Hybrid principal component analysis and K-nearest neighbour to detect the catfish disease.pdf

by

Submission date: 03-Feb-2022 01:21PM (UTC+0700)

Submission ID: 1753997473

File name: Hybrid principal component analysis and K-nearest neighbour to detect the catfish disease.pdf (784.94K)

Word count: 2828

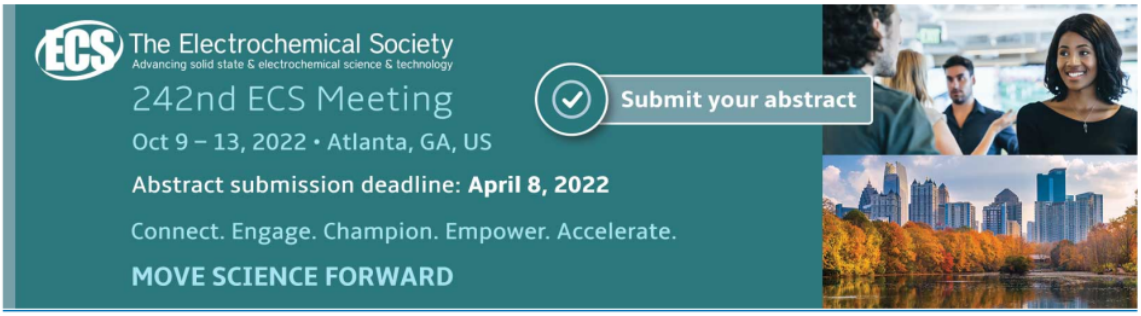
Character count: 15276

Journal of Physics: Conference Series

PAPER · OPEN ACCESS

Hybrid principal component analysis and K-nearest neighbour to detect the catfish disease

To cite this article: D S Maylawati *et al* 2019 *J. Phys.: Conf. Ser.* **1402** 077008



Hybrid principal component analysis and K-nearest neighbour to detect the catfish disease

D S Maylawati[1,2,*], R Andrian[1], S Sunarto[1], M Wildanuddin[1] and A Wahana[1]

[1] Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia
[2] Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia

*diansm@uinsgd.ac.id

Abstract. Catfish cultivation in Indonesia is a very promising business opportunity with a big profits. Every year, market demand continues to increase. However, this is contrary with the lack of catfish farmer's knowledge, so that catfish yields are not optimal. This is because, for certain types of catfish such as Sangkuriang catfish, it is easy to contract certain diseases. This study aims to create an automated system that capable of detecting catfish disease based on its symptoms with image recognition techniques. Early detection of catfish disease can help to find out the causes and prevention, so that the yield remains optimally. The method that used in this study is Principal Component Analysis (PCA) for feature extraction in images combined with K-Nearest Neighbour (KNN) with Euclidean Distance to classify catfish diseases among others: white spots, edema (abdominal swelling), jaundice, and bent spinal disease (scoliosis and lordosis). Based on the results of the experiment using 30 images data for training and 20 images data for testing, 18 image data is classified correctly. This result proves that PCA and KNN able to detect catfish disease well with percentage of accuracy around 90%.

1. Introduction

Catfish cultivation in Indonesia is a very promising business opportunity with a big profits. Every year, market demand continues to increase. Based on the Directorate General of Aquaculture of the Ministry of Maritime Affairs and Fisheries, in 2017, the catfish production reached 1.8 million tons or increase 131.7% from the previous year's achievement. Moreover in several big city in Indonesia, the demand of catfish increase up to 30% in the last 2 years [1]. However, this is contrary with the lack of catfish farmers knowledge, so that catfish yields are not optimal [2,3]. This is because, for certain types of catfish such as Sangkuriang catfish, it is easy to contract certain diseases [4].

