Performance Improvement of Leaf Identification System Using Principal Component Analysis

Abdul Kadir, Lukito Edi Nugroho, Adhi Susanto and Paulus Insap Santosa

Gadjah Mada University, Indonesia
{akadir, lukito, insap}@mti.ugm.ac.id, susanto@te.ugm.ac.id

Abstract

This paper reports the results of experiments in improving performance of leaf identification system using Principal Component Analysis (PCA). The system involved combination of features derived from shape, vein, color, and texture of leaf. PCA was incorporated to the identification system to convert the features into orthogonal features and then the results were inputted to the classifier that used Probabilistic Neural Network (PNN). This approach has been tested on two datasets, Foliage and Flavia, that contain various color leaves (foliage plants) and green leaves respectively. The results showed that PCA can increase the accuracy of the leaf identification system on both datasets.

Keywords: *Convexity, Leaf identification system, PCA, PNN, solidity*

1. Introduction

Several approaches have been introduced to classify a leaf [1], such as k-Nearest Neighbor Classifier (k-NN), Probabilistic Neural Network (PNN), Genetic Algorithm (GA), Support Vector Machine (SVM), and Principal Component Analysis (PCA). Most of researchers used green color leaves or ignored color information on leaves. For example, Zulkifli [2] proposed General Regression Neural Network to classify 10 kinds of plants with green color leaves. Wu et al. [3] used PNN to classify 32 kinds of green leaves. They shared dataset called Flavia. Several researchers used the dataset to test their classifiers. For example, Singh et al. [4] suggested SVM to implement a classifier that was reported could improve the accuracy, Shabanzade et al. [5] used Linear Discriminant Analysis (LDA) to test part of Flavia.

Actually, color as features in leaf identification system has been introduced by Man et al. [6]. They used the first order, the second order and the third of color moments in HSV color space. They claimed that the system can recognize 24 categories of plants with the average accuracy up to 92.2%. The color information also have been inserted into plant retrieval system by Kebabci et al. [7].

Several leaf classification systems have incorporated texture features to improve the performance, such as in [6] that used entropy, homogeneity and contraction derived from co-occurrence matrix came from Digital Wavelet Transform (DWT), in [8] that used lacunarity to capture texture of leaf and in [9] that used GLCM.

To implement the classification system, several features were extracted that combined shape, color, vein, and texture features. Then, all those features were processed by using Principal Component Analysis (PCA) to obtained orthogonal features. After that, the results were used in the classification system. This approach gave improvement in average accuracy for classifying 60 kinds of foliage plants and 32 kinds of green leaves.

2. Basic of the Proposed System

2.1 Features of the System

First of all, several features were incorporated in the classification system will be described. The features can be categorized into: 1) shape features, 2) color features, 3) texture features and 4) vein features. The Shape features were eccentricity, roundness, dispersion, solidity, convexity, and features called Generic Fourier Descriptors (GFDs). The color features were employed: mean, standard deviation and kurtosis of leaf' colors. Texture features were extracted from gray-level co-occurrence matrix (GLCM). Finally, vein features were two kinds of features extracted from vein of leaf.

2.1.1 Eccentricity

The eccentricity is defined as

$$eccentricity = \frac{w}{l},$$

where w is the length of the leaf' minor axis and l is the length of the leaf' major axis. This feature can be used to differentiate the rounded leaf and the long one.

2.1.2 Roundness

The roundness or circularity ratio is defined as

$$roundness = \frac{A}{p^2},$$

where A is the area of the leaf and P is the perimeter of the leaf. This feature also can be used to differentiate the rounded leaf and the long one.

2.1.3 Dispersion

Dispersion is ratio between the radius of the maximum circle enclosing the region and the minimum circle that can be contained in the region. Mathematically, it is notated as below

$$dispersion = \frac{\max(\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2})}{\min(\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2})},$$

In the above formula, $(\bar{x}, \bar{y})$ is the centroid of the leaf and (x_i, y_i) is the coordinate of a pixel in the leaf contour.

2.1.4 Solidity

Solidity is a feature derived from convex hull and defined as [10]

$$solidity = \frac{area}{convex\ area},$$

where area is the area of the leaf and convex area is the area of the convex hull. A convex hull is the smallest convex set containing all points in an object. Several algorithm, such as Graham Scan [11] can be used to generate convex hull. Figure 1 shows a sample of a convex hull, like a rubber that conveys a leaf.

