TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 4, August 2020, pp. 1897~1903
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i4.14069

Rice seed image classification based on HOG descriptor with missing values imputation

Huy Nguyen-Quoc, Vinh Truong Hoang
Faculty of Computer Science, Ho Chi Minh City Open University, Vietnam

Article Info

Article history:

Received Sep 9, 2019
Revised Feb 16, 2020
Accepted Mar 18, 2020

Keywords:

HOG descriptor
KNN imputation
Linear interpolation
Missing value imputation
Rice seed image classification
Zero imputation

ABSTRACT

Rice is a primary source of food consumed by almost half of world population. Rice quality mainly depends on the purity of the rice seed. In order to ensure the purity of rice variety, the recognition process is an essential stage. In this paper, we firstly propose to use histogram of oriented gradient (HOG) descriptor to characterize rice seed images. Since the size of image is totally random and the features extracted by HOG can not be used directly by classifier due to the different dimensions. We apply several imputation methods to fill the missing data for HOG descriptor. The experiment is applied on the VNRICE benchmark dataset to evaluate the proposed approach.

Corresponding Author:

Vinh Truong Hoang,
Faculty of Computer Science, Ho Chi Minh City Open University,
97 Vo Van Tan Street, Ward 6, District 3, HCM City, Vietnam.
Email: vinh.th@ou.edu.vn

1. INTRODUCTION

Rice cultivation is the main agriculture in many countries. In order to produce high quality rice seeds, classification is the most important stage. The intention of classification stage is to ensure the purity of the rice seeds by removing bad ones or rice seeds from another varieties. This process is usually controlled by some standards to ensure the rice seed quality and purity before selling to the farmers for mass growing. Nowadays, in Vietnam, the classification stage is usually done by eyes of skillful farmer based on some features, this method is time consuming and may give low quality seeds. With the development of information technology, the computer vision has been applied in agriculture by various applications of pattern recognition such as detection of diseases in human and plants, manufacturing automatic inspection, automatic characterization, fruit, vegetable and grain quality assessment [1-3]. The analysis and the detection of rice seeds is carried by an automatic computer-aided vision system. The step of analyzing and recognizing images requires defining descriptors which represent different classes of seed textures and can discriminate against them.

Furthermore, texture analysis is an intensive research topic over the years. A wide variety of local image descriptors have been proposed and this significantly contributed the progress in image analysis and other machine vision tasks. According to a recent survey of Humeau-Heurtier [4, 5] about texture analysis, texture attributes can be divided into seven categories defined in terms of statistical, structural, transform-based, model-based, graph-based, learning-based, and entropy-based. Several texture analysis approaches based on global feature, include color Gabor filtering [6], Markov random field model [7]. Some of the effective local feature methods are color scale invariant feature transform (SIFT) [8], color pyramid of histograms of oriented gradients (PHOG) [9], discriminative color descriptors (DCD) [10], three-dimensional adaptive sum and difference histograms (3D-ASDH) [11], color local binary pattern [12, 13] and affine wavelet [14]. Among of them, histogram of oriented gradients (HOG) [15] is successfully applied for image classification and object detection. Duong and Hoang [16] apply to extract rice seed images based on features coded in multiple color spaces using HOG descriptor. Phan et al. [17] evaluate and compare different local image descriptors (GIST, SIFT, morphological features) and classifier (random forest, KNN, SVM) for rice seed varieties identification. They showed that the random forest gives the best results for discriminating rice seed images. More recently, Vu et al. [18] propose to use morphological and geometrical features to classify three groups of rice seed varieties.

In the past, many works propose to fuse features extracted from local image descriptors in order to enhance the performance. For example, Lurstwut and Pornpanomchai [19] recently present a method to evaluate rice seed germination images based on neural networks. They extract and fuse three features (color, morphology and texture) from rice seed images in order to evaluate the germination. Mebatsion et al. [20] combine fourier descriptors and three geometrical for automatic classification of non-touching cereal grains. Szczypiskietal [21] identify the barley varieties based on image attributes extracted from shape, color and texture of individual kernels. Chaugule and Mali [22] propose a new feature extraction approach for classifying paddy seeds based on seed color, shape, and texture from horizontal vertical and front rear angles. Kuoetal [23] recognize rice grains image by using the sparse-representation-based classification on the 30 varieties rice reproduced in a local greenhouse at Taiwan Li et al. [24] use the laser scanning system to acquire the three-dimensional point cloud of a rice seed. The length, width, thickness and shape of rice seed are computed based on the oriented bounding box. Hoai et al. [25] introduce a comparative study of hand-crafted descriptors and convolutional neural networks (CNN) for rice seed images classification.

