Dimensional Reduction of Underwater Shrimp Digital Image Using the Principal Component Analysis Algorithm

Arif Setiawan1,5*, Hadiyanto Hadiyanto2,4, and Catur Edi Widodo3,4

1Doctoral Program of Information Systems, School of Postgraduate Studies, Diponegoro University, Jl. Imam Bardjo SH, Semarang, 50241, Indonesia
2Center of Biomass and Renewable Energy (CBIORE), Department of Chemical Engineering, Diponegoro University. Jl. Prof. Soedarto SH, Tembalang, Semarang, 50271, Indonesia
3Department of Physics, Faculty of Science and Mathematics, Diponegoro University, Jl. Prof. Soedarto, Tembalang, Semarang, 50275, Indonesia
4School of Postgraduate Studies, Diponegoro University, Jl. Imam Bardjo SH, Semarang, 50241, Indonesia
5Department of Information System, Faculty of Engineering, Muria Kudus University, Jl. Lingkar Utara, Kayuapu Kulon, Gondangmanis, Kudus, 59327, Indonesia

Abstract. Shrimps are aquaculture products highly needed by the people and this is the reason their growth needs to be monitored using underwater digital images. However, the large dimensions of the shrimp digital images usually make the processing difficult. Therefore, this research focuses on reducing the dimensions of underwater shrimp digital images without reducing their information through the application of the Principal Component Analysis (PCA) algorithm. This was achieved using 4 digital shrimp images extracted from video data with the number of columns 398 for each image. The results showed that 12 PCs were produced and this means the reduced digital images with new dimensions have 12 variable columns with data diversity distributed based on a total variance of 95.61%. Moreover, the original and reduced digital images were compared and the lowest value of MSE produced was 94.12, the minimum value of RMSE was 9.54, and the highest value of PSNR was 8.06 db, and they were obtained in the 4th digital image. The experiment was conducted using 3 devices which include I3, I7, and Google Colab processor computers and the fastest computational result was produced at 2.1 seconds by the Google Colab processor. This means the PCA algorithm is good for the reduction of digital image dimensions as indicated by the production of 12 PC as the new variable dimensions for the reduced underwater image of shrimps.

1 Introduction

This Shrimp is one of the top aquaculture foods needed in the world [1]. This seafood contributes significantly to the economy of coastal areas and is reported to be widely

* Corresponding author: arif.setiawan@umk.ac.id

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
developed in developing countries due to its great influence on marine product exports [2]. It is important to note that vannamei are the popular species of shrimps popularly cultivated [3] due to the fact that they require a relatively short time for cultivation as indicated by the 3 months for one harvest [4]. There is, therefore, a need to monitor shrimps in different aquaculture locations using smart systems in order to increase their yield [5]. This led to the wide application of computer vision technology for non-invasive monitoring of shrimps [6] using several types of data such as digital images, video, and sound [7]. Meanwhile, one of the devices usually used to obtain the image data of shrimps from under the water without touching them is a camera [8].

Digital images have high dimensions [9] and are considered a vector with a 2-dimensional pixel value where each pixel consists of an RGB bit value with a high-dimensional data space [10]. This simply implies a digital image is a two-dimensional function with spatial coordinates x and y. It consists of a finite number of elements with each having a specific area [9]. Moreover, digital images represent the quantity characteristics of each image element called pixels as well as the lighting of objects obtained through a camera. They also provide vague quantitative descriptive information [11]. It is very necessary to reduce the dimensions of digital images in order to ensure easier processing without reducing the information contained [12]. Therefore, this research proposes the application of the Principal Component Analysis (PCA) method to reduce digital images. PCA is a method normally applied to reduce a variable in high-dimension digital image data [13] without reducing the information from the initial data [14]. It is also a statistical-based approach usually used to find a variable in a set of interrelated variables [15].