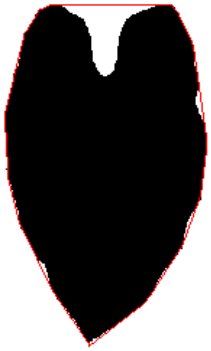

Figure1. Convex Hull (the red color)

2.1.5 Convexity

Convexity is also a feature derived from convex hull. It is defined as [10]

$$convexity = \frac{convex\ perimeter}{perimeter},$$

where convex perimeter represents the perimeter of the convex hull and perimeter is the circumstance of the leaf.

2.1.6 GFD

Generic Fourier Descriptors (GFDs) are descriptors derived from Polar Fourier Transform, proposed by Zhang [12]. The definition of PFT:

$$PFT(\rho,\phi) = \sum_r \sum_i f(\rho,\theta).\exp\left[j2\pi\left(\frac{r}{R}\rho + \frac{2\pi}{T}\phi\right)\right],$$

where

- $0<r<R$ and $\theta_i = i(2\pi/T)$ $(0<i<T)$; $0<\rho<R$, $0<\phi<T$,
- R is radial frequency resolution, and
- T is angular frequency resolution.

Firstly, an image I = {f(x, y); 0<x<M, 0<y<N} is converted to polar space Ip = {f(r,θ); 0<r<R, 0<θ<2π}, where R is the maximum radius from center of the shape. In this case, the origin of polar space is the center of space to get translation invariance. The centroid (x_c, y_c) was calculated by using formula

$$x_c = \frac{1}{M}\sum_{i=0}^{M-1} x,\quad y_c = \frac{1}{N}\sum_{i=0}^{N-1} y,$$

where M is the total of rows of the image and N is the total of columns of the image. Meanwhile, r and ɵ are computed by using:

$$r = \sqrt{(x-x_c)^2 + (y-y_c)^2}, \theta = \arctan\frac{y-y_c}{x-x_c},$$

Actually, PFT is not invariant to rotation and scaling. To achieve rotation invariance, the phase information in the coefficient should be ignored. Hence, only the magnitudes are used. To obtain the scale invariance, the first magnitude value is normalized by the area of the circle and all the magnitude values are normalized by the magnitude of the first coefficient. So, the GFDs are

$$GFD_s = \left\{ \frac{PF(0,0)}{2\pi r^2}, \frac{PF(0,1)}{PF(0,0)}, ..., \frac{PF(0,n)}{PF(0,0)}, ..., \frac{PF(m,0)}{PF(0,0)}, ..., \frac{PF(m,n)}{PF(0,0)}, \right\}$$

where m is the maximum number of the radial frequencies and n is the maximum number of angular frequencies. In this research, m = 6 and n = 4.

2.1.7 Color Moments

Color moments represents color features that are extracted from color information on the leaf by using statistical calculations such as mean (μ), standard deviation (σ), skewness (θ), and kurtosis (γ). In this research, those calculations were applied to each component in RGB color space. The four features are calculated as follows:

$$\mu = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} P_{ij}$$

$$\sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} - \mu)^2}$$

$$\theta = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} - \mu)^3}{MN\sigma^3}$$

$$\gamma = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (P_{ij} - \mu)^4}{MN\sigma^4} - 3$$

In this case, M is the height of the image, N is the width of the image, and P_{ij} is the value of color on rows i and column j.

2.1.8 Texture Features

This research used GLCM to capture textural information in the leaf. Five features were derived from GLCM : 1) Angular Second Moment (ASM), 2) contrast, 3) Inverse Different Moment (IDM), 4) entropy and 5) correlation. The five features are calculated as follows.

$$ASM = \sum_{i=1}^{L} \sum_{j=1}^{L} \left(GLCM(i, j)^2 \right)$$

$$Contrast = \sum_{i=1}^{L} \sum_{j=1}^{L} (i - j)^2 \left(GLCM(i, j) \right)$$

$$IDM = \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{(GLCM(i,j))^2}{1+(i-j)^2}$$

$$Entropy = -\sum_{i=1}^{L}\sum_{j=1}^{L}GLCM(i,j)x\log(GLCM(i,j))$$

$$Correlation = \sum_{i=1}^{L}\sum_{j=1}^{L}\frac{(ij)(GLCM(i,j)-\mu_1'\mu_2')}{\sigma_i'\sigma_j j'}$$