However, for the real-world application, we recognize that HOG feature vectors extracted from images with random sizes have different numbers of dimension which is impossible to classify. The reason is because of the difference of the image size. Current solution for this problem is resize all the image set to one general size, but this method may cause problems like low resolution, information loss, etc. Another approach can be used to solve this problem is missing value imputation [26]. This process allows to replace the missing value data with substituted values. There have been many approaches developed for classifying the incomplete data. The first one is to remove the missing value patterns directly. However, this approach can only be realized when the missing data set is small. In the last few years, missing value imputation problem has attracted more attention by many researchers. The investigations cover a wide range of techniques, from statistical imputation techniques and machine learning-based imputation methods the statistical imputation methods use popular statistical methods such as the replacement by mean of the available data and regression models of missing values [5, 26, 27]. Lin and Tsai [28] review and analyze 111 journal papers published from 2006 to 2017 related to solve the problems of incomplete dataset including the choice of datasets, missing rates and missingness mechanisms, the missing value imputation techniques and evaluation metrics employed.

In order to tackle the limit of HOG features extracted from random size images, we propose to apply missing values imputation method to gain the same dimensional feature vector of all images including KNN imputation, zero imputation and linear interpolation. The following of this paper is organized as follows. Section 2 introduces research methods which are HOG descriptor and missing values imputation methods. Section 3 then describes the experimental results. Finally, conclusion and future works are presented in section 4.

2. RESEARCH METHOD

In this section, we briefly present the histograms of oriented gradient which is used to extract features from rice seed images. Then, several missing value imputation methods are discussed.

2.1. Histograms of oriented gradient descriptor

Histograms of oriented gradient (HOG) descriptor is widely used in object detection and classification, especially for person detection. It is first proposed by Dalal and Triggs [15]. Before computing HOG, several processing steps are adopted in order to reduce noise and increase the performance. Then, the gradient magnitude $M_{(x,y)}$ and angle of gradient $\alpha_{(x,y)}$ vector at each pixel are computed in an 8 × 8 pixels cell, this step is also called gradient computation. The gradient computation of pixel coordinate at (x, y) is fomulated as follows:

$$\Delta_x = |G(x-1, y) - G(x+1, y)| \qquad (1)$$

$$\Delta_y = |G(x, y-1) - G(x, y+1)| \quad (2)$$

$$M_{(x,y)} = \sqrt{\Delta_x^2 + \Delta_y^2} \quad (3)$$

$$\alpha_{(x,y)} = arctan\left(\frac{\Delta_y}{\Delta_x}\right) \quad (4)$$

where, grayscale value at coordinate (x, y) is defined as $G(x, y)$, Δ_x and Δ_y are horizontal and vertical gradient. The dimension of HOG feature vector depends on the cell size and the number of bin orientation used for building the intervals of the angles of the gradient. 64 adopted gradient features are divided into 9-bin histogram which is mainly used to build the intervals of the angles of the gradient from 0 to 180 degrees for unsigned gradients (or from 0 to 360 degrees in case of signed gradients). So, there will be 20 degrees per bin. For each gradient feature, its magnitude will be added into the corresponding angle in the histogram. Finally, histogram from all block (each block contains 2×2 cells and has 50% overlap) are normalized and combined into a feature vector. This descriptor has been applied in various applications such as face recognition [29, 30], computer-aided diagnosis of tuberculosis [31], medical image analysis [32] and traffic analysis [33].

2.2. Missing value imputation methods

We propose to adopt three missing value imputation methods which is presented in the following.