The PCA algorithm has been previously applied by Itkonen [16] to analyze the toll road driving style characterization using naturalistic driving data with a focus on the covariance between the driving style features. Moreover, it was also applied by Particles to compress and reconstruct digital images. This author showed that images with very large dimensions have high feature representations and also require large space in storage media. This led to the application of PCA to compress and reconstruct the image without reducing the available information [17]. Khaing [10] also used PCA to reduce the dimensions of digital images. The research explained that high-resolution images experience several problems when processing their data and this is the reason the PCA algorithm was used at the preprocessing stage to compress the image. The results showed that compressed digital images are transmitted at a faster time. Nasution [18] applied PCA for feature extraction of human facial images in order to determine the similarity of facial patterns. Mustaqeem [19] also combined PCA with a support vector machine (SVM) to predict weaknesses in a tool program with the initial step being the reduction of the variables. Islam [20] used PCA to detect brain tumors using digital images. The process involved combining the method with the K-Means algorithm to cluster brain tumor images such that the super-pixel image features were initially extracted using the algorithm. The results showed that the process led to better accuracy and required a shorter computation time. Moreover, Peretti [21] also applied PCA to measure the feasibility of images using pharmacokinetic parameters. The scaled sub-profile modeling conducted through this method was used to explore the use of image parametric in analyzing patient disease.

These studies proved that the PCA algorithm is effective in reducing the dimensions of digital images as an initial step in data processing [22]. Therefore, this research proposes the application of this PCA algorithm to reduce the dimensions of digital images of shrimp under water. This is considered the first step to measuring shrimp weight in an aquaculture environment through a non-invasive model. The goal of this study is to (a) reduce the dimensions of the digital image of the shrimp under the water without reducing the information from the image and (b) compare the original image with the reduced image produced from the application of the PCA algorithm.
2 Materials and Methods

2.1 Stages of the Study Process

This study was conducted in several stages which include (a) reading the vector value of the digital shrimp image, (b) testing the image vector variable using KMO and Bartlett's Test, (c) standardizing the shrimp digital image data, (d) calculating covariance/correlation matrices, (e) calculating eigenvalues, (f) calculating component matrices and component matrix rotations, (g) calculating principal components as new dimension variables for the reduced digital images, (h) splitting the RGB image of shrimp into 3 channels which include red, green, and blue channels, (i) implementing the PC obtained into 3 image channels, and (j) comparing the original and reduced images. These processes are described in the following Figure 1.

Fig. 1. Stages of the study process

2.2 Shrimp Image Data Collection

The data used is a digital image of live shrimp underwater obtained using an underwater camera. The data was retrieved in an mp4 format and used to produce a sample of 4 digital shrimp images in a jpeg format [23]. The dimension of each digital image used was read in the form of a matrix as indicated in Equation 1. Its explains that the $X$ digital image matrix has dimensions of $m \times n$, $m$ indicates the number of columns and $n$ indicates the number of rows in the pixel [24]. In this study it will be tested to reduce the number of columns contained in each image using the PCA algorithm.

$$ X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} $$

2.3 Testing Variables Using Barlett's Test and Kaiser-Meyer-Olkin (KMO)

The Barlett's Test is normally used to determine the correlated variables such that when the sig value is above 0.05, the variables do not meet the requirements for the factor analysis process and vice versa [25].

The KMO values range between 0 and 1 and they are normally used to assess the feasibility of factor analysis to be conducted such that when the value is greater than 0.5, then the factor analysis in the PCA algorithm is feasible. However, when the value is less than 0.5, there is a need to retest by reducing the number of variables used up to the moment the value is greater than 0.5. The KMO calculation process is described in the following Equation 2 [26, 27].
\[ KMO = \frac{\sum_{i} \left[ \Sigma_{i \neq k} r_{ik}^2 \right]}{\sum_{i} \Sigma_{i \neq k} r_{ik}^2 + \sum_{i} \Sigma_{i \neq k} a_{ik}^2} \]  

(2)

Where, \( r_{ik}^2 \) is the square of the simple correlation matrix and \( a_{ik}^2 \) is the square of the partial correlation matrix [28].

### 2.4 Data Standardization

Data normalization was conducted at this stage and the normalized value was found to be the value of each variable in an image vector column. Meanwhile, the average of each variable is the standard deviation [29]. The data standardization process was calculated using Equation 3 while the standard deviation value was determined through Equation 4.

\[ z = \frac{x - \mu}{\sigma} \]  

(3)

\[ S = \sqrt{\frac{\sum(x_i - \mu)^2}{n}} \]  

(4)

Where, \( S \) is the standard deviation value, \( x_i \) is the value \( x \) to \( i \) of one variable, \( \mu \) is the average \( x \) value, and \( n \) is the amount of data.

