancer and Fig. 4 presents example of thermal image of chemotheraphy condition where breast cancer has evidently higher temperature and the temperature distribution is very asymmetric. Fig. 5 shows the true color of advanced breast cancer thermogram converted into grayscale image.

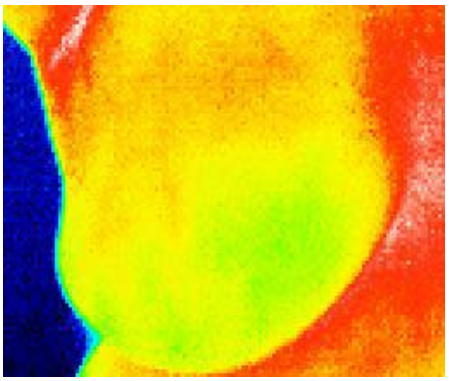

Fig. 2: Present example of thermal image of healthy breast

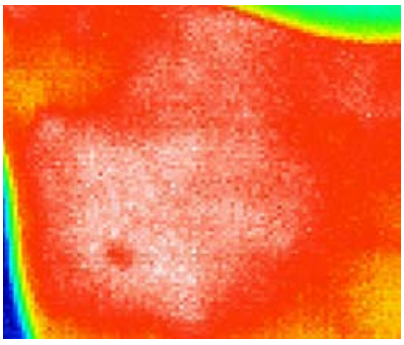

Fig. 3: Present example of thermal image of advanced breast cancer (white area spread on the breast)

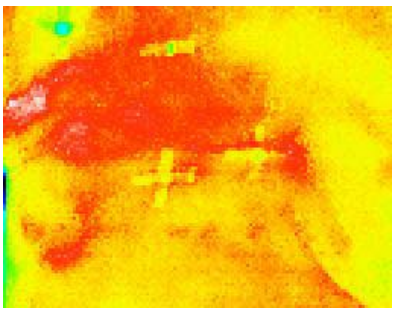

Fig. 4: Present example of thermal image of chemotheraphy condition (red area on the breast)

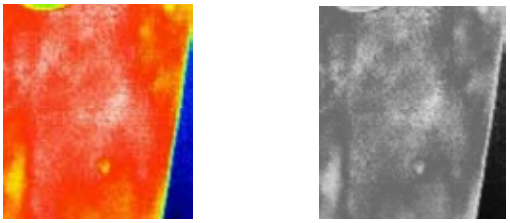

Fig. 5: (left) true color image, (right) converted in to grayscale

For each thermal image, we can create mean value, entropy value, skewness value, kurtosis value, and variance value accounted from grayscale images as described above. Fig. 6 through Fig. 9 show plots of raw data thermogram in two dimensions which is measured from two statistical features. This Fig.s show the randomized input of three kinds of thermograms. The normal thermograms are identified by red circle sign, chemotherapy thermograms are identified by green dot sign, and at last advanced thermograms are identified by a blue star sign.

