Dimensionality Reduction for CCD Sensor-Based Image to Control Fall Armyworm in Agriculture

Alex B. Bertolla[1,2], Paulo E. Cruvinel[1,2]
[1]Embrapa Instrumentation, São Carlos, SP, Brazil
[2]Post-Graduation Program in Computer Science - Federal University of São Carlos, SP, Brazil
E-mail: alex.bertolla@embrapa.br, paulo.cruvinel@embrapa.br

***Abstract*—Digital imaging sensors, such as Charge-Coupled Devices, have been used for large-scale agricultural pest control. The ability to process and analyze the amount of data generated by these sensors has become a challenge, especially due to the high dimensionality of the collected features. In the literature, it is possible to find various research on dimensionality reduction and algorithms. This article presents a study on the dimensionality reduction of features from a digital image acquired with a Charge-Coupled Devices sensor in an agricultural field, in order to choose the optimal number of principal components for reducing feature dimensionality. In this context, it has become very important to define a method for selecting the optimal number of principal components for dimensionality reduction, while retaining only the necessary information associated with the main variables that describe the object of interest (Fall armyworms - *Spodoptera frugiperda*). The results showed, for example, that by using Hu invariant moments for feature extraction, dimensionality reduction was possible for all analyzed cases, leading to 80% of the original data. In this context, it was possible to preserve the semantic characteristics collected by the sensor and prepare them for classification.**

***Keywords*—CCD sensor, digital image, feature extraction, dimensionality reduction, principal component analysis.**

I. Introduction

Charge-Coupled-Devices (CCD) are the most used imaging sensors for digital image acquisition. They have built-in frame capture systems and the analog-to-digital conversion is done in the sensor itself [1].

CCD's sensors have been used in such ways to acquire images for different purposes. In agriculture, those sensors are usually used to capture images of pests and diseases [2], [3].

Due to the complex and high dimensions of the data captured by those sensors, storing and processing the amount of data acquired has become a challenging task [4], known as the curse of dimensionality [5]. This phenomenon is related to the fact, that with a certain degree of accuracy from a function estimation, the number of variables increases as the number of samples also has to increase [6].

To solve the issue of the curse of dimensionality, different methods based on dimensionality reduction techniques have been proposed [7]. These methods transform the original high-dimensional data into a new reduced dataset, removing the redundant and non-relevant features [8]. Dimensionality reduction algorithms allow an efficient reduction of the number of variables, and if applied before machine learning models can avoid overfiting.

In the literature, it is possible to find diverse research available about dimensionality reduction techniques for different types of data, such as Principal Component Analysis (PCA) introduced in 1901 by Karl Pearson [9], and its variations [10], Linear Discriminant Analysis (LDA) [11], Singular Value Decomposition (SVD) [12] and ISOMAP [13].

PCA is a linear dimension reduction technique and is the most predominant method applied [14], and was considered to compose this work.

This paper presents a method for the dimensionality reduction optimization when using a CCD sensor-based images to control Fall armyworms in agriculture. In fact, the task of image classification allows the machine to understand what type of information is contained in an image, on the other hand, semantic segmentation methods allow the precise location of different kinds of visual information, as well as each begins and ends.

After the introduction, this document is organized as follows: Section II describes the work methodology; Section III shows the results and the discussion of the experiments; and finally, Section IV presents the conclusion of this paper.

II. Materials and Methods

A. Digital Image Sensor and Dataset

A digital image can be defined as a bi-dimensional function $f(x, y)$, where (x, y) are the intensity positions, defined as pixel [15]. CCD's sensors can capture images in different color spaces, however, the most common color space is the Red Green, and Blue (RGB), which represents the visible spectrum [16].

Table I presents the features of the images acquired using the CCD sensor.

TABLE I
Image features acquired by CCD sensor

Image type	JPG / JPEG
Color space	RGB
Width	3072 pixels
Height	2048 pixels
Resolution	72 pixels per inch (ppi)
Pixel size	0,35mm

Regarding the image acquisition, a dataset was generated using a CCD sensor. This dataset is composed of the Fall

armyworm images in real maize crops, where the pest was found both in leaves and cobs maize.