- KNN imputation **(**KNNI**)** is an imputation method based on the K-nearest neighbors' algorithm by using the correlation structure of the data. The missingvalue is imputed by take the weighted mean of K nearest values [34, 35]. This method is mostly used than mean imputation and other methods because it can handle both categorical data and continuous data with multiple missing values and higher accuracy. Based on Branden and Verboven, we adopt K = 10.
- Linear interpolation is a method of constructing new data point based on known data points. It is one of the simplest interpolation methods by taking two known data points to compute the missing value. The linear interpolation at the point $A(x_A, y_A)$ can be formulated as:
-

$$y_A = y_B + (y_C - y_B)\frac{x_A - x_B}{x_C - x_B} \quad (5)$$

where, $B(x_B, y_B)$ and $C(x_C, y_C)$ are known points. A is usually between B and C.

- Zero imputation is a simplest method that the missing values are substituted by zero. The aims of this method are to fully capture all axes of feature vector.

3. RESULTS AND ANALYSIS

3.1. Dataset

We use the benchmark rice seed (VNRICE) dataset which consists of six common Vietnam rice seed varieties, including BC-15, Huong Thom-1, Nep-87, Q-5, Thien Uu-8, Xi-23. These rice seeds are sampled from a rice seed production company where the rice varieties were grown and harvested following certain conditions for standard rice seeds production. All images are acquired by a CMOS image sensor color camera. Figure 1 shows example images from this dataset. Each column illustrates each category of VNRICE dataset. We see that this is really a challenge task even for human since the images look similar. The k-nearest neighbor (kNN) classifier associated with the L1-distance and the SVM classifier are considered in order to classify the rice seed images. The achievement of the classification is measured by the accuracy rate which was performed by split-sample validation with holdout sampling. A half of the data were used as the input of the classifier to build the training model while the rest were used to test it. Table 1 present the characteristic of VNRICE dataset via split-sample validation method.

Table 1. Characteristic of VNRICE dataset

Rice variety	# Training set	# Testing set	Total images
BC-15	917	917	1,834
Huong Thom-1	1,048	1,048	2,096
Nep-87	699	700	1,399
Q-5	962	962	1,924
Thien Uu-8	513	513	1,026
Xi-23	1,113	1,113	2,226

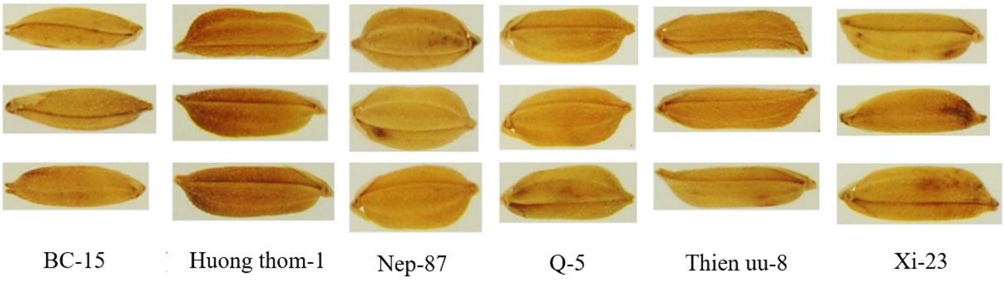

Figure 1. Example images from six rice seed varieties

3.2. Experimental setup and results

Most classifiers require the same dimension of input training data while the HOG feature vectors are mainly depended on the image size. The VNRICE dataset contains many images with different sizes. We first resize all the images into the same size before extracting the HOG features. The minimum and maximum of height (H) and width (W) of rice images from VNRICE are $W_{min} = 169$, $H_{min} = 46$, $W_{max} = 380$, $H_{max} = 103$. The classification results are presented in Table 2 by two classifiers. The second column represents the height and width of images after processing with random sizes. The third column shows the dimension of HOG features obtained corresponding with the resized image. The 6-NN classifier is considered since it gives the best performance in range of k ∈ {1,2,..,50}.