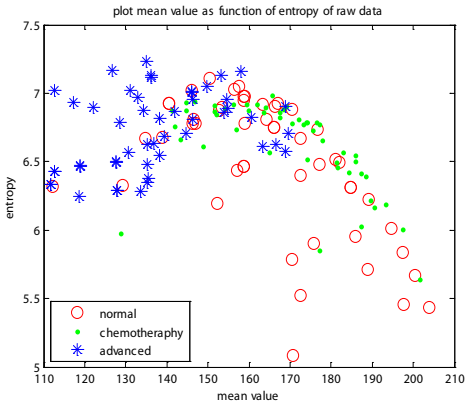

Fig. 6: Plot mean value as function of entropy of raw data thermograms

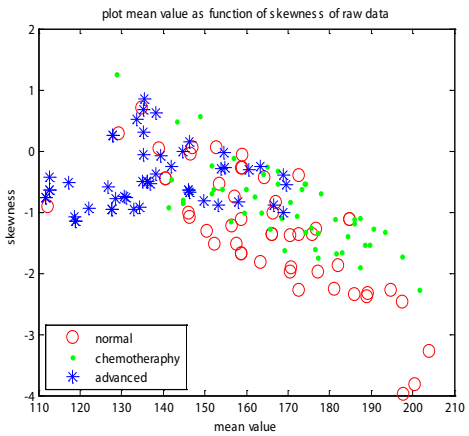

Fig. 7: Plot mean value as function of skewness of raw data thermograms

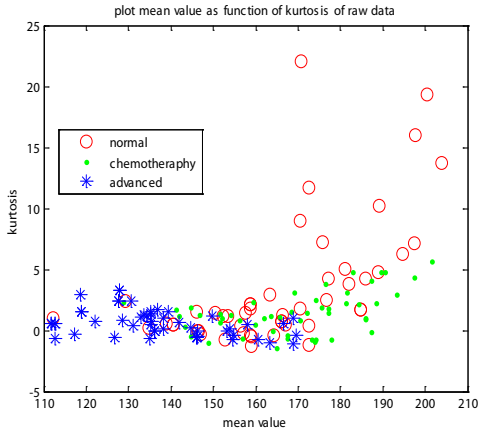

Fig. 8: Plot mean value as function of kurtosis of raw data thermograms

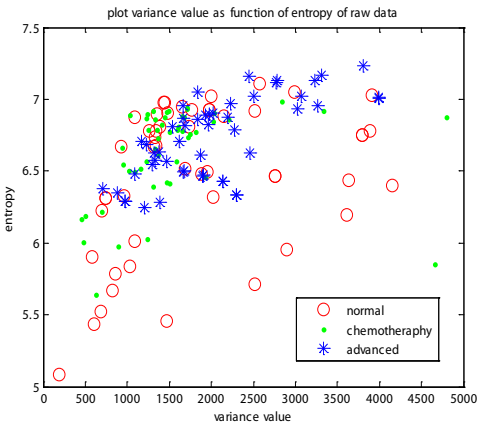

Fig. 9: Plot variance value as function of entropy of raw data thermograms

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, all the y values have y subtracted from them, and the z values have $\bar{z}$ subtracted from them. This produces a data set whose mean is zero. Plots of the raw data subtract the mean were showed in Fig. 10 through Fig. 13.