In the digital era today, utilization of technology is very useful, also in Image Recognition research. Image recognition is a process to identify and detect an object or feature that contained in digital resource especially image and video [5–7]. There are many methods and algorithms for image segmentation, detection, processing, and recognition, such as Data-Driven Markov Chain Monte Carlo (DDMCMC) algorithm [8], Otsu algorithm for image segmentation [9], Canonical Correlation Analysis [10], McCall's method [11], Discrete Cosine Transform (DCT) [12,13], Discrete Wavelet Transform (DWT) [14–16], Huffman method [12,13], and the most common image processing algorithm is Principal Component Analysis (PCA) [5,6,8,17]. Moreover, in the development of image recognition and detection research, there are many data mining algorithms that are combined with image processing

method, such as Artificial Neural Network (ANN) [18–20], K-Nearest Neighbour (KNN) [21,22], Fuzzy C-means [15,23] and also Deep Learning (Deep Neural Network) [24–26] that popular today.

Therefore, based on the facts and problems above, this study aims to create an automated system that capable of detecting catfish disease based on its symptoms with image recognition techniques. Early detection of catfish disease can help to find out the causes and prevention, so that the yield remains optimally. This system implements PCA and KNN algorithm. PCA is a common and simple algorithm for image processing, and also can decompose a large number of variables that correlate to a small number of uncorrelated variables, and can reduce high dimensions to lower dimensions without losing important information on the data [27]. While, KNN is one classification algorithm that can be used easily, effectively, and efficiently.

2. Methodology

The method that used in this study is Principal Component Analysis (PCA) for feature extraction in images combined with K-Nearest Neighbour (KNN) with Euclidean Distance to classify catfish diseases.

2.1. Principal Component Analysis (PCA)

The PCA represents an image in a feature vector, for example I is a matrix representation of the image with N1xN2 characteristic vectors of image I that obtained by stringing rows into one row vector. The second line is arranged after the first row, followed by the third row and so on until there is only one line. So, the dimensions of the feature vector are the number of pixels in the image, N1 and N2. The goal is to reduce the dimensions of the feature vector. The PCA algorithm that illustrated in Figure 1 has a several steps as follow [28]:

- Normalize matrix X, with k = X_k / length (X_k), where X_k is the k-column vector of X
- Build a covariance matrix with W = XT * X
- Calculate the eigenvalue (E) and Eigenvector (Q) from the covariance matrix W
- Sort the eigenvalues (E) and Eigenvectors (Q) of small magnitude based on the order of eigenvalues.
- Calculate the principal component P = X * Q * E

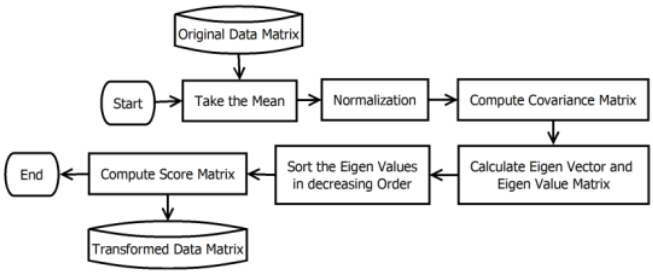


Figure 1. PCA algorithm flowchart.

2.2. K-Nearest Neighbour (KNN)

KNN is a classification method in which there is new data that is not yet had the class. Then, the class or the label will selected from the data for the number of *k* that closest to the new data [29]. The highest class of data closest to the number of *k* is selected as the predicted class for the new data. The usual KNN method is used by researchers as a comparison of the methods developed. Value of *k* is usually determined in an odd number to avoid the appearance of the same distance in the classification process. Figure 2 is a flowchart diagram that describes KNN algorithm. The nearest value is produced by calculating similarity / dissimilarity distance. There are many distance metrics calculation methods,

among others: Euclidean Distance, Square Euclidean Distance, Mahalonobis Distance, Manhattan distance, Elastic Distance and so on [30–33]. KNN is a supervised algorithm because it uses labels or targets in the classification process. KNN will choose training data as a reference in the process of finding distance between data.

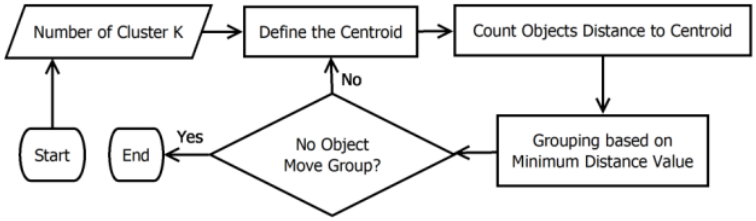


Figure 2. KNN algorithm flowchart.

