GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

ISSN 2545-4803 on line

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI)

YEAR 2020

VOLUME III, Number 2

GOCE DELCEV UNIVERSITY - STIP, REPUBLIC OF NORTH MACEDONIA
FACULTY OF COMPUTER SCIENCE

ISSN 2545-4803 on line

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS

(BJAMI)

YEAR 2020 VOLUME III, Number 2

AIMS AND SCOPE:
BJAMI publishes original research articles in the areas of applied mathematics and informatics.

Topics:
1. Computer science;
2. Computer and software engineering;
3. Information technology;
4. Computer security;
5. Electrical engineering;
6. Telecommunication;
7. Mathematics and its applications;
8. Articles of interdisciplinary of computer and information sciences with education, economics, environmental, health, and engineering.

Address of the editorial office
Goce Delcev University – Stip
Faculty of philology
Krste Misirkov 10-A
PO box 201, 2000 Štip,
Republic of North Macedonia

BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI), Vol 3

ISSN 2545-4803 on line
Vol. 3, No. 2, Year 2020

EDITORIAL BOARD

CONTENT

UDC: 004.932.75:519.614]:003.324.1

EIGEN SPACE BASED CHARACTER RECOGNITION APPLICATION FOR LEARNING CHINESE HANDWRITING IN PALEMBANG

SHINTA PUSPASARI AND HENKY HONGGO

Abstract. Eigen space model was proposed in this study for Chinese Handwriting Recognition. The principle component on image extracted, namely eigen vector and eigen value. These features will be used to recognize each character that is written in real time by using some kind of an input device. An application system is developed for learning Chinese characters which is designed to be able to recognize handwritten characters, pronounce, and translate it into Indonesian. The system was designed for beginner in Chinese handwriting learning and it only recognizes the basic of Chinese characters. Images of Chinese handwriting used as the input and database contained 265 binary images for feature extraction. The feature extraction and the feature matching algorithms were designed based on Eigen Space Model and Euclidean Distance Method. This application used a canvas painted as an input media that made users able to perform a direct input in real time. The system performs recognition using the proposed method to extract the image of Chinese handwriting with a characteristic value of the tested image and the image of 5 samples per Chinese handwriting image data and the average level of Chinese handwriting recognition percentage reached up to 70%. The writing of Chinese characters follows a certain pattern and sequence, differing only slightly at the time of writing causing different characters to be recognized by PCA. In term of effectiveness, the system is also evaluated by user. The result shows that it is a promising tool for learning Chinese handwriting.

1. Introduction

Palembang is the center of the Srivijaya kingdom, which was the largest kingdom in the archipelago in the past and whose cooperation reached the Chinese territory so that ethnic Chinese also became part of the Palembang society. Chinese language helps the daily language of some ethnic Chinese communities in Palembang. Mandarin is one of the many international languages learned by all ages, from children to adults, which is more than 1 billion people speaking and knowing Mandarin. Mandarin has differences with the English and Indonesian language in the way of writing and pronunciation. How to write Chinese characters is almost the same as making calligraphic strokes; they have indentations and are very different at the time of writing the alphabet letters, like the letter *A* for example. Thus, studying Chinese characters has a more complex level than learning English and Indonesian, especially in the way of writing and pronunciation with the tones or pinyin.

In the field of language education, Chinese characters have become one of the subjects at school level from elementary school through high school; the difficulties faced by people who just learn Chinese characters are located at the time to understand and master how to write Chinese handwriting correctly and how to pronounce the Chinese characters with or without pinyin correctly [1].

To learn Chinese handwriting can be an advantage as Mandarin is an international language. Technology in the form of a Chinese handwriting learning application is designed to assist users in learning the meaning, writing, and pronunciation of Chinese

handwriting. In the application that is designed, there are several methods that can be used for character recognition; the methods include neural networks, fuzzy logic, as well as principal component analysis [2]. The Method of Principal Component Analysis based on Euclidean space has been widely used as a character recognition method; the basic principle of the method of Eigen Space is using a principle component of the data (images) and maintaining the characteristics of the data set, namely Eigen values and Eigen vectors [3].

These features have proven their effectiveness for classification, face image classification for example, namely Eigenface [4]. Eigenface is a method for extracting features that are very successfully applied in the calculation of digital imagery for face recognition [5]. Eigen space has proven its effectiveness for character recognition [6-10].

This study tried to implement the Eigen space model for character recognition in an application for learning Chinese characters. The character is processed as an image which is input through a canvas media on the application using an input media such as a mouse or a digital pen. For extraction and testing, this study used a data set containing 53 Chinese characters on 256 binary images. The result shows that the application is adequate for learning Chinese characters based on its character recognition accuracy.