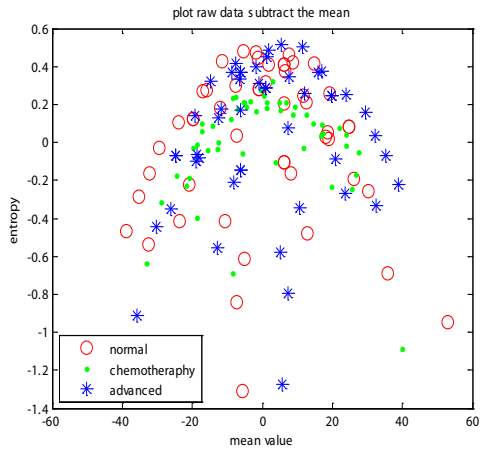

Fig. 10: Plot mean value as function of entropy of raw data subtract the mean

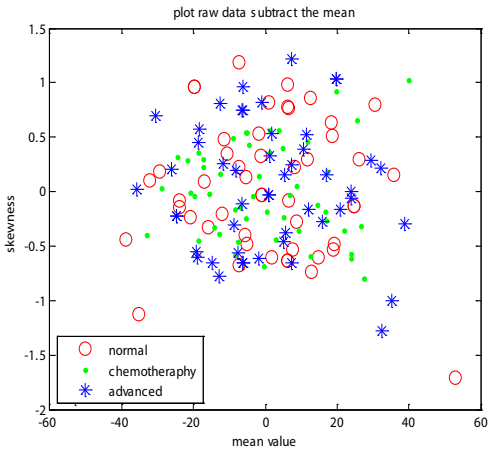

Fig. 11: Plot mean value as function of skewness of raw data subtract the mean

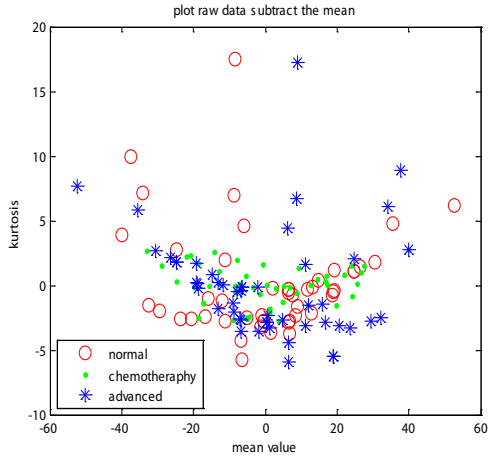

Fig. 12: Plot mean value as function of kurtosis of raw data subtract the mean

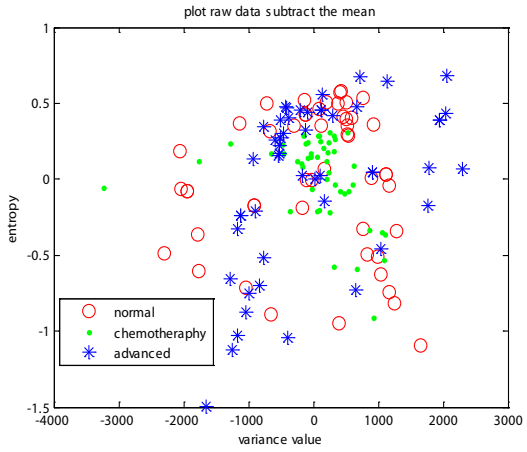

Fig. 13: Plot variance value as function of entropy of raw data subtract the mean

Calculating the covariant matrix, the eigenvectors and eigenvalues from raw data subtract the mean. Plots data subtract the mean with rotation matrix P were showed in Fig. 14 through Fig. 17.

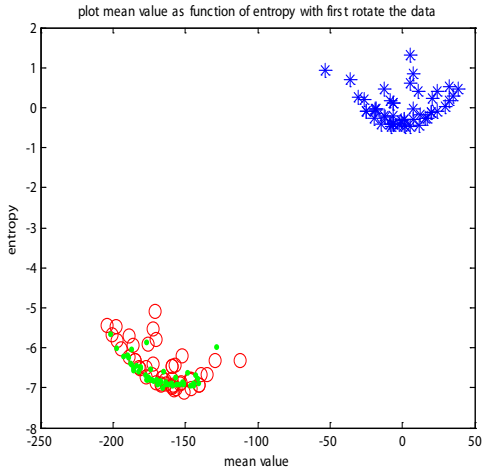

Fig. 14: Plot mean value as function of entropy of raw data subtract the mean with rotation matrix P

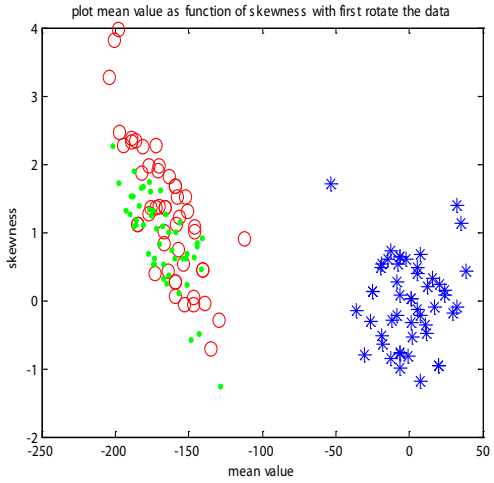

Fig. 15: Plot mean value as function of skewness of raw data subtract the mean with rotation matrix P

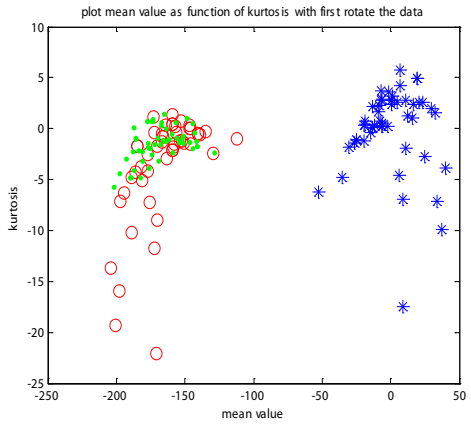

Fig. 16: Plot mean value as function of kurtosis of raw data subtract the mean with rotation matrix P

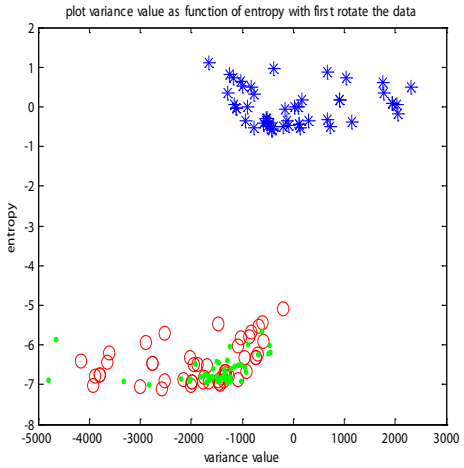

Fig. 17: Plot variance value as function of entropy of raw data subtract the mean with rotation matrix P

For the better result, the new coordinates with rotation matrix Pt in terms of the old ones in Fig. 18 through Fig. 21 could be also solved.