Table 2. Accuracy on the VNRICE dataset with random selected images size

No.	Image size	Dimension of HOG feature vector	6-NN (%)	SVM (%)
1	46 × 46	576	72.73	74.72
2	50 × 100	1,980	77.14	65.10
3	46 × 169	2,880	76.29	69.63
4	103 × 103	4,356	78.43	74.99
5	46 × 380	6,624	74.85	77.70
6	100 × 150	6,732	78.54	80.52
7	77 × 229	7,776	77.91	81.88
8	79 × 228	7,776	77.49	81.96
9	103 × 169	7,920	**80.08**	82.08
10	101 × 300	14,256	77.25	85.19
11	169 × 169	14,400	79.62	83.58
12	103 × 380	18,216	75.74	86.23
13	120 × 400	24,696	75.51	**86.95**
14	380 × 380	76,176	78.30	85.89
	Average accuracy		77.13	79.74

We observe that the best accuracy is reached when we resize image to 103×169 for 6-NN classifier and 120×400 for SVM classifier. It is impossible to determine which size is optimal to transform after 14 trials. Additionally, the dimension space of HOG features increases when the image size is bigger. In order to apply the missing value imputation, we extract the HOG features features from the original image with random sizes. The dimension of shortest and longest vector is 3,888 and 12,636 respectively. The three imputation methods presented in section 2 are then applied to fill the missing values. It is worth to note that the imputation stage is before the cross validation, so these methods are also in unsupervised learning context. All feature vectors are then assigned all missing values as NaN (not a number) from its original length to 12,636.

For zero imputation, it just simply replaces all NaNs by 0. When using KNN imputation method, it requires at least one completed feature vector to fill the missing values. In case of linear interpolation method, at least two vectors are required in order to proceed. Figure 2 illustrates the accuracy of three imputation methods on VNRICE dataset by 6-NN classifier. The original dimension is based on the shortest vector with 3,888 features. The missing values are filled by 100 features each time. From this chart, we see that linear interpolation clearly outperforms than two other methods. Surprisingly, the zero imputation gives better results than 10-NN imputation method from 7,888 features. Furthermore, Tables 3 and 4 detail the classification performance of each method on three different dimensions. Linear interpolation method with SVM classifier gives the best accuracy with 99.94%. We improve more than 20% compared to the average accuracy with random resized image in Table 2.

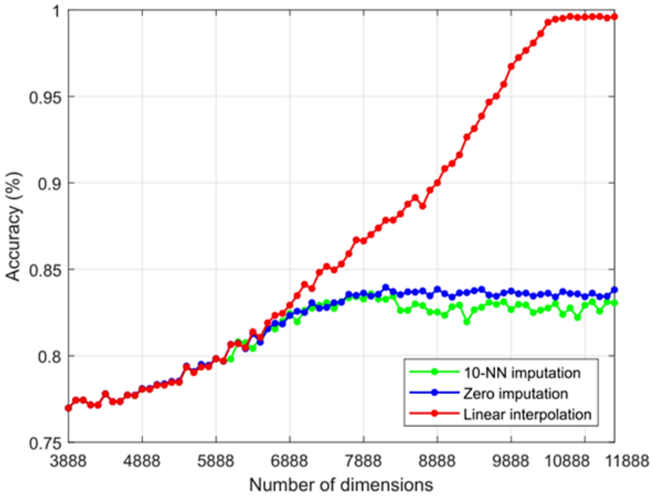

Figure 2. Fill missing values methods experiment results chart

Table 3. The detailed classification performance of three missing value imputation methods on three different dimensions

	Number of dimensions	3,888	8,888	11,288
Accuracy (%)	10-NN imputation	**76.97**	82.53	83.06
	Zero imputation	**76.97**	83.85	83.82
	Linear interpolation	**76.97**	**90.00**	**99.61**

Table 4. Comparison of three imputation method on random selected dimension

Image size	Fill missing values method	Number of dimensions	Accuracy (%)	
			6-NN	SVM
Original image with random sizes	Zero imputation	9,000	83.47	83.60
		12,636	83.90	83.41
		15,000	83.65	83.14
	10-NN imputation	9,000	82.34	83.93
		12,636	82.81	84.23
	Linear interpolation	9,000	90.78	98.52
		11,340	**99.66**	**99.94**

4. CONCLUSION

In this paper, we propose to apply several missing value imputation methods to tackle the image size of HOG descriptor. By using three missing value imputation methods, the proposed approach is efficient by clearly improving the classification performance on VNRICE image dataset. The future of this work is now continued to apply the proposed approach to other types of textual data. Since the filled value might be noised and irrelevant, we further propose to apply feature selection method to remove the noised features from imputation method.