The rest of this paper is organized as follows: section 2 describes the Eigen space model, section 3 illustrates the Eigen space-based approach for character recognition, and section 4 describes the experimental result of this study. Finally, at the end of the paper we present the conclusions of this study.

2. Eigen Space Model

Eigen space is a space dimension of the mapping of a particular characteristic. The term itself is used at the beginning of Eigen values of the 20th century by a German scientist David Hilbert. Eigen vectors derived from the German language and can be interpreted as a characteristic or critical value. Eigen space for face recognition (Eigenface) is used to reduce the dimension of a high-dimensional face image into the space of lower facial features [11]. So, the time required for computing the process becomes much shorter, especially when it is used to process the collection of a large data base. Due to these characteristics, Eigen space is widely used in various engineering applications, biometric, and many others [12-19]. Eigen space obtained by processing the Eigenvectors and Eigen values contains a certain characteristic value. Eigen space mapping results are sets of Eigenvectors and have the same Eigen values [20]. In Eigen space, the character recognition process will be carried out, due to contain all relevant information required for matching.

2.1 Eigen Values and Eigen Vectors

Eigen vector is the set of vectors commonly referred to as a feature vector of a linear system of equations [21]. Linear transformation performed on a system of linear equations will not affect the Eigen vectors and Eigen values of the system. Eigen vector direction will not change even if it is transformed.

2.2 Eigen Space Model

An effective method for similarity measure between images is using Euclidean Distance. The closer the distance, the more similar it is. An image can be represented as *n*-dimensions space. The similarity between two images, *p* and *q*, could be defined as the distance between two points in it and is notated by *D(p,q)*. An image *u* will be more similar to image *v,* then *w* if *D(u,v) < D(u,w)* [10].

In *n*-dimensions Euclidean space, R^n, the distance between the points *x* and *y* can be formulized as follows, where *n* is the number of its points,

$$D(x,y) = |x-y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \qquad (2.1)$$

3. Character Recognition

3.1 Feature Extraction Phase

1. Creating a matrix representation of the image *I*.
   An image of size *W* x *H* (width x height) represented in the form of a matrix that contains the elements of color intensity values in an image pixel. The matrix can be represented in an *n*-dimensional vector which is also called a vector image of x_i, where *n* is *WxH*, and *i* is the index of the image. *I* is the image matrix size *Nxn* matrix whose each row contains the image vectors x_i, where *N* is the number of images. Matrix representation of *I* is as follows,

$$I = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & .. & x_{1n} \\ x_{21} & x_{22} & x_{23} & .. & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdot\cdot & x_{3n} \\ .. & .. & .. & .. & .. \\ x_{N1} & x_{N2} & x_{N3} & \cdot\cdot & x_{Nn} \end{bmatrix} \qquad (3.1)$$

2. Let us create the mean matrix of *I*, μ
   $$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (3.2)$$
   by the above formula, this matrix will be obtained,
   $$\mu = [\mu_1 \quad \mu_2 \quad \mu_3 \quad \cdot\cdot \quad \mu_n] \qquad (3.3)$$
3. Let us create the *A* Matrix
   The difference matrix (*A*) is created by reducing each matrix element *I* in the *i*-th column with μ_i. *A* matrix representation is as follows,

$$A = \begin{bmatrix} x_{11} - \mu_1 & x_{12} - \mu_2 & x_{13} - \mu_3 & .. & x_{1n} - \mu_n \\ x_{21} - \mu_1 & x_{22} - \mu_2 & x_{23} - \mu_3 & .. & x_{2n} - \mu_n \\ x_{31} - \mu_1 & x_{32} - \mu_2 & x_{33} - \mu_3 & \cdot\cdot & x_{3n} - \mu_n \\ .. & .. & .. & .. & .. \\ x_{N1} - \mu_1 & x_{N2} - \mu_2 & x_{N3} - \mu_3 & \cdot\cdot & x_{Nn} - \mu_n \end{bmatrix} \qquad (3.4)$$

4. Find the covariance matrix using the following equation,
   $$C = A^T * A \qquad (3.5)$$
   where *C* is a matrix of dimension *n* x *n*.

5. Finding Eigen values and Eigen vectors of the covariance matrix using the following equation,

$$C\,v = \lambda\,v \qquad (3.6)$$

where v is the Eigen vector and λ is the Eigen value.

6. Features extraction of images dataset to perform the multiplication process of extracting the image made the difference between the matrix A and Eigen vector v.