where

$$\mu_i' = \sum_{i=1}^{L}\sum_{j=1}^{L} i*GLCM(i,j)$$

$$\mu_j' = \sum_{i=1}^{L}\sum_{j=1}^{L} j*GLCM(i,j)$$

$$\sigma_i^{\,2} = \sum_{i=1}^{L}\sum_{j=1}^{L} GLCM(i,j)(i-\mu_i')^2$$

$$\sigma_j^{\,2} = \sum_{i=1}^{L}\sum_{j=1}^{L} GLCM(i,j)(i-\mu_i')^2$$

L is the number of rows/columns in GLCM. The co-occurrence matrix GLCM(i,j) counts the co-occurrence of pixels with grey value i and j at given distance d. The direction of neighboring pixels to represents the distance can be selected, for example 135^o, 90^o, 45^o, or 0^o. In this research, the four directions were calculated. When calculating ASM, IDM, entropy, contrast and correlation, the average ones were used as features. This action was done to overcome rotation dependency in GLCM.

2.1.9 Vein Features

Based on pixels on the vein, 2 vein features can be derived. In this research, vein was generated by using a morphological operation called opening [3]. That operation was performed on the gray scale image using disk-shaped structuring element of radius 1 and 2 and subtracted remained image by the margin. By using that operation, a structure like vein was obtained. Then, 2 features were calculated by using following formulas:

$$V_1 = \frac{A_1}{A},\ V_2 = \frac{A_2}{A}\ ,$$

V_1 and V_2 represent features of the vein, A_1 and A_2 are total pixels of the vein, and A denotes total pixels on the part of the leaf.

2.2 PNN for Leaf Identification

The leaf identification system was developed by using Probabilistic Neural network (PNN). PNN is a kind of neural networks that can learn fast from training data and guarantees to converge to an optimal classifier as the size of the representative training set increases [13].

PNN is an implementation of statistical algorithm called kernel discriminant analysis, in which the operations are organized into a multilayered feed-forward network with four layers: 1) input layer, 2) pattern layer, 3) summation layer, and 4) output layer [13]. When an input is presented, the first layer computes distances from the input vector to the training input vectors and produces a vector whose elements indicate how close the input is to a training

input. The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilities [14]. Finally, a competitive transfer function on the output of the second layer picks the maximum of these probabilities, and produces a 1 for that class and a 0 for the other classes. The architecture for this system is shown in Figure 2.

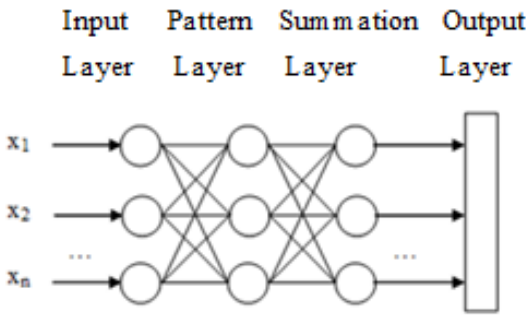

Figure 2. Architecture of PNN

2.3 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical method that the main goal is to reduce dimension of data. According to Slens [15], with minimal effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it.

PCA has been used for several applications, such as for breast thermal classification [16] and analysis of socioeconomic factors and their association with malaria in children [17]. In those researches, PCA was used to reduce dimension of data.

The algorithm of PCA is described as follow [18].

1. Calculate the mean of data and store to μ:

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

2. Subtracted all data with μ : ($\bar{x} \leftarrow x - \mu$).

3. Calculate the covariance of the data:

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

4. Calculate the eigenvalues and the eigenvectors based on the covariance matrix.

5. Calculate the final data by using the following formula:

$$finalData - eigenvectors * \bar{x}$$

To reduce the features to n dimensions, n eigenvectors of the n highest of eigenvalues are selected.

3. Model of the Proposed System

The model of the proposed system is shown in Figure 3. First of all, the leaf of query is preprocessed to separate the leaf from its background.