3. Result and discussion

3.1. Analysis of PCA algorithm

Main process of PCA is to extract image features and transform into vector matrices. Before conducting the main process, images will be conducted pre-process, such as grey scaling, cropping, and resizing that illustrated in Figure 3 (a) and (b). Then, based on PCA algorithm, the average of each row matrices vector value will be counted with formula (1). Next, ZeroMeans (formula (2)), Covariant (formula (3)), and Eigen vector value (formula (4)) will be calculated. Last process, it will be found the projection matrix value based on Eigen vector value. This projection matrix is the result of feature extraction from the entire image data, so that the next process that is processed is only this projection matrix which has a much smaller size compared to the vector matrix of all previous images.

$$\mu_i = \frac{(x_{1,i+x_{2,i}+\ldots+x_{m,i}})}{m} \quad (1)$$

$$ZeroMean = MatricVector - MeanMatricVector \quad (2)$$

$$c(x_i, x_i) = \sqrt{\frac{\sum_{j-1}^{m}(x_{j,i}-\mu_i)(x_{j,i}-\mu_i)}{m-1}} \quad (3)$$

$$\lambda\,\vartheta = C\,\vartheta \quad (4)$$

(a) (b)

Figure 3. (a) Catfish image before pre-processing; (b) Catfish image after pre-processing.

3.2. Analysis of KNN algorithm for catfish disease

Based on projection matrix value of image that resulted from PCA algorithm, KNN will classify the image with calculate Euclidean Distance value. The result of Euclidean Distance will be reduced by the projection matrix and the results will be squared. The catfish disease will be classified into 4 classes, among others: white spots, edema (abdominal swelling), jaundice, and bent spinal disease (scoliosis and lordosis).

3.3. Experiment result

The experiment is conducted using 30 images of training data (Figure 4) and 20 images of testing data (Figure 5). The result of experiment showed that 18 images of testing data are classified correctly and 2 image of testing data are classified in wrong class. The result of experiment is provided in Table 1.

Figure 4. Image of catfish for training data.

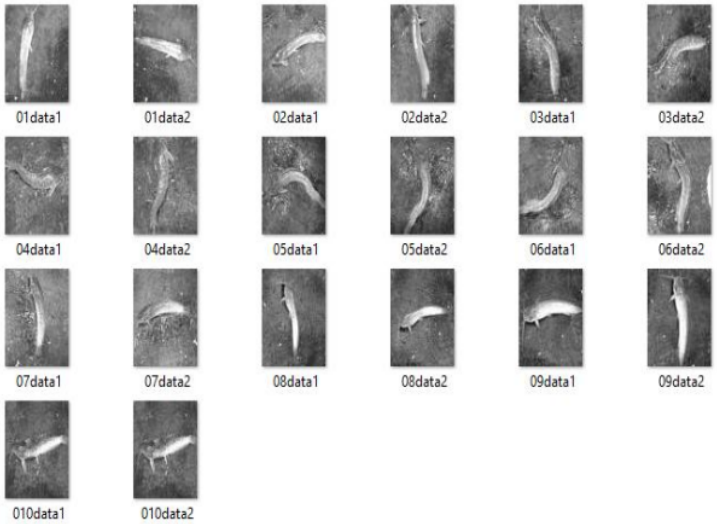


Figure 5. Image of catfish for testing data.

Table 1. Testing result of catfish disease detection.

No	Data	Diseases	Prediction Result	Decision
1	01data1	White Spots	White Spots	Correct
2	01data2	White Spots	White Spots	Correct
3	02data1	White Spots	White Spots	Correct
4	02data2	White Spots	White Spots	Correct
5	03data1	*Edema*	White Spots	Not correct
6	03data2	*Edema*	*Edema*	Correct
7	04data1	*Edema*	*Edema*	Correct
8	04data2	*Edema*	*Edema*	Correct
9	05data1	Jaundice	Jaundice	Correct
10	05data2	Jaundice	Jaundice	Correct
11	06data1	Jaundice	Jaundice	Correct
12	06data2	Jaundice	*Edema*	Not correct
13	07data1	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct
14	07data2	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct
15	08data1	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct
16	08data2	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct
17	09data1	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct
18	09data2	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct
19	010data1	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct
20	010data2	*Scoliosis/Lordosis*	*Scoliosis/Lordosis*	Correct