B. Feature Extraction

The Hu invariant moments descriptor was considered for the extraction of the geometric features of the pest. For the calculation of the seven invariant moments of Hu, it is necessary, a priori, to calculate the two-dimensional moments, that is, the central moments and normalized central moments [17]. Two-dimensional moments are understood to be the polynomial functions projected onto a 2D image, $f(x, y)$, and size $M \times N$ and order $(p + q)$.

The normalized central moments allow the central moments to be invariant to scale transformations, being defined by:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{1}$$

where γ is defined as:

$$\gamma = \frac{p+q}{2} + 1 \tag{2}$$

for $p + q = 2, 3, ...$, positive integers $\in \mathbb{Z}$.

In this way, the invariant moments can be calculated considering:

$$\phi_1 = \eta_{20} + \eta_{02} \tag{3}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{4}$$

$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{5}$$

$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{6}$$

$$\begin{gathered}\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) \\ [(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03}) \\ (\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{gathered} \tag{7}$$

$$\begin{gathered}\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + \\ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{gathered} \tag{8}$$

$$\begin{gathered}\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) \\ [(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{gathered} \tag{9}$$

Neither of the seven Hu invariant moments is directly related to the size of an object in an image. However, the size of an object can be indirectly inferred through either the first or fourth moment [18].

After the features are extracted using the methods considered, a single feature vector is organized. Then, to reduce its dimensionality, PCA is applied [19].

C. Principal Components Analisys

PCA considers an array $\mathbf{X}$ of data with n samples representing the number of observations and m independent variables [20], that is:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \tag{10}$$

Herein, the principal components are obtained for a set of m variables X_1, X_2, ..., X_m with means μ_1, μ_2, ..., μ_m and variance σ_1^2, σ_2^2, ..., σ_m^2, which are independent and have covariance between the n-th and m-th variable [8], in the form:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1m}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \cdots & \sigma_{nm}^2 \end{bmatrix} \tag{11}$$

where $\mathbf{\Sigma}$ represents the covariance matrix. To do this, the pairs of eigenvalues and eigenvectors are found (λ_1, e_1), (λ_2, e_2), ..., (λ_m, e_m), where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ and associated with Σ [21], where the i-th principal component is defined by:

$$Z_i = e_{i1}X_1 + e_{i2}X_2 + ... + e_{im}X_m \tag{12}$$

where Z_i is the i-th principal component. The objective is to maximize the variance of Z_i, as:

$$Var(Z_i) = Var(e_i'\mathbf{X}) = e_i' Var(\mathbf{X}) e_i = e_i' \mathbf{\Sigma} e_i \tag{13}$$

where i = 1, ..., m. Thus, the spectral decomposition of the matrix $\mathbf{\Sigma}$ is given by $\mathbf{\Sigma} = \mathbf{P\Lambda P'}$, where $\mathbf{P}$ is the composite matrix by the eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{\Lambda}$ the diagonal matrix of eigenvalues of $\mathbf{\Sigma}$ [22]. Thus, it has to be:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix} \tag{14}$$

In general, the principal component of greatest importance is defined as the one with the greatest variance which explains the maximum variability in the data vector. The second most important component is the component that has the second highest variance, and so on, up to the least important component [12].

Likewise, the normalized eigenvectors represent the main components that constitute the feature vector with reduced dimension. Besides, such reduced components are used to describe the acquired images. Additionally, the reduced features are used for the recognition of the patterns of Fall armyworm (*Spodoptera frugiperda*), i.e., useful consideration for both cases, leaf or cob maizes.

III. Results and Discussion

For this study, an image dataset composed of 2280 images acquired with CCD's sensor was used. These images represent the Fall armyworm (*Spodoptera frugperda*) acquired in a real environment of maize crop in its five different stages of growth, grouped in 456 images for each stage. Figure 1(a to e) illustrates one example of each stage of growth, also named Instar, Figure 1(f) illustrates two different Instar in the same image.