3.2 Testing Phase

1. On the image to be tested, do the steps that are almost the same as the extraction phase, namely make the matrix representation of the same test images as the first step.
2. Having obtained the matrix of test images, the difference will be a new matrix U, namely the difference between the matrix of test images and the average matrix of the image matrix I (the results of the second step).
3. Having obtained the matrix of the difference U, the matrix of test images and the average matrix of the image matrix I (the results of the second step), the image extraction process will be done by matrix multiplication of the difference between U and the Eigen vectors (results from the fifth step), so the matrix will obtain the test images (image Extraction test).
4. Find the shortest distance using the equation (2.1) Euclidean distance. Having obtained the matrix image of the sample and the matrix of test images (Image Extraction Test), we will look for the distance between the image of the sample matrix (Sample Image Extraction) and the matrix of test images (Image Extraction Test) using the equation Euclidean distance. The smaller the distance between the image of the sample and the test images, the higher the level of similarity is.

4. Experimental Results

The application for learning Chinese characters is built using Visual Basic Programming Language for its interface and using MatLab library for image processing. The dataset is used while testing is as follows,

1. The size of the sample and the test image is 80 x 80 pixels.
2. The dataset sample consists of 265 sample images and can be added.
3. It contains 53 Chinese characters and consists of five image samples for each.
4. The sample and test images are in binary format.

The following Figure 1 illustrates the user interface of the learning application in Indonesian because the app was designed for Indonesians for learning Chinese characters. The application is in Indonesian because it is developed specially for Indonesians because there are many students who speak and learn Chinese characters. The first button is for Chinese handwriting recognition, the second one for Adding New Sample Character in order to improve the recognition accuracy, and the last button is for closing the application.

Chinese Character Learning user interface shown in Figure 1 contained a functioning combo box to select a Chinese characters character to be tested, and five picture boxes to display the Chinese characters character sample selected by the application user.

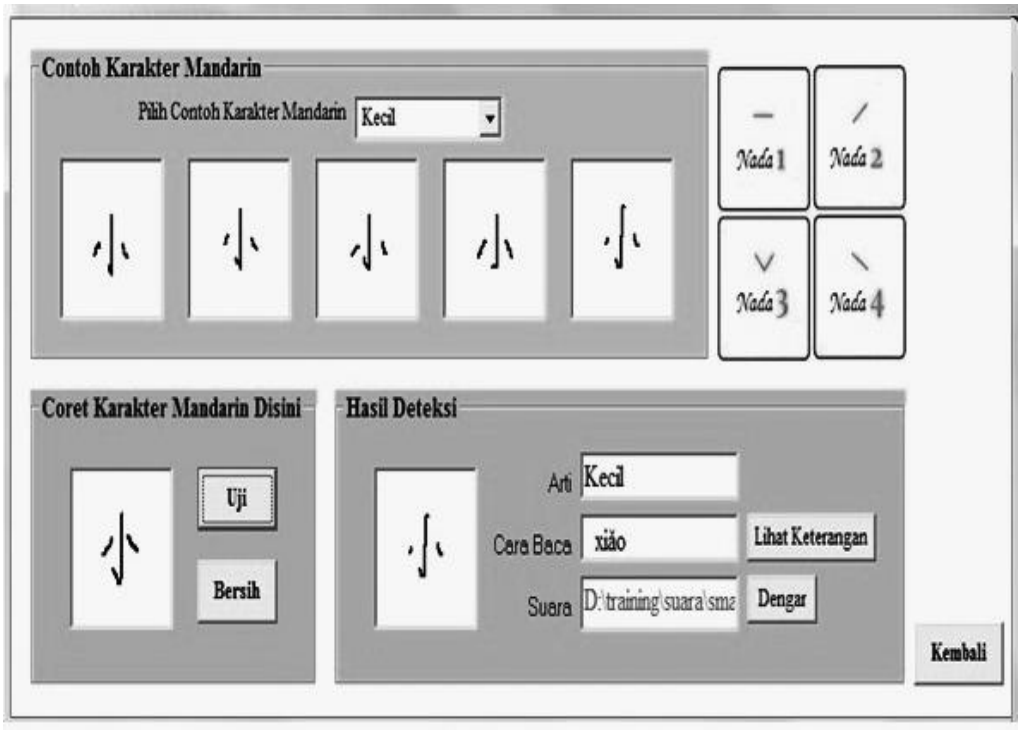

Figure 1. Chinese Character Learning Application User Interface

On the Character Recognition user interface as in Figure 1, there is a picture box canvas measuring 80 x 80 pixels which can be crossed out. To perform detection of the character, the user can write off a Chinese character that will be recognized as input. Then, when the user presses the net, the image on the canvas will be cleaned again. Once the user selects the test button, the image on the canvas will be stored directly and tested with Matlab. When running the application to test a character for the first time, the system will make the process of loading data in the form of the results of calculations using Eigen values and vectors. This process is only done on the first testing, as the next testing does not fit the dataset anymore. Detection results show the results of handwriting recognition such as its meaning, how to read it, sounds, and images that are most similar to the Chinese characters character input.