ACKNOWLEDGEMENTS

This work was supported by Ho Chi Minh City Open University under Grant No E.2019.06.1. The authors also would like to thank Dr. Hong Phan from the Vietnam National University of Agriculture for providing the Rice Seed Image dataset.

REFERENCES

[1] J. Gomes and F. Leta,“Applications of computer vision techniques in the agriculture and food industry: A review,” *European Food Research and Technology*, vol. 235, pp. 989–1000, 2014.

[2] D. I. Patricio and R. Rieder, “Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review,” *Computers and Electronics in Agriculture*, vol. 153, pp. 69–81, 2018.

[3] A. C. Tyagi, “Towards a second green revolution,” *Irrigation and Drainage*, vol. 65, no. 4, pp. 388–389, 2016.

[4] A.Humeau-Heurtier, “Texture feature extraction methods: A survey,” *IEEE Access*, vol. 7, pp. 8975–9000, 2019.

[5] C. F. Tsai, M. L. Li, and W. C. Lin, “A class center based approach for missing value imputation,” *Knowledge-Based Systems*, vol. 151, pp. 124–135, Jul 2018.

[6] A. Sinha, S. Banerji, and C. Liu, “Novel color Gabor-LBP-PHOG (GLP) descriptors for object and scene image classification,” *Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing*, no. 58, pp. 1-8, 2012.

[7] P. Vacha, M. Haindl, and T. Suk, “Colour and rotation invariant textural features based on Markov random fields,” *Pattern Recognition Letters*, vol. 32, no. 6, pp. 771–779, Apr 2011.

[8] A. Bosch, A. Zisserman, and X. Muñoz, “Scene classification via pLSA,” *European conference on computer vision*, Springer, pp. 517–530, 2006.

[9] A. Sinha, S. Banerji, and C. Liu, “New color GPHOG descriptors for object and scene image classification,” *Machine Vision and Applications*, vol. 25, no. 2, pp. 361–375, Feb 2014.

[10] R. Khan, J. van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat, “Discriminative colordescriptors,” *Proceedings of 23th IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 2866–2873, 2013.

[11] F. Sandid and A. Douik, “Robust color texture descriptor for material recognition,” *Pattern Recognition Letters,* vol. 80, pp. 15–23, Sep 2016.

[12] T. Maenpaa, M. Pietikainen, and J. Viertola, “Separating color and pattern information for color texture discrimination,” *Object recognition supported by user interaction for service robots*, 2002.

[13] T. Maenpaa and M. Pietik¨ainen, “Classification with color and texture: jointly or separately?” *Pattern Recognition*, vol. 37, no. 8, pp. 1629–1640, Aug 2004.

[14] K. Meethongjan, M. Dzuikifli, P. K. Ree, and M. Y. Nam, “Fusion affine moment invariants and wavelet packet features selection for face verification,” *Journal of Theoretical & Applied Information Technology*, vol. 64, no. 3, pp. 606-615, 2014.

[15] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 886–893, June 2005.

[16] H. Duong and V. T. Hoang, “Dimensionality reduction based on feature selection for rice varieties recognition,” *2019 4th International Conference on Information Technology (InCIT)*, pp. 199–202, Oct 2019.

[17] P. T. T. Hong, T. T. T. Hai, L. T. Lan, V. T. Hoang, V. Hai, and T. T. Nguyen, “Comparativestudy on vision based rice seed varieties identification,” *2015 Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pp. 377–382, 2015.

[18] H. Vu, V. N. Duong, and T. T. Nguyen, “Inspecting rice seed species purity on a large dataset using geometrical and morphological features,” *Proceedings of the Ninth International Symposium on Information and Communication Technology - SoICT 2018*, pp. 321–328, 2018.

[19] B. Lurstwut and C. Pornpanomchai, “Image analysis based on color, shape and texture for rice seed (Oryza sativa L.) germination evaluation,” *Agriculture and Natural Resources*, vol. 51, no. 5, pp. 383–389, Oct 2017.