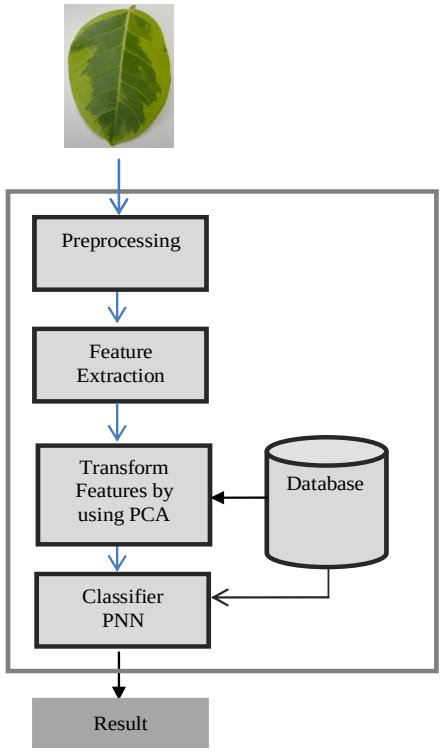

Figure 3. Model of the Leaf Identification System

Then, all features described previously are extracted and then are transformed by using PCA by incorporating the data in the database for transforming the features. The result is inputted to the classifier. Finally, the classifier decides the plant based on the training data in the database.

The database was prepared once. All leaf images for training data purpose were preprocessed and then all features were extracted. After that, all features were transformed by using PCA and stored into the database. Table 1 shows detail of features included in the system. Total of features was 54.

4. Experimental Results

In order to test the system, dataset called Foliage that was prepared for the experiments was used. The dataset contains 60 kinds of foliage leaves, with various colors and shapes (Figure 4).

Experiments were done by using 95 leaves per plant for training purpose and 20 leaves per plant for testing purpose. Without PCA, the accuracy was 93%. When PCA was used, the results are shown in Table 2.