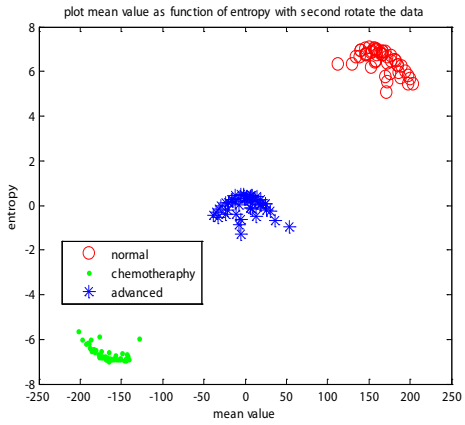

Fig. 18: Plot mean value as function of entropy of raw data subtract the mean with rotation matrix Pt

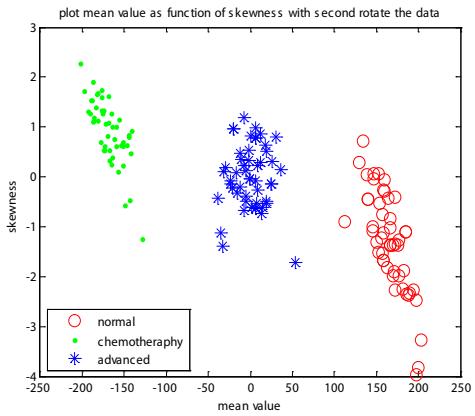

Fig. 19: Plot mean value as function of skewness of raw data subtract the mean with rotation matrix Pt

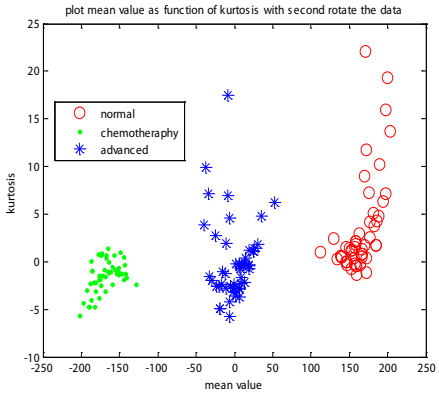

Fig. 20: Plot mean value as function of kurtosis of raw data subtract the mean with rotation matrix Pt

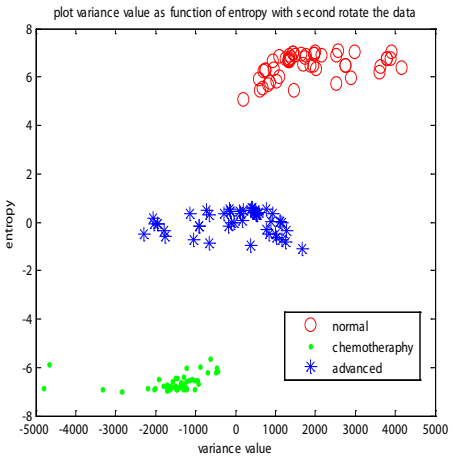

Fig. 21: Plot variance value as function of entropy of raw data subtract the mean with rotation matrix Pt

It turns out that these axes works much better in recognizing faces, because the PCA analysis has provided the original images in terms of the differences and similarities between them. The PCA analysis has identified the statistical patterns in the data. Plots for the new coordinates in 3 dimensions rotation were showed in Fig. 22 through Fig. 25.