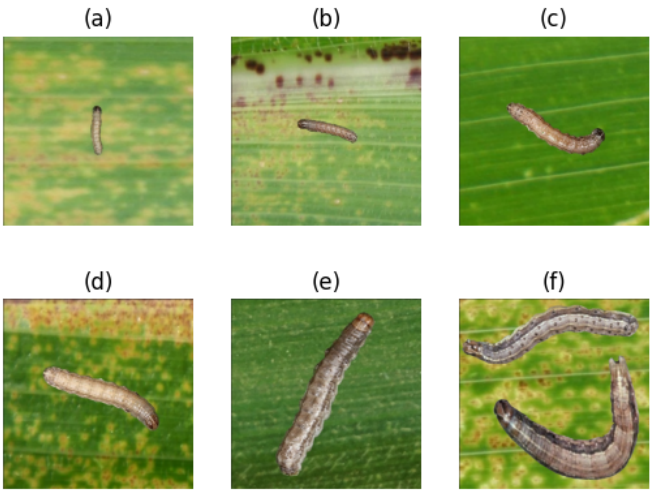

Fig. 1. Fall armyworm (*Spodoptera frugperda*) in different stages of growth.

Considering all images contain at least one specimen of the pest in different stages, in other words, different sizes, the Hu invariant moments descriptor has been considered for instance. Thus, for each image of the Fall armyworm, a feature vector was generated, containing the seven Hu invariant moments (ϕ_1, ϕ_2, ϕ_3, ϕ_4, ϕ_5, ϕ_6 and ϕ_7), which are related to the shape and geometrical features of this pest. The features contained in these vectors will allow the classification of the Fall armyworm (*Spodoptera frugperda*) in its different stages of growth.

Table II presents the seven Hu invariant moments, as examples, from three different images, which were processed using the dataset.

TABLE II
Feature vector composed of Hu invariant moments. Example of three images

Hu invariant moments	Images		
	Image 1	Image 2	Image 3
ϕ_1	6.692	6.6178	6.524
ϕ_2	13.581	13.424	19.102
ϕ_3	24.321	23.944	22.370
ϕ_4	25.919	26.245	23.445
ϕ_5	51.517	-52.023	46.665
ϕ_6	34.307	-33.305	-34.728
ϕ_7	-51.458	-51.556	47.656

As the values of the feature vectors were in different scales, it was necessary to normalize them. To generate a database of characteristics of the fall armyworm (*Spodoptera frugperda*), the feature vectors referring to each image were saved on disk.

Table III presents the normalized seven Hu invariant moments from three different images.

TABLE III
Normalized feature vector. Example of three images

Hu invariant moments	Images		
	Image 1	Image 2	Image 3
ϕ_1	0.274	0.162	0.021
ϕ_2	-1.048	-1.121	1.496
ϕ_3	1.035	0.817	-0.092
ϕ_4	1.408	1.669	-0.575
ϕ_5	0.719	-1.607	0.610
ϕ_6	0.808	-1.289	-1.333
ϕ_7	-0.863	-0.865	1.199

To remove duplicate information and also non-significant information, this stage first proceeds to the dimensionality reduction of the feature vector through PCA. To achieve the appropriate number of principal components that explain the original data, seven principal components were measured.

Through the variance ratio metric, it was possible to measure how much each of the seven principal components was explained from the original data. Figure 2 shows the variance ratio.

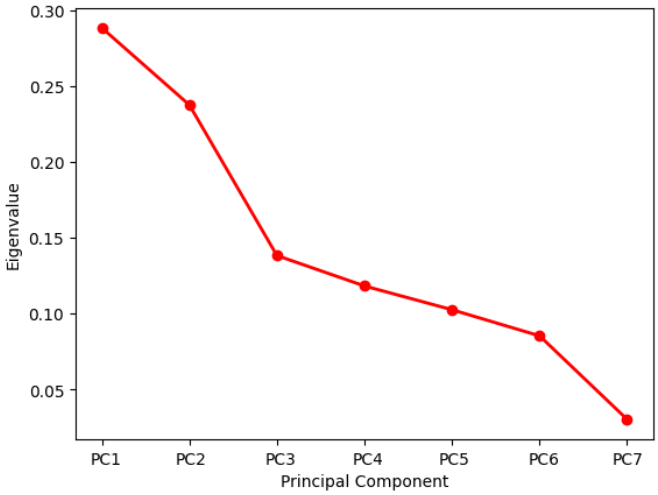

Fig. 2. Scree plot.

Analyzing the scree plot in Figure 2, it is possible to infer that by applying two to four principal components it is possible to explain almost from 55% to 80% of the variability of the original data. Considering that, the experiments were based on four principal components.