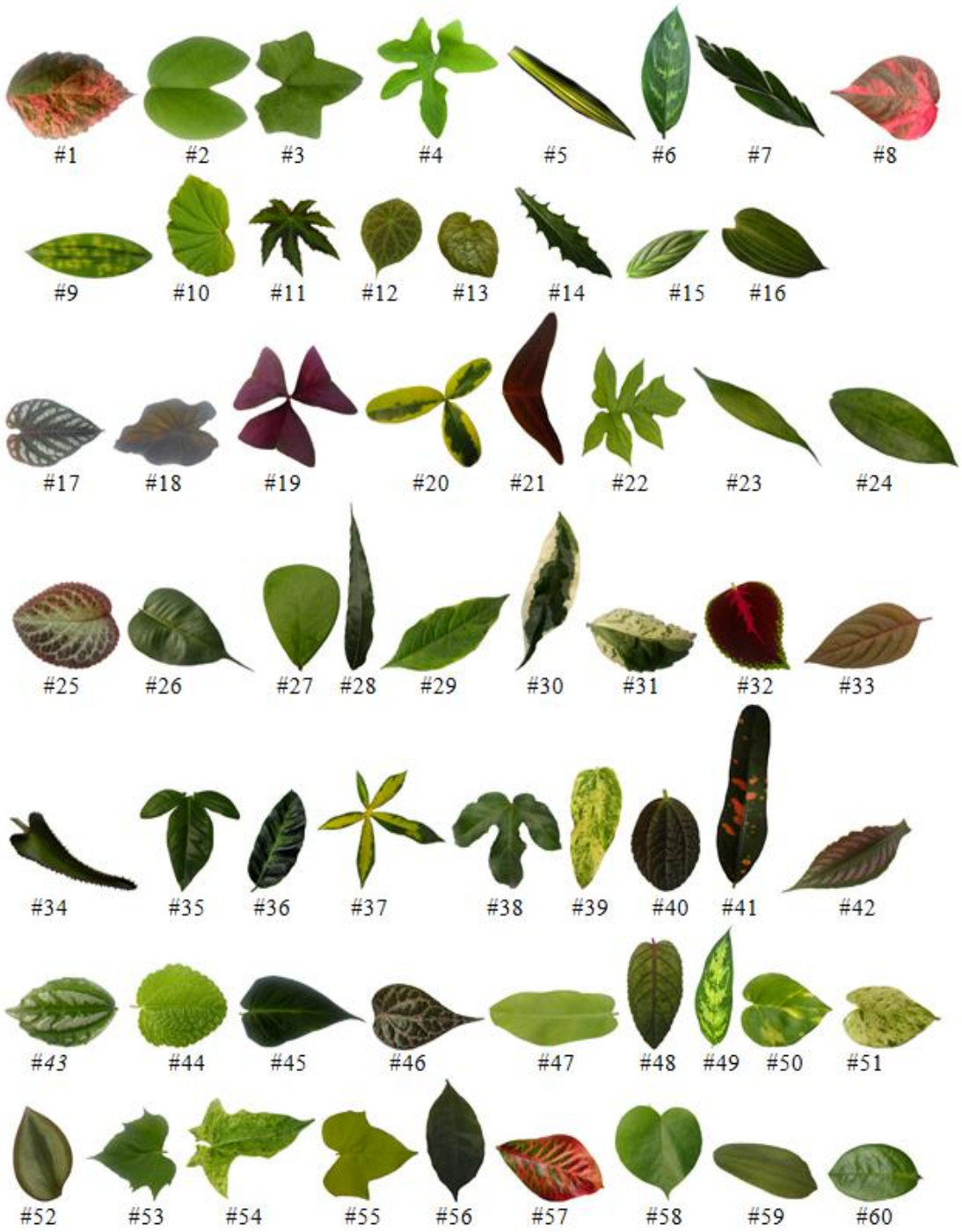

Figure 4. Sixty Kinds of Leaves in the Dataset Foliage

Table 1. Features in the System

Features	Total
Shape features: • eccentricity: 1 • roundness: 1 • dispersion: 1 • solidity: 1 • convexity: 1 • GFDs (without $\frac{PF(0,0)}{2\pi r^2}$): 30	35
Vein features	2
Texture features: • ASM: 1 • Contrast: 1 • IDM: 1 • Entropy: 1 • Correlation: 1	5
Color moments (R,G,B): • mean : 3 • standard deviation : 3 • skewness : 3 • kurtosis : 3	12
All features	54

Table 2. Results when PCA was used in the System by using Dataset Foliage

Number of Features	Accuracy
54	94.6667%
50	94.8333%
40	94.4167%
30	94.9167%
25	95.7500%
20	94.1667%
15	94.0000%
10	86.1667%
6	62.0833%
5	57.6667%

The first result in Table 2 (number of features = 54) shows that without reducing the features when using PCA, the accuracy of the system was improved. When the dimensions were reduced to 15 features, the system still gave better performance than when PCA was not included. The other interesting result, by reducing the features to 25, the system gave optimum result (97.75% of accuracy).

To compare to other researches, the experiments were also tried to use dataset Flavia [3]. The experiments involved 32 kinds of leaves with 40 leaves per plant for training

data and 10 leaves per plant for testing data. The results are shown in Table 3. Meanwhile, without PCA, the accuracy was 93.75%.

Table 3. Results when PCA was used in the System by using Dataset Flavia

Number of Features	Accuracy
54	93.4375%
50	92.8125%
40	92.8125%
30	93.4375%
25	95.0000%
20	94.3750%
15	91.5625%
10	81.5625%
6	53.4365%
5	50.6250%

Based on Table 3, the optimum accuracy on the dataset Flavia was reached when the dimension was reduced to 25 features.

Table 4 list the accuracy of the proposed system and other methods proposed by other researchers that used dataset Flavia. It is shown that the proposed system gave better accuracy than the others.

Table 4. Accuracy Comparison

Methods	Accuracy
PNN in [3]	90.3120%
SVM in [4]	81.5600%
PNN-PCNN in [4]	91.2500%
Fourier Moment in [4]	46.3000%
PNN-PFT in [10]	94.6875%
The proposed system	95.0000%

5. Conclusions

A leaf identification system that incorporates PCA has been developed. The experiments show that PCA can improve the accuracy of the system, from 93% to 95.7500% when dataset Foliage was used and from 93.4375% to 95% when dataset Flavia was used. That accuracy was reached when the dimension was reduced from 54 features to 25 features. Compared to other methods that used dataset Flavia, the proposed system gave better accuracy. However, some other works will be explored to obtain better performance.

References

[1] M. Kumar, M. Kamble, S. Pawar, P. Patil and N. Bonde, “Survey on Techniques for Plant Leaf Classification”, International Journal of Modern Engineering Research, vol. 1, no. 2, **(2011)**, pp. 538-544.