4. Conclusion

Early detection of disease in catfish can reduce the risk of large crop failure. This research uses the recognition technique using PCA and KNN to classify the disease of catfish. Based on the experiment result that has been done, in the process of extracting traits to detect the type of catfish disease that uses the PCA (Principal Component Analysis) method in image processing, it can reduce image data that is quite large. As well as in the classification of data it has been effective in classifying image data of catfish. Then, in the classification process using the KNN method, it has been around 90% effective in determining the type of catfish disease. For further research, not only on catfish, but can also detect diseases that may arise in other types of fish. This can support the creation of increased yields in the fisheries sector. And also, this research can be developed using deep learning that popular in big data era.

Acknowledgments

We would like to say thank you for Research and Publication Centre of UIN Sunan Gunung Djati Bandung that give the full support and fund for this publication research.

References

[1] Sari S M 2018 *Permintaan Tinggi, Produksi Lele Melesat Lebih 2 Kali Lipat* [Online] Retrieved from: http://industri.bisnis.com/read/20180129/99/731845/permintaan-tinggi-produksi-lele-melesat-lebih-2-kali-lipat Accessed on 16-Oct-2018

[2] Yana Y 2016 *Manfaat Ikan Lele Untuk Kesehatan* [Online] Retrieved from: http://manfaat.co.id/6-manfaat-ikan-lele-untuk-kesehatan Accessed on 04-Aug-2016

[3] D SP 2013 *Budi Daya Lele* (Jakarta: PT Agro Media Pustaka)

[4] Rustandi 2014 *Budidaya Lele Sangkuriang* [Online] Retrieved from: http://penyuluhpi.blogspot.co.id/2014/06/budidaya-lele-sangkuriang.html Accessed on: 04-Aug-2016

[5] Porter T and Bow S T 2007 Pattern Recognition and Image Preprocessing *Math. Gaz.*

[6] Pratiwi D E and Harjoko A 2013, "Implementasi Pengenalan Wajah Menggunakan PCA (