As discussed in the prior section, neither of the seven invariant moments is directly related to the size of an object. However, the first and the fourth moments can be used to infer the size of an object in an image.

Moreover, through the maximum variation ratio metric it is possible to measure the weight of each of Hu invariant moments in each principal component.

Table IV presents the maximum variance to each of the four principal components concerning the original data.

TABLE IV
MAXIMUM VARIATION OF DATA IN RELATION TO EACH PRINCIPAL COMPONENT. BASED ON FOUR PRINCIPAL COMPONENTS.

Hu invariant moments	Principal components			
	PC 1	PC 2	PC 3	PC 4
ϕ_1	-0.147	-0.630	-0.120	-0.568
ϕ_2	0.501	-0.295	0.036	0.230
ϕ_3	-0.395	-0.424	0.087	0.027
ϕ_4	-0.378	0.501	-0.158	-0.479
ϕ_5	-0.376	-0.277	-0.310	0.286
ϕ_6	-0.355	-0.011	0.865	0.127
ϕ_7	0.398	-0.078	0.325	-0.543

Even though some values presented in Table IV are negative, the weights for each principal component are considered absolute values. For example, in PC2, the first moment (ϕ_1) has the highest weight.

Figure 3 illustrates the maximum variance to each of the four principal components concerning the original data with the absolute values.

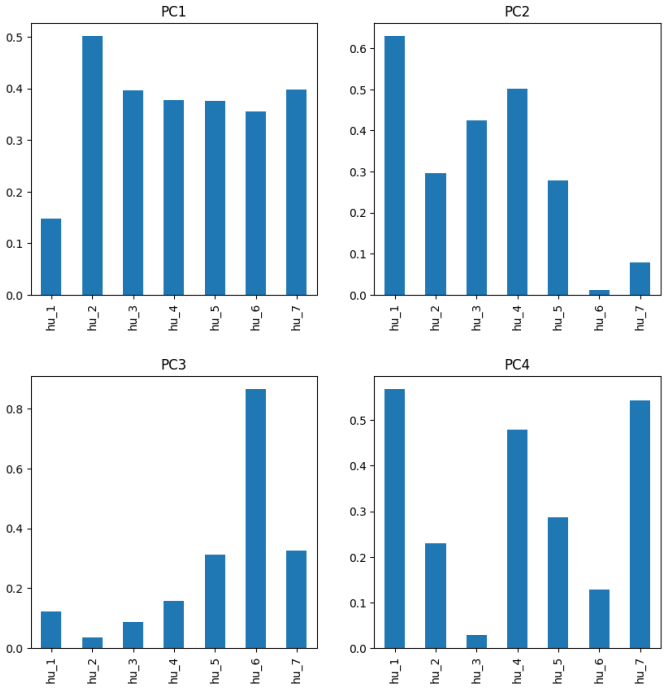

Fig. 3. Maximum variation of data in relation to each principal component, based on four principal components.

The experiment with four principal components showed that, as can be visualized in Figure 3, to have the most representative weights either from the first moment or the fourth moment, it was necessary to work with two or four principal components.

Based on prior experiments, the dimensionality reduction of the feature vector was performed. Table V presents the values of the four principal components.

TABLE V
FEATURE VECTOR COMPOSED OF FOUR PRINCIPAL COMPONENTS. EXAMPLE OF THREE IMAGES

Principal components	Images		
	Image 1	Image 2	Image 3
PC1	0.333	2.121	-0.742
PC2	-2.280	-1.156	1.306
PC3	0.551	-1.243	-0.528
PC4	0.181	-1.193	0.012

The distribution of the variation of the four principal component values is illustrated in Figure 4.