Calculated from Table 1, it was found that the average percentage of detection results is up to 70%, from 10 Chinese characters characters; each character performed the test 10 times. The writing technique is influenced by the style and manner of writing. The more strokes on the character, the higher the bias that appears at recognition because of the similarities between the characters. Table 2 describes the analysis of the results of character detection, pronounced as xiǎo, based on the Euclidean distance. At first and second trial, PCA failed to detect the written character as xiǎo for the stroke is closer to èr and shàng.

Table 1. *Result of 10 Times of Ten Characters Testing*

No.	Character	Pinyin	Accuracy
1.		yī	100%
2.		xiǎo	80%
3.		rén	90%
4.		dà	90%
5.		shí	90%
6.		shuǐ	50%
7.		tóu	50%
8.		tǔ	90%
9.		zuǒ	40%
10.		mǎ	40%

The writing of Chinese characters follows a certain pattern and sequence, differing only slightly at the time of writing, thus causing different characters recognized by PCA. In user testing for obtaining user feedback regarding the interface and functionality of the application, a questionnaire was given to the user to assess as well as provide input on programs that had been made. The result shows that this application can help in learning Chinese characters. This character recognition computer application based on Euclidean space is useful in learning Chinese characters, it has an attractive appearance, and it is easy to use. The results can be improved by adding more samples to the dataset. Misclassification also happened because the user could not draw the character correctly, being beginner in learning Chinese characters.

Table 2. *Euclidean Distance Measure*

No.	Character xiao	Top 10 Euclidean Distance Measure
1		**12.0416 = xiǎo** 12.8841 = yī 12.9615 = yī 13.1149 = yī 13.2665 = yī 13.3417 = yī 13.8203 = shí 13.8203 = bā 13.9642 = shí 14.0000 = èr
2		**11.9583 = xiǎo** **12.6491 = xiǎo** 12.9615 = yī 12.9615 = yī 13.4164 = yī 13.4536 = shàng 13.5277 = xià 13.5647 = yī 13.6015 = shí 13.7113 = shàng
3		13.2288 = shí 13.4907 = yī 13.5277 = shí 13.5647 = yī 13.6382 = yī 13.7113 = yī 13.7840 = yī 13.8924 = xià 13.9642 = shí **14.2478 = xiǎo**

5. Conclusions

This study tried to implement the Eigen space method for character recognition in a computer application system for learning Chinese handwriting. Eigen space mapping results are sets of the Eigen vectors that have the same Eigen values. The Eigen space character recognition process will be carried out due to it containing all relevant information required for matching. The experimental result shows that the Eigen space model is effective for Chinese handwriting recognition for a learning application, with the average percentage of detection of 70% using a data set for extraction and testing; it contained 53 Chinese characters on 256 binary images. The writing technique is influenced by the style and manner of writing. The more strokes on the character, the higher the bias might appear at recognition because of the similarities between characters. The writing of Chinese characters follows a certain pattern and sequence, differing only slightly at the time of writing, thus causing different characters recognized by PCA. In terms of effectiveness, the application is also evaluated by users. The result shows that this application is adequate for learning Chinese handwriting, it has an attractive appearance, and it is easy to use especially by beginners in Chinese characters learning. It is a promising tool for learning Chinese handwriting in future.

References

[1] K. Y. FUNG and K. C. FUNG, "HCI Technology with Mastery Learning Approach for Children Learning Chinese Characters Writing in Hong Kong," 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), Tartu, Estonia, 2020, pp. 226-227, doi: 10.1109/ICALT49669.2020.00073.

[2] B. M. Vinjit, M. K. Bhojak, S. Kumar and G. Chalak, "A Review on Handwritten Character Recognition Methods and Techniques," 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 1224-1228, doi: 10.1109/ICCSP48568.2020.9182129.

[3] Pentland, B. Moghaddam, T. Starner, *"View-Based and Modular Eigenspaces for Face Recognition"*, Proceedings of the IEEE Computer Society International Conference on Computer Vision and PatternRecognition, 1994.

[4] J.R. del solar, and P. Navarrete, *"Eigenspace-Based Face Recognition: A Comparative Study of Different Approaches"*, IEEE Transaction on System, Man and Cybernetic, Vol. 16, No. 7, hal. 817-830. 2005.