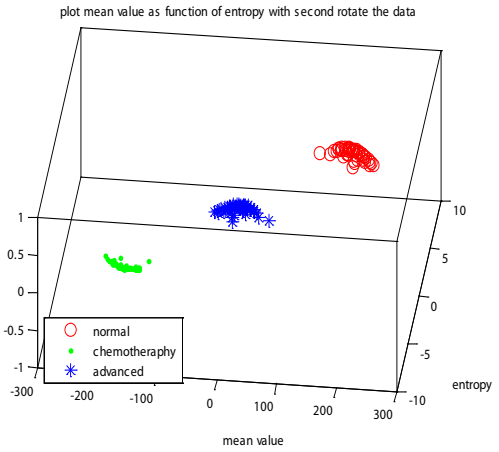

Fig. 22: Plot mean value as function of entropy of raw data subtract the mean with rotation matrix Pt in 3D rotation

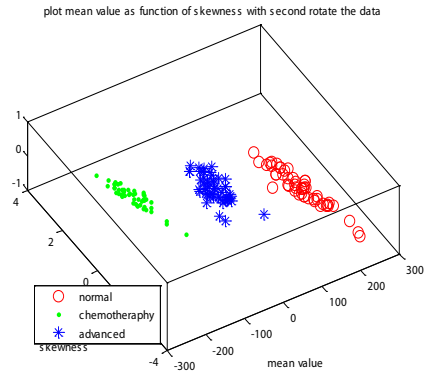

Fig. 23: Plot mean value as function of skewness of raw data subtract the mean with rotation matrix Pt in 3D rotation

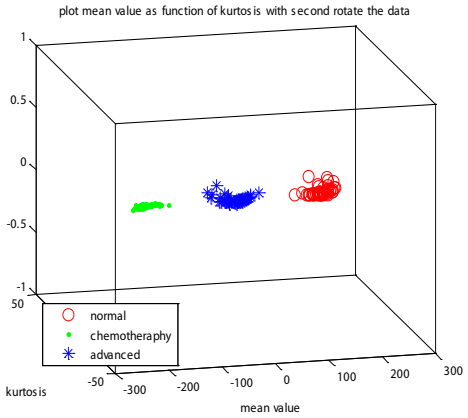

Fig. 24: Plot mean value as function of kurtosis of raw data subtract the mean with rotation matrix Pt in 3D rotation

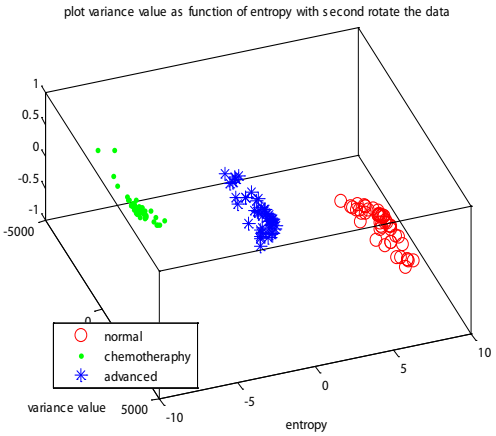

Fig. 25: Plot variance value as function of entropy of raw data subtract the mean with rotation matrix Pt in 3D rotation

Table 1 shows the measurement of mean value and entropy value of average normal, chemotherapy, and advanced thermograms.

Table 1: the average value of thermograms

Type of thermo-gram	mean	variance	skew-ness	kurtosis	entropy
Normal	170,8	1720,6	-1,09	1,82	6,61
Chemo-therapy	164,5	1815,9	-1,27	3,00	6,56
Advanced	137,6	2177,1	-0,57	0,45	6,82

Table 2 shows the coefficient correlation of thermograms. There is no any change value of x,y (denoting normal, chemotherapy thermo-grams) in first and second transformation, yet change value to zero for advanced thermograms. The result in Table 2 shows an exist linear dependence of two variables x and y, x and z, or y and z. One of the variables (zt1) altered transformation as an independent variable related to the other x and y which provides zero value from it.

Table 2: Coefficient correlations of thermograms

Type of thermo-gram	mean	variance	skew-ness	kurtosis	entropy
Normal	170,8	1720,6	-1,09	1,82	6,61
Chemo-therapy	164,5	1815,9	-1,27	3,00	6,56
Advanced	137,6	2177,1	-0,57	0,45	6,82

Note: x is variable of the normal thermograms, y is variable of the chemotherapy thermograms, and z is variable of the advanced thermograms. The transformation variables denote xt1, zt1, yt1, xt2, yt2, zt2 for first and second transformation respectively.