[2] Z. Zulkifli, "Plant Leaf Identification using Moment Invariants & General Regression Neural Network", Master Thesis, Universiti Teknologi Malaysia, **(2009)**.
[3] G. Wu, F. S. Bao, E. Y. Xu, Y. X. Wang, Y. F. Chang and Q. L. Xiang, "A Leaf Recognition Algorithm for Plant Classification using Probabilistic Neural Network", IEEE 7th International Symposium on Signal Processing and Information Technology, **(2007)**.
[4] K. Singh, I. Gupta and S. Gupta, "SVM-BDT PNN and Fourier Moment Technique for Classification of Leaf Shape", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 3, no. 4, **(2010)**, pp. 67-78.
[5] M. Shabanzade, M. Zahedi and S. A. Aghvami, "Combination of Local Descriptors and Global Features for Leaf Classification", Signal Image Processing: An International Journal, vol. 2, no. 3, **(2011)**, pp. 23-31.
[6] Q. K. Man, C. H. Zheng, X. F. Wang and F. Y. Lin, "Recognition of Plant Leaves using Support Vector", International Conference on Intelligent Computing, **(2008)**, pp. 192-199, Shanghai.
[7] H. Kebabci, B. Yanikoglu and G. Unal, "Plant Image Retrieval using Color, Shape, & Texture Features", The Computer Journal Advance Access, vol. 53, no. 1, **(2010)**.
[8] A. Kadir, L.E. Nugroho, A. Susanto and P. I. Santosa, "Leaf Classification using Shape, Color, and Texture Features", International Journal of Computer Trends and Technology, vol. 1, no. 3, **(2011)**, pp. 225-230.
[9] A. Kadir, L.E. Nugroho, A. Susanto and P. I. Santosa, "Neural Network Application on Foliage Plant Identification", International Journal of Computer Applications, vol. 29, no. 9, **(2011)**, pp. 15-22.
[10] J. C. Russ, "The Image Processing Handbook", CRC Press, Boca Raton, **(2011)**.
[11] M. T. Goodrich and R. Tammasia, "Algorithm Design", John Wiley & Sons, **(2002)**.
[12] D. Zhang, "Image Retrieval Based on Shape", Unpublished Dissertation. Monash University, **(2002)**.
[13] V. Cheung and K. Cannons, "An Introduction to Probabilistic Neural Networks", http://www.psi.toronto.edu/~vincent/research/presentations/PNN.pdf, **(2003)**.
[14] H. Demuth, M. Beale and M. Hagan, "Neural Network Toolbox 6 User's Guide", Natick, The MathWorks, Inc., **(2009)**.
[15] J. Slens, "A Tutorial on Principal Component Analysis", http://www.snl.salk.edu/~shlens/pca.pdf, **(2009)**.
[16] O. D. Nurhayati, A. Susanto, T. S. Widodo and M. Tjokronagoro, "Principal Component Analysis Combined with First Order Statistical Method for Breast Thermal Images Classification", International Journal of Computer Science and Technology IJCST), vol. 2, no. 2, **(2011)**, pp. 12-18.
[17] A. C. Krefis, N. G. Schwarz, B. Nkrumah, S. Acquah, W. Loag, N. Sarpong, Y. Adu-Sarkodie, U. Ranft and J. May, "Principal Component Analysis of Socioeconomic Factors and Their Association with Malaria in Children from the Ashanti Region", Ghana, Malaria Journal, vol. 9, **(2010)**.
[18] L. I. Smith, "A Tutorial on Principal Component Analysis", http://www.cs.otago.ac.nz/ cosc453/student_tutorials/principal_components.pdf, (2002).

Authors

Abdul Kadir

Received B.Sc. in Electrical Engineering from Gadjah Mada University, Indonesia, in 1987, M. Eng. in Electrical Engineering from Gadjah Mada in 1998, and Master of Management from Gadjah Mada University in 2004. His research interests include image processing, pattern recognition and web-based applications.

Lukito Edi Nugroho

Received B.Sc. in Electrical Engineering from Gadjah Mada University, Indonesia, in 1989, M.Sc. from James Cook University of North Queensland in 1994, Ph.D from School of Computer Science and Software Engineering, Monash University, in 2002. His research interests are software engineering, information systems and multimedia.

Adhi Susanto

Professor emeritus at Gadjah Mada University, Indonesia. He received Bachelor in Physics in 1964 from Gadjah Mada University, Indonesia, Master in Electrical Engineering in 1966 from University of California, Davis, USA, and Doctor of Philosophy in 1986 from University of California, Davis, USA. His research interests areas are electronics engineering, signal processing and image processing.

Paulus Insap Santosa

Obtained his undergraduate degree from Universitas Gadjah Mada, Indonesia, in 1984, master degree from University of Colorado at Boulder in 1991, and doctorate degree from National University of Singapore in 2006. His research interests include human computer interaction and technology in Education.