Principal Component Analysis)," *Ijeis*, 2013.
[7] Taofik A, Ismail N, Gerhana N A, Komarujaman K and Ramdhani M A 2018 Design of Smart System to Detect Ripeness of Tomato and Chili with New Approach in Data Acquisition *IOP Conference Series: Materials Science and Engineering* **288** 1 012018
[8] Tu Z, Chen X, Yuille A L and Zhu S C 2005 Image parsing: Unifying segmentation, detection, and recognition *Int. J. Comput. Vis.*
[9] Vala M H J and Baxi A 2013 A Review on Otsu Image Segmentation Algorithm *Int. J. Adv. Res. Comput. Eng. Technol.*
[10] Sen Sun Q, Zeng S G, Liu Y, Heng P A and Xia D S 2005 A new method of feature fusion and its application in image recognition *Pattern Recognit.*
[11] Sugiantoro B 2017 Analisa Usabilitas Sistem Deteksi Akses Pornografi Pengguna Internet Menggunakan Metode Mccall'S *J. Online Inform.*
[12] Hardianto R H, Novamizanti L and Prasasti A L 2017 Implementasi dan Analisis Kompresi Hybrid pada Citra Medis Digital Hasil Rontgen Kanker Payudara *J. Terap. Teknol. Inf.*
[13] Krasmala R, Budimansyah A and Lenggana U T 2017 Kompresi Citra Dengan Menggabungkan Metode Discrete Cosine Transform (DCT) dan Algoritma Huffman *J. Online Inform.*
[14] Chandra J and Wahono R S 2015 Integrasi Discrete Wavelet Transform dan Singular Value Decomposition pada Watermark ing Citra untuk Perlindungan Hak Cipta *J. Intell. Syst.*
[15] Pratama A A, Suciati N and D. Purwitasari, "Implementasi Fuzzy C-Means untuk Pengelompokan Citra Batik Berdasarkan Motif dengan Fitur Tekstur," *J. Tek. POMITS* **1** 1
[16] Faza A M, Slamet C and Nursantika D 2018 Analisis Kinerja Kompresi Citra Digital dengan Komparasi DWT, DCT dan Hybrid (DWT-DCT) *J. Online Inform.*
[17] Okfan Rizal Ferdiansyah A A and Ema Utami 2015 Implementasi Principal Component Analysis Untuk Sistem Temu Balik Citra Digital *Creat. Inf. Technol. J.* **2** 3
[18] Effendy N and Imanto R 2008 Deteksi Pornografi Pada Citra Digital Menggunakan Pengolahan Citra dan Jaringan Syaraf Tiruan *Proc. Natl. Conf. Inf. Technol. Res.*
[19] Setiawan A F and Agung A K 2018 Klasifikasi Pola Sidik Jari Menggunakan Jaringan Syaraf Tiruan Backpropagation Untuk Analisa Karakteristik Seseorang *ANTIVIRUS J. Ilm. Tek. Inform.*
[20] Gustina S, Fadlil A and Umar R 2016 Identifikasi Tanaman Kamboja menggunakan Ekstraksi Ciri Citra Daun dan Jaringan Syaraf Tiruan *Annual Research Seminar*
[21] Syahid D, Jumadi and Nursantika D 2016 Sistem Klasifikasi Jenis Tanaman Hias Daun Philodendron Menggunakan Metode K-Nearest Neighboor (KNN) Berdasarkan Nilai Hue Saturation Value (HSC) *J. Online Inform.* **1** 1 20–23
[22] Gerhana Y A, Zulfikar W B, Ramdani A H and Ramdhani M A 2018 Implementation of Nearest Neighbor using HSV to Identify Skin Disease *IOP Conf. Ser. Mater. Sci. Eng.* **288** 1 012153
[23] Bezdek J C, Ehrlich R and Full W 1984 FCM: The fuzzy c-means clustering algorithm *Comput. Geosci.* **10** 2–3 191–203
[24] Ciregan D, Meier U and Schmidhuber J 2012 Multi-column deep neural networks for image classification *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*
[25] Technologii C H, Poc S and Multime G A 2013 Deep Learning for Image Recognition *arxiv.org*
[26] Kalash M, Rochan M, Mohammed N, Bruce N D B, Wang Y and Iqbal F 2018 Malware Classification with Deep Convolutional Neural Networks *2018 9th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2018 - Proceedings*
[27] Fakhrina F A, Suprapto and Nurwasito H 2016 *Perancangan Aplikasi Pengolahan Citra Digital Untuk Mendeteksi Hama Dan Penyakit Pada Tanaman Jagung*
[28] Mauridhi H P and Muntasa A 2010 *Konsep Pengolahan Citra Digital dan Ekstraksi Fitur* (Yogyakarta: Graha Ilmu)
[29] Yeni K and Syafa'ah N 2014 Sistem Pendukung Keputusan Untuk Menentukan Jurusan Pada Siswa Sma Menggunakan Metode Knn Dan Smart *JSII*

[30] Zhang S 2012 Nearest neighbor selection for iteratively kNN imputation *J. Syst. Softw.*
[31] Cai P, Wang Y, Lu G, Chen P, Ding C and Sun J 2016 A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting *Transp. Res. Part C Emerg. Technol.*
[32] Geler Z, Kurbalija V, Radovanović M and Ivanović M 2016 Comparison of different weighting schemes for the kNN classifier on time-series data *Knowl. Inf. Syst.*
[33] Weinberger K Q and Saul L K 2009 Distance metric learning for large margin nearest neighbor classification *Journal of Machine Learning Research* pp 207-244

Hybrid principal component analysis and K-nearest neighbour to detect the catfish disease.pdf

ORIGINALITY REPORT

16%	%	%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to UIN Sunan Gunung DJati Bandung Student Paper	7%
2	Submitted to Universitas Negeri Jakarta Student Paper	3%
3	Submitted to Universitas Pendidikan Indonesia Student Paper	1%
4	Submitted to Binus University International Student Paper	1%
5	Submitted to The University of Buckingham Student Paper	1%
6	Submitted to University of Surrey Student Paper	1%
7	Submitted to Heriot-Watt University Student Paper	1%
8	Submitted to Universitas Islam Indonesia Student Paper	1%

9	Submitted to Far Eastern University Student Paper	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		