[20] H. Mebatsion, J. Paliwal, and D. Jayas, “Automatic classification of non-touching cereal grains in digital images using limited morphological and color features,” *Computers and Electronics in Agriculture*, vol. 90, pp. 99–105, Jan 2013.

[21] P. M. Szczypi´nski, A. Klepaczko, and P. Zapotoczny, “Identifying barley varieties by computer vision,” *Computers and Electronics in Agriculture*, vol. 110, pp. 1–8, Jan. 2015.

[22] A. A. Chaugule and S. N. Mali, “Identification of paddy varieties based on novel seed angle features,” *Computers and Electronics in Agriculture*, vol. 123, pp. 415–422, 2016.

[23] T. Y. Kuo, C. L. Chung, S. Y. Chen, H. A. Lin, and Y. F. Kuo, “Identifying rice grains using image analysis and sparse-representation-based classification,” *Computers and Electronics in Agriculture*, vol. 127, pp. 716–725, Sep 2016.

[24] H. li, Y. Qian, P. Cao, W. Yin, F. Dai, F. Hu, and Z. Yan, “Calculation method of surface shape feature of rice seed based on point cloud,” *Computers and Electronics in Agriculture*, vol. 142, pp. 416–423, Nov. 2017.

[25] D.P.V. Hoai, T. Surinwarangkoon, V.T. Hoang, H.-T. Duong, K. Meethongjan, “A comparative study of rice variety classification based on deep learning and hand-crafted features”, *ECTI Transactions on Computer And Information Technology*, vol. 14, no. 1, 2020.

[26] P. J. Garcıa Laencina, J. L. Sancho Gómez, and A. R. Figueiras Vidal, “Pattern classification with missing data: a review,” *Neural Computing and Applications*, vol.19, no.2, pp.263–282, Mar 2010.

[27] A. M. Sefidian and N. Daneshpour, “Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model,” *Expert Systems with Applications*, vol. 115, pp. 68–94, 2019.

[28] W. C. Lin and C. F. Tsai, “Missing value imputation: a review and analysis of the literature (2006–2017),” *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1487–1509, Feb 2019.

[29] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, “Face recognition using Histograms of Oriented Gradients,” *Pattern Recognition Letters*, vol. 32, no. 12, pp. 1598–1603, Sep 2011.

[30] H. T. M. Nhat and V. T. Hoang, “Feature fusion by using lbp, hog, gist descriptors and canonical correlation analysis for face recognition,” *2019 26th International Conference on Telecommunications (ICT)*, pp. 371–375, April 2019.

[31] A. Chauhan, D. Chauhan, and C. Rout, "Role of Gist and PHOG Features in Computer-Aided Diagnosis of Tuberculosis without Segmentation," *PloS ONE*, vol. 9, no. 11, 2014.

[32] T. J. Alhindi, S. Kalra, K. H. Ng, A. Afrin, and H. R. Tizhoosh, "Comparing LBP, HOG and Deep Features for Classification of Histopathology Images," *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro: IEEE, pp. 1–7, Jul 2018.

[33] T. Surinwarangkoon, S. Nitsuwat, and J. Elvin, "A traffic sign detection and recognition system," *International Journal of Circuits, Systems and Signal Processing*, vol. 7, no. 1, pp. 58–65, 2013.

[34] K. V. Branden and S. Verboven, "Robust data imputation," *Computational Biology and Chemistry*, vol. 33, no. 1, pp. 7–13, Feb 2009.

[35] L. A. Hunt, "Missing Data Imputation and Its Effect on the Accuracy of Classification,", in: F. Palumbo, A. Montanari, and M. Vichi, "Studies in Classification, Data Analysis, and Knowledge Organization," *Data Science*, Cham: Springer International Publishing, pp. 3–14, 2017.

BIOGRAPHIES OF AUTHORS

Huy Nguyen-Quoc is an undergraduate student from Ho Chi Minh City Open University, Vietnam. His research interests include pattern recognition and image analysis.

Vinh Truong Hoang is an assistant professor and Head of Image Processing and Computer Graphics Department at the Ho Chi Minh City Open University, Vietnam. His research interests include image analysis and feature selection.