[5] M.A. Turk and A.P. Pentland, "*Face Recognition Using EigenFaces"*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, hal. 586-591, 1991.

[6] P. K. Sharma, M. Deori, B. Kaur, C. Dey and K. Das, "Implementation of ANN based Assamese Character recognition system with reinforcement of Radon Transform and PCA based feature extraction," *2013 1st International Conference on Emerging Trends and Applications in Computer Science*, Shillong, 2013, pp. 167-171, doi: 10.1109/ICETACS.2013.6691416.

[7] A. K. Verma, N. Verma, S. V. Singh and R. Sharma, "Robust script identification using Wavelet and EHD features in PCA space," 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, 2016, pp. 1-6, doi: 10.1109/ICRAIE.2016.7939585.

[8] G. Keerthi Prasad, I. Khan, N. R. Chanukotimath and F. Khan, "On-line handwritten character recognition system for Kannada using Principal Component Analysis Approach: For handheld devices," 2012 World Congress on Information and Communication Technologies, Trivandrum, 2012, pp. 675-678, doi: 10.1109/WICT.2012.6409161.

[9] A. B. Jagtap and R. S. Hegadi, "Offline Handwritten Signature Recognition Based on Upper and Lower Envelope Using Eigen Values," 2017 World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, 2017, pp. 223-226, doi: 10.1109/WCCCT.2016.62.

[10] D. Nasien, F. Candra, Delsavonita, D. Yulianti, R. R. Andhi and M. H. Adiya, "Off-line Handwritten Korean Letter using Principle Component Analysis and Back Propagation Neural Network," 2018 2nd International Conference on Electrical Engineering and Informatics (ICon EEI), Batam, Indonesia, 2018, pp. 72-74, doi: 10.1109/ICon-EEI.2018.8784136.
[11] M. Kocaleva and V. Gicev, "EIGENVALUES AND EIGENVECTORS OF A BUILDING MODEL AS A ONE-DIMENSIONAL ELEMENT", *BJAMI*, vol. 3, no. 1, pp. pp. 43-54, Jun. 2020.
[12] M. A. Hussain, "Eigenspace based accurate iris recognition system," 2010 Annual IEEE India Conference (INDICON), Kolkata, 2010, pp. 1-3, doi: 10.1109/INDCON.2010.5712679.
[13] F. Alonso-Fernandez, R. A. Farrugia and J. Bigun, "Eigen-patch iris super-resolution for iris recognition improvement," 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, 2015, pp. 76-80, doi: 10.1109/EUSIPCO.2015.7362348.
[14] R. Kumar, A. Sreeram, A. Purushothaman and S. Ganapathy, "Unsupervised Neural Mask Estimator for Generalized Eigen-Value Beamforming Based Asr," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7494-7498, doi: 10.1109/ICASSP40776.2020.9054550.
[15] S. Kaushik and R. Singh, "A new hybrid approach for palm print recognition in PCA based palm print recognition system," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2016, pp. 239-244, doi: 10.1109/ICRITO.2016.7784958.
[16] M. B. S. Divya and N. B. Prajwala, "Facial Expression Recognition by Calculating Euclidian Distance for Eigen Faces Using PCA," 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 0244-0248, doi: 10.1109/ICCSP.2018.8524332.
[17] Husnawati and S. Puspasari, "Principal Component Analysis Method for Image Segmentation of Smart Meter Digital", Proceeding of 4th Forum in Research, Science, and Technology International Conference, Indonesia, 2020.
[18] V. Singh Tomar, N. Gupta and U. Singh, "Expressions Recognition Based on Human Face," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 590-595, doi: 10.1109/ICCMC.2019.8819714.
[19] S. Mukhopadhyay and S. Sharma, "Real Time Facial Expression and Emotion Recognition using Eigen faces, LBPH and Fisher Algorithms," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 212-220, doi: 10.1109/Confluence47617.2020.9057985.
[20] S.L. Eddins, R. Woods, and R. Gonzales. *Digital Image Processing Using Matlab*, Prentice Hall. 2004.
[21] K. Nishino, Y. Sato and K. Ikeuchi, "Eigen-texture method: Appearance compression based on 3D model," Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 1999, pp. 618-624 Vol. 1, doi: 10.1109/CVPR.1999.787003.

Shinta Puspasari
Universitas Indo Global Mandiri
Department of Informatics,
Faculty of Computer Science,
Jalan Jenderal Sudirman 629, Palembang, Indonesia
Email address: shinta@uigm.ac.id

Henky Honggo
Zao An Rubric Contributor of Tribun Sumsel
Palembang Indonesia
Email address: henkyhonggo@gmail.com