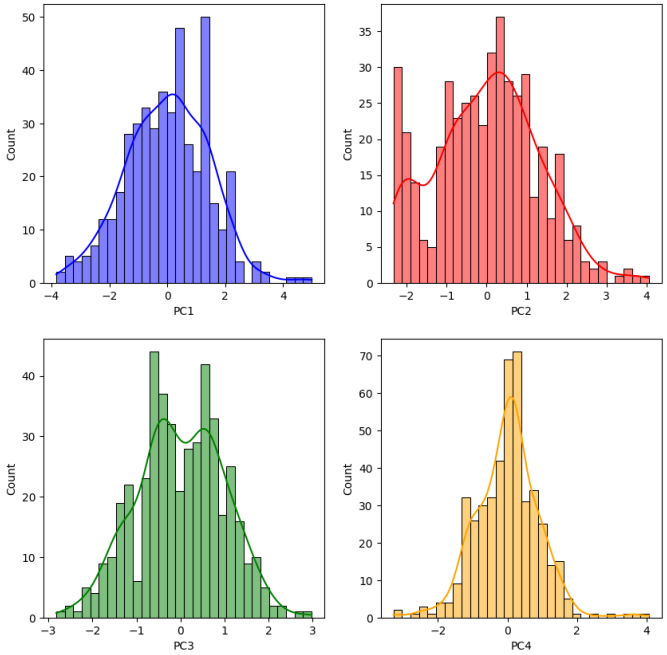

Fig. 4. Histogram of distribution of the four principal components values.

Once the values of the four principal components are obtained, it is necessary to evaluate if it is sufficient to work with two principal components, or whether it should be considered four principal components. For this purpose, it should be considered the maximum variation of each principal component to the original data, how much the principal components could explain the original data, and also the minimum error.

This information can be observed in Figure 5, which illustrates a boxplot chart of the four principal component values and their distribution.

From Figure 5, it is possible to visualize that when working with two or three principal components the error would be minimal, but with two principal components only 55% of the original data is explained, and working with three principal components the first and fourth moments are not representative.

This experiment demonstrated that working with four principal components was the ideal option. Because both the first and fourth moments are very representative, four principal

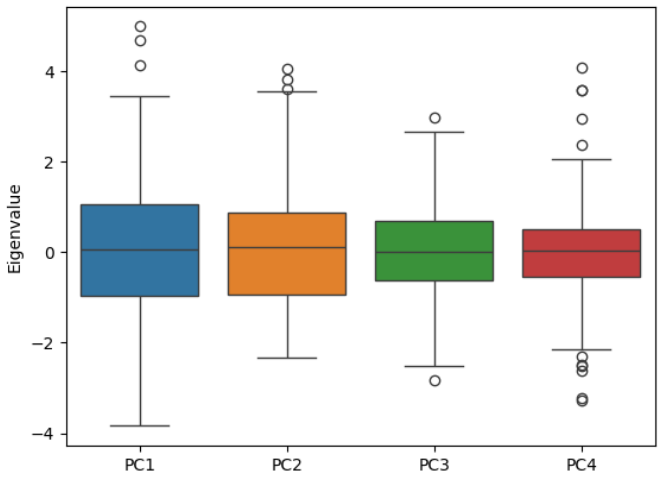

Fig. 5. Boxplot of four principal components values.

components can explain 80% of the original data, and even with a low increase in the error, it is not considerable to decrease the estimation.

IV. Conclusions

This paper presented a study of dimensionality reduction using Principal Components Analysis (PCA), considering feature vectors composed of extracted Hu invariant moments.

Before measuring the number of principal components necessary to represent the original data from the Fall armyworm digital images, the feature vectors were normalized, to obtain all the seven Hu invariant moments.

The measure of the explained variance ratio to the original data was applied to verify the quantity number of principal components necessary to explain the maximum of the original data.

In addition, the first and fourth invariant moments were used to infer the estimated size of the Fall armyworm (*Spodoptera frugperda*) in the images. Likewise, the measure of the maximum variation of each principal component, concerning to each Hu invariant moment, was performed to find how much these moments contribute to recognizing the main elements acquired with the CCD's sensor.

The measurements show that computing two to four principal components was sufficient to explain 55% to 80% of the original data, and either the first or fourth moments were contained in two and four principal components.

Finally, despite seven invariant moments being used, such analysis led to the conclusion that when using 4 principal components, one may achieve the explanation of 80% for the original data, with low error, as well as, not a significative variation.

For future works, it is suggested to extend this research to an unsupervised method to reach the selection of the number of principal components to remain with the semantic features from a recognized agricultural pest.

Acknowledgment

This research was partially supported by the São Paulo Research Foundation (FAPESP 17/19350-2). We thank the Brazilian Corporation for Agricultural Research (Embrapa) and the Post-Graduation Program in Computer Science from the Federal University of São Carlos (UFSCar). The Authors also recognize the helpful discussions with MSc Bruno M. Moreno to finalize the manuscript.