### 2.5 Calculating the Covariance/Correlation Matrix

The dimension of the covariance/correlation matrix is \( n \times n \) based on the number of image variables and the covariance/correlation matrix was calculated using Equation 5.

\[ Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y) \]  

(5)

Where, \( \mu_x \) is the mean of sample \( x \), \( \mu_y \) is the mean of sample \( y \), \( x_i \) is the value observation \( i \) of variable \( x \), and \( y_i \) is the observation \( i \) of variable \( y \) [30, 23].

### 2.6 Calculating Eigenvalues and Eigenvector

The eigenvalues and eigenvectors were determined using the covariance matrix as indicated in Equation 6 for eigenvalue and Equation 7 for eigenvector.

\[ \text{Determinan} (A - \lambda I) = 0 \]  

\[ AX = \lambda X \]  

\[ AX - \lambda X = 0 \]  

\[ (A - \lambda) x = 0 \]  

\[ (A - \lambda I)x = 0, x \neq 0 \]  

(6)

(7)

Where, \( A \) is the \( n \times n \) matrix, \( \lambda \) and \( \lambda \) are eigenvalues, \( x \) is a non-zero matrix, and \( I \) is the identity matrix. The eigenvalue of the \( X \) matrix is symbolized by \( \lambda_1, \lambda_2, \lambda_3 ... \lambda_n \), and the eigenvector is represented by \( x_1, x_2, x_3 ... x_n \) [31].
2.7 Rotation of Component Matrix

The matrix was rotated using varimax which normally maximizes the weighting factor as indicated in Equation 8. This step is necessary to determine the new reduces variables. Where \( X'_{mxn} \) is the rotation matrix, \( X_{mxn} \) is the original matrix, dan \( T_{nxn} \) is the multiplication matrix.

\[
X'_{mxn} = X_{mxn} * T_{nxn}
\]  
(8)

2.8 Calculating the Value of the Principal Component (PC)

The principal component is a newly selected variable with an eigenvalue above 1. It is a reduced image dimension that contains information on variables that are not correlated with each other. It is important to note that the PC process reduced the size of the original matrix while maintaining data variation which involves data that is not redundant or uncorrelated [32]. The eigenvalue was also calculated using the following Equation 9 [23].

\[
PC(\%) = \frac{\text{eigenvalue}}{\text{variance covarian}} \times 100\%
\]  
(9)

2.9 Evaluation Method

The results of this experiment were evaluated using Mean Square Error (MSE), Root Mean Square Error (RMSE), and Peak Signal-to-Noise Ratio (PSNR) through Equations 10, 11, and 12 respectively.

\[
MSE = \frac{1}{n} |x - x'|^2 = \frac{1}{n} \sum_{i=1}^{n} (x - x')^2
\]  
(10)

\[
RMSE = \sqrt{\frac{1}{n} |x - x'|^2} = \frac{1}{n} \sum_{i=1}^{n} (x - x')^2
\]  
(11)

\[
PSNR = 10. \log_{10} f(x) = \left( \frac{\text{MAX}^2}{\text{MSE}} \right)
\]  
(12)

3 Results and Discussion

This study was conducted using SPSS software for the PCA algorithm analysis and Python programming software to test the shrimp digital images. A total of four digital image data of shrimp underwater were used for this purpose as indicated in Figure 2. The matrix value of each image was read and this led to the production of 224 rows x 398 columns dimensions. In the process of reading the image using the OpenCV and Pandas libraries. The pixel data is read using the Pandas library in Python using the pd.dataframe() function to find out the value of the numerical component matrix of each image. After obtaining the numerical value of the image dimensions, the numerical image data is normalized using the Standarscaler() function and then processed using the PCA algorithm.