V. Conclusions

The experimental results show that first order statistical measurement namely mean value, variance value, skewness value, kurtosis value, and entropy value combined with principal component analysis method is promising to distinguish the types of thermal images. Further experiments should be coupled with spectral and structural methods for analysis.

VI. Acknowledgement

The first author would like to acknowledge the Dr.Sardjito hospital, Yogyakarta for supporting this work, Electrical Engineering and Technology Information Department, University of Gadjah Mada, for permitting to carry my PhD work from Dept of Computer Engineering, Diponegoro University, Semarang, Indonesia.

References

[1] A. K., "Fundamentals of Digital Image Processing", Prentice-Hall, Inc., A Division of Simon & Schuster Engelwood Cliffs, New Jersey, 1989.

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

[3] Keyserling, J.R, P.D.Ahlgren, E.Yu, N.B.Belliveau, "Overview of functional infrared imaging as part of a multi-imaging strategy for breast cancer detection and therapeutic monitoring", Proc. 2nd Joint IEEE EMBS/BMES Conf., Houston, TX, pp. 1126-8, 2002.

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

[5] Lindsay I Smith, "A tutorial on Principal Components Analysis", 2002.

[6] Šebök, M., M. Šimko, M. Chupáč., "Infrared Measurement of Temperature and Spectral Filters Application". Measurument Science Review, 5 (3), pp 46-49, 2005

[7] Stéphane Lafon, Yosi Keller, Ronald R. Coifman., "Data fusion and multicue data matching by diffusion maps". IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(11):1784–1797, 2006.

[8] Stéphane Lafon, Ronald R. Coifman., "Diffusion maps. Applied and Computational Harmonic Analysis : Special issue on Diffusion Maps and Wavelets", 21(1):5–30, July 2006.

[9] Stoitsis,J.,I.Valavanis, S.G.Mougiakakou, S.Gole-matti, A.Nikita, K.S.Nikita, "Computer aided diagnosis based on medical image processing and artificial intelligence methods", Elsevier Journal,2006.

[10] Tan, T.Z., C.Quek, G.S. Ng, E.Y.K.Ng, "Breast Cancer Diagnosis Using Thermography and Complementary Learning Fuzzy Neural Network", School of Computer Engineering, Nanyang Tech-nological University, 2004.

Born in Semarang, 2nd October 1979. Graduated from the Department of Telecommunication Engi-neering, Faculty of Engi-neering, Telecommunicati-on Engineering Institute, Indonesia in 2002. Master from Post Graduate Program, Department Electrical Engineering, Gadjah Mada University, Indonesia in 2008. In the period of 2008— present she is a Ph. D. Candidate in the Department of Electrical Engineering, Faculty of Engineering, Gadjah Mada University, Yogyakarta, Indonesia. From 2009— present she is a lecturer in the Department of Computer System, Faculty of Engineering, Diponegoro University, Indonesia. Current and interest research is in biomedical image processing. Oky Dwi Nurhayati, S.T, M.T., is also a member of IAENG (International Association of Engineer, and IMAVIS (International Journal on Image and Vision Computing).

Dr. Adhi Susanto. Born in Banjar, Indonesia, 1940. M. Sc. (1966) and Ph.D. (1988) from University of California Davis.

A Professor in The Dept. of Electrical Engineering and Information Technology, Faculty of Engineering, Gadjah Mada University, Jogjakarta, Indonesia. Current research interest is Electronics Engineering, Image Processing, Signal Processing, Adaptive System, Classification and Pattern Recognation Techniques. A member in IEEE and Planetary Society.

Dr. Thomas Sri Widodo. Born in Klaten, Indonesia, 1950. A Professor in the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Gadjah Mada University, Yogyakarta. INDONESIA. Dipl. Ing. ENSERG, Grenoble, France, 1985 Doctorat d'Etudes Approfondies, Univ. Montpellier 2, France, 1986 Docteur de l'Universite, Univ. Montpellier2, France, 1988. Current research interest is system, signal and electronics

Dr. Maesadji Tjokronagoro. A Professor in The Dept of Medical Science, Faculty of Medicine, Gadjah Mada University, Jogjakarta, Indonesia. Current research interest are Medical Image Processing, Radiotherapy, and Oncology.