References

[1] C. Zhang, H. Hu, D. Fang, and J. Duan, "The ccd sensor video acquisition system based on fpga&mcu," in *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, vol. 9. IEEE, 2020, pp. 995–999.

[2] S. Sankaran, R. Ehsani, and E. Etxeberria, "Mid-infrared spectroscopy for detection of huanglongbing (greening) in citrus leaves," *Talanta*, vol. 83, no. 2, pp. 574–581, 2010.

[3] J. L. Miranda, B. D. Gerardo, and B. T. T. III, "Pest detection and extraction using image processing techniques," *International journal of computer and communication engineering*, vol. 3, no. 3, pp. 189–192, 2014.

[4] C. Ji *et al.*, "Big data processing: Big challenges and opportunities," *Journal of Interconnection Networks*, vol. 13, no. 03n04, p. 1250009, 2012.

[5] W. K. Vong, A. T. Hendrickson, D. J. Navarro, and A. Perfors, "Do additional features help or hurt category learning? the curse of dimensionality in human learners," *Cognitive science*, vol. 43, no. 3, p. e12724, 2019.

[6] A. L. M. Levada, "Parametric PCA for unsupervised metric learning," *Pattern Recognition Letters*, vol. 135, pp. 425–430, 2020.

[7] F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review*, vol. 40, p. 100378, 2021.

[8] S. Ayesha, M. K. Hanif, and R. Talib, "Overview and comparative study of dimensionality reduction techniques for high dimensional data," *Information Fusion*, vol. 59, pp. 44–58, 2020.

[9] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[10] A. L. M. Levada, "PCA-KL: a parametric dimensionality reduction approach for unsupervised metric learning," *Advances in Data Analysis and Classification*, vol. 15, no. 4, pp. 829–868, 2021.

[11] F. Pan, G. Song, X. Gan, and Q. Gu, "Consistent feature selection and its application to face recognition," *Journal of Intelligent Information Systems*, vol. 43, pp. 307–321, 2014.

[12] S. Nanga *et al.*, "Review of dimension reduction methods," *Journal of Data Analysis and Information Processing*, vol. 9, no. 3, pp. 189–231, 2021.

[13] Z. A. Sani, A. Shalbaf, H. Behnam, and R. Shalbaf, "Automatic computation of left ventricular volume changes over a cardiac cycle from echocardiography images by nonlinear dimensionality reduction," *Journal of Digital Imaging*, vol. 28, pp. 91–98, 2015.

[14] R. M. Wu *et al.*, "A comparative analysis of the principal component analysis and entropy weight methods to establish the indexing measurement," *PLoS One*, vol. 17, no. 1, p. e0262261, 2022.

[15] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Pearson Prentice Hall, 2004.

[16] G. Saravanan, G. Yamuna, and S. Nandhini, "Real time implementation of rgb to hsv/hsi/hsl and its reverse color space models," in *2016 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2016, pp. 0462–0466.

[17] W. Zhao and J. Wang, "Study of feature extraction based visual invariance and species identification of weed seeds," in *2010 Sixth International Conference on Natural Computation*, vol. 2. IEEE, 2010, pp. 631–635.

[18] L. Zhang, F. Xiang, J. Pu, and Z. Zhang, "Application of improved hu moments in object recognition," in *2012 IEEE International Conference on Automation and Logistics*. IEEE, 2012, pp. 554–558.

[19] K. Hongyu, V. L. M. Sandanielo, and G. J. de Oliveira Junior, "Principal analisys components: teorical summart, aplication and interpretation, (análise de componentes principais: resumo teórico, aplicação e interpretação)," *E&S Engineering and science*, vol. 5, no. 1, pp. 83–90, 2016.

[20] B. Zhao, X. Dong, Y. Guo, X. Jia, and Y. Huang, "PCA dimensionality reduction method for image classification," *Neural Processing Letters*, pp. 1–22, 2022.

[21] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "PCA-based feature reduction for hyperspectral remote sensing image classification," *IETE Technical Review*, vol. 38, no. 4, pp. 377–396, 2021.

[22] B. M. S. Hasan and A. M. Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.