The digital image data of each shrimp was further tested using Barlett and KMO tests. The results showed that the Barlett significant value is 0.000 (sig < 0.05) and this means there is a significant correlation between the variables. Moreover, the KMO test produced 0.967 which is greater than 0.5 (> 0.5) and this indicates the variables are adequate for this study. It was discovered that these tests meet the requirements of the PCA algorithm. The comprehensive results are presented in the following Table 1.
This was followed by the standardization of data as well as the calculation of the covariance matrix and eigenvalues. The covariance matrix has a nxn dimension where n is the number of digital image variables. Meanwhile, the experiment conducted using 398 columns showed 12 principal components (PC) and their initial eigenvalues including the percentage of variance and cumulative are presented in Table 2. The results showed that 12 PCs were formed and these include PC1 with 125.854 eigenvalues and 62.297 variance, PC2 with 19.727 and 9.864, PC3 with 14.377 and 7.188, PC4 with 6.875 and 3.437, PC5 with 6.36 and 3.18, PC6 with 4.494 and 2.747, PC7 with 3.824 and 1.912, PC8 with 2.509 and 1.254, PC9 with 2.082 and 1.041, PC10 with 1.621 and 0.811, PC11 with 1.383 and 0.692, and PC12 with 1.123 and 0.561 respectively. The distribution of data diversity for all the PCs from the total variance was found to be 95.054%. Meanwhile, the distribution of eigenvalues and the percentage of variances are comprehensively described in Figure 3 (a) and (b).

### Table 1. Barlett's Test and KMO test results

<table>
<thead>
<tr>
<th>Kaiser-Meyer-Olkin Measure of Sampling Adequacy</th>
<th>.967</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartlett's Test of Sphericity</td>
<td></td>
</tr>
<tr>
<td>Approx. Chi-Square</td>
<td>92547.232</td>
</tr>
<tr>
<td>df</td>
<td>4950</td>
</tr>
<tr>
<td>Sig.</td>
<td>.000</td>
</tr>
</tbody>
</table>

### Table 2. Eigenvalues, percentage of variance and cumulative

<table>
<thead>
<tr>
<th>Component</th>
<th>% of Variance</th>
<th>Total</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC 1</td>
<td>62.927</td>
<td>125.854</td>
<td>62.927</td>
</tr>
<tr>
<td>PC 2</td>
<td>9.864</td>
<td>19.727</td>
<td>72.791</td>
</tr>
<tr>
<td>PC 3</td>
<td>7.188</td>
<td>14.377</td>
<td>79.979</td>
</tr>
<tr>
<td>PC 4</td>
<td>3.437</td>
<td>6.875</td>
<td>83.417</td>
</tr>
<tr>
<td>PC 5</td>
<td>3.18</td>
<td>6.36</td>
<td>86.597</td>
</tr>
<tr>
<td>PC 6</td>
<td>2.747</td>
<td>5.494</td>
<td>89.344</td>
</tr>
<tr>
<td>PC 7</td>
<td>1.912</td>
<td>3.824</td>
<td>91.256</td>
</tr>
<tr>
<td>PC 8</td>
<td>1.254</td>
<td>2.509</td>
<td>92.51</td>
</tr>
<tr>
<td>PC 9</td>
<td>1.041</td>
<td>2.082</td>
<td>93.551</td>
</tr>
<tr>
<td>PC 10</td>
<td>0.811</td>
<td>1.621</td>
<td>94.362</td>
</tr>
<tr>
<td>PC 11</td>
<td>0.692</td>
<td>1.383</td>
<td>95.054</td>
</tr>
<tr>
<td>PC 12</td>
<td>0.561</td>
<td>1.123</td>
<td>95.615</td>
</tr>
</tbody>
</table>
Fig. 3. (a) Distribution of eigenvalues (b) Percentage of variance value

The component matrix and its rotation were calculated. It is important to note that the component matrix value is a member of the formed factor when it is $\geq 0.5$ but not a member when it is $< 0.5$. Moreover, when there was more than one component value that are $\geq 0.5$ in one dimension, they were rotated until there were no two factors with values $\geq 0.5$. The experiment was continued by splitting the RGB image pixels into 3 image channels which include the red, green, and blue image channels. The eigenvalues of each image channel were calculated and the results with the distribution of their variances. Figure 4 shows the distribution of eigenvalues and their variances for each of the red, green and blue image channels.

Fig. 4. The distribution of eigenvalues and their variances for 3 channel images
The 12 PC obtained were analyzed based on these channels and the results are indicated in Figure 5 (a). The image dimensions were continually reduced using these PCs after which each of the images was tested to determine the changes in the size. The comparison of the original and reduced images is presented in Figure 5 (b). The images were evaluated using MSE, RMSE, and PSNR. The results showed that the minimum MSE value was 94.12, the lowest RMSE value was 9.54, and the highest PSNR value was 8.06 and they were all obtained in the 4th digital image. This means this image has the reduction which is closest to the original image. The findings from this process are presented in Table 3 while the MSE, RMSE, and PSNR values are described in Figure 6 (a) and Figure 6 (b). Meanwhile, the PCs were applied to the digital images produce a new digital image with decreasing file size, image 1 file size decreased by 17 kb, from original size = 19 kb to 2 kb, image 2 file size decreased by 17 kb, from original size = 24 kb to 4 kb, image 3 file size decreased by 17 kb, from original size = 20 kb to 3 kb, image 4 file size decreased by 17 kb, from original size = 19 kb to 2 kb. The comparison of the file size for the original and reduced images is presented in Table 3 and Figure 6 (c). The computing process was conducted using a computer infrastructure with processor specifications of I3, I7, and through the application of python on Google Colab. The results obtained are presented in Figure 6 (d) and Google Colab was observed to have least processing time.

Table 3. Values of MSE, RMSE, PSNR and file size of image

<table>
<thead>
<tr>
<th></th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>94.35</td>
<td>96.92</td>
<td>94.12</td>
<td>95.61</td>
</tr>
<tr>
<td>RMSE</td>
<td>9.71</td>
<td>9.84</td>
<td>9.54</td>
<td>9.77</td>
</tr>
<tr>
<td>PSNR</td>
<td>7.14</td>
<td>5.86</td>
<td>8.06</td>
<td>7.1</td>
</tr>
<tr>
<td>Original size (kb)</td>
<td>19</td>
<td>21</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Reduction size (kb)</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Previous research conducted by Itkonen [16] explained that the PCA algorithm obtained a percentage principal component value of 45%, in this study the percentage principal component value was 95%. Another study conducted by Khaing [10] explained that the PCA algorithm reduces file size by 66.64%, in this study the value of decreasing image file size is 89.47%. The results of this study indicate that the PCA algorithm can be used to reduce the dimensional data of a digital image of a shrimp underwater with a significant decrease in dimensions, without changing the digital image information. Reducing the digital image file size can help the image processing process to be faster, for further study purposes. This study is the first step of the underwater shrimp detection process using digital images with complex noise.

4 Conclusion

The PCA algorithm was used to analyze a digital image in this study and 12 variables known as principal components were found to have eigenvalues above 1. These PCs were further used to reduce the dimensions of the digital image of shrimp underwater using 4 samples. It is important to note that the principal component formed has a data diversity distribution from the total variance of 95.054% while the 12 personal components had 95.61%. The PCs were applied to the digital images to produce a new digital image from an initial average file size of 20 kb to 3 kb. Moreover, the images were evaluated and the minimum MSE value was found to be 94.12, the lowest RMSE value was 9.54, and the highest PSNR value was 8.06 and they were all obtained in the 4th digital image. This shows that the PCA algorithm is good for the reduction of digital image dimensions without reducing the information. It also indicates the 4th image is the closest to the original image. The findings also showed that Google Colab had the best average processing time of 2 seconds compared to the other I3 and I7 computer processors. This study only uses 4 images data and has not used digital
images with a large amount of data, in the future study a research development will be carried out using a large amount of digital image data of shrimp underwater, and the introduction of living shrimp objects underwater will be carried out. Digital image analysis and machine learning approach.

The authors are grateful to the Information System Doctoral Program at Diponegoro University and Muria Kudus University for facilitating this study.

References

2. A. Priadana, A. W. Murdiyanto, and A. W. Murdiyanto, Shrimps clusterization by size using digital image processing with CCA and DBSCAN, 8, April, 106–112 (2020)


17. T. Particles, B. Ohtani, S. Chandren, and O. J. Gurney-champion, Image Compression and Reconstruction Based on Image Compression and Reconstruction Based on PCA (2021)


25. N. Iofrida et al., International Journal of Gastronomy and Food Science Italians’ behavior when dining out: Main drivers for restaurant selection and customers segmentation, 28, March (2022)


27. Kulshrestha, Performing the KMO and Bartlett’s Test for Factors Estimating the Warehouse Efficiency, Inventory and Customer Contentment for E-retail Supply Chain, 09 (2019)

28. N. Shrestha, Factor Analysis as a Tool for Survey Analysis, 9, 1, 4–11 (2021)

29. D. A. L. R. Ubinfeld, Data Standardization, April (2020)

