LONTAR KOMPUTER VOL. 9, NO. 1, APRIL 2018 **p-ISSN 2088-1541**
DOI : 10.24843/LKJITI.2018.v09.i01.p03 **e-ISSN 2541-5832**

Poisonous Shrimp Detection System for Litopenaeus Vannamei using k-Nearest Neighbor Method

Abdullah Husin[1], Othman Mahmod[2], Lisa Afrinanda [3]

[1,3]Department of Information System, Universitas Islam Indragiri, Indonesia
[2]Department of Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Malaysia
[1]abdialam@yahoo.com
[2]mahmod.othman@utp.edu.my
[3]Lisaafrinanda@gmail.com

Abstract

One of the important seafoods in the food consumption of humans is shrimp. Although shrimp contains proteins that are needed by the human body, sometimes it contains toxins. This is due to environmental factors or catching processes that may use toxins. Therefore, the community should take precautions when consuming shrimp. White shrimp (Litopenaeus vannamei) is one type of shrimp that is preferred because of its delicious taste. The purpose of this research is to develop a computerized system for poisonous white shrimp detection. The category of white shrimps consists of two kinds, i.e., fresh white shrimps that are caught in a natural way (class A), and poisonous white shrimps that are caught by using toxin (class B). The features used are RGB colors (red, green, and blue) and texture (energy, contrast, correlation, and homogeneity). A similarity-based classification is performed by the k-Nearest Neighbor (k-NN) algorithm. The experiment was conducted on a dataset consisting of 90 white shrimp images. The holdout validation method was used to evaluate the system. The original dataset was divided into two parts, whereby 60 images were used as training samples and 30 images were used as testing images. Based on the evaluation results, it can be concluded that the classification accuracy is 73.33%. The benefit of the developed system is to help the community in selecting good and safe white shrimps.

Keywords: *White shrimp, Classification, k-Nearest Neighbor, Holdout*

1. Introduction

Indonesia is one of the largest shrimp producing countries in the world. About 77% of the global shrimp production is produced by Asian countries, including Indonesia. Based on the 2013 data from the Ministry of Marine Affairs and Fisheries Indonesia, it is known that the achievement of fishery exports in Indonesia is approximately 802,000 tons at a price of US$2.6 billion. The achievement is largely sourced from shrimp commodities, which is US$997 million [1].

White shrimp, or *Litopenaeus vannamei* (see Figure 1), is one of the best-selling shrimps and is in great demand due to its taste, and it is often offered as the main menu at restaurants. White shrimps are fast growing in Indonesia and have several advantages over other types of shrimp as they have a fast growth cycle. The shrimps are usually caught in several ways: (1) by natural means such as nets and non-toxic baits; or (2) by toxins, for example, decis, tuba, and other toxins.

Figure 1. White Shrimp (*Litopenaeus Vannamei)*

Generally, the detection of toxic shrimp is performed by consumers in plain view. Poisonous shrimp detection tools are rarely used by ordinary people; most tools have never been applied in the marketplace. The detection of poisonous shrimp in plain view is less precise and inconsistent due to the limitations of the senses and negligence of humans. This often adversely affects consumers such as dry throat complaints, stomach aches, and itching [2].

The rapid development of computer hardware and software supported by pattern recognition and image processing has resulted in technological advances in the detection of objects through images. Therefore, it is expected that determining the classification of white shrimps can be realized with the help of computers and technology. This system is expected to be useful for the community, by helping to detect the type of both poisonous shrimps and natural shrimps.

2. Research Methodology

This study aims to develop a poisonous shrimp detection system for white shrimp variants. To achieve this objective, the following steps need to be taken as follows:

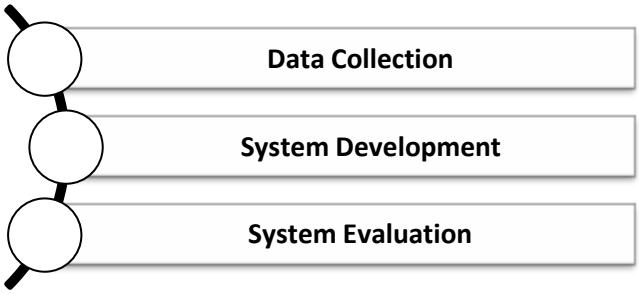

Figure 2. Research Methodology

2.1. Data Collection

A total of 90 (ninety) random shrimp samples were taken by image acquisition using Xiaomi Note 1 digital camera (13 MP camera resolution). A tripod was used to ensure that the image capture used the same distance of 40 cm. There are 2 (two) categories of samples taken, namely poisonous white shrimps and natural white shrimps. Each category consisted of 45 samples. All of the images were converted to bitmap file type and changed to 640 x 480 pixels resolution. Furthermore, preprocessing was applied to the images by using processing techniques.

2.2. System Development

A classification system was constructed consisting of two sub-systems, namely a class builder subsystem used to build a knowledge database, and a subclassification system used to predict unknown shrimp categories. The attributes or features used are color (red, green, and blue) and texture (energy, contrast, correlation, and homogeneity). These features are significant to be used in performing image classification [3]. The process of creating the database began with the process of feature extraction. After the feature extraction of sample images were performed, the feature vector of each sample image was added and stored in a knowledge database. The classification process was executed by using the k-Nearest Neighbor [4] method. The feature

vector of an unknown image was compared to the feature vector of a sample image stored in the knowledge database. The similarities were then calculated by using the distance between two feature vectors. The smaller obtained vector distance indicates that the unknown image is more similar to the certain sample image.

2.3. System Evaluation

System evaluation was performed to estimate the classification performance by using the holdout method [5]. In the holdout method, the original dataset with the known class labels was partitioned into two parts, namely training data used to build the knowledge database, and test data used to test the performance of the system. The ratio of training data and test data is 2 to 1. The original dataset consisting of 90 labeled images was divided into two parts. The first partition consisted of 60 images stored in the knowledge database, while the second partition comprised 30 images used as the test data.

After implementation of the system, the estimation of classification performance was obtained by using the holdout method. The recapitulation of the classification results was contained in the confusion matrix [6], whether the images were categorized correctly or categorized incorrectly. Then, the confusion matrix was used to measure the performance of the classification system. An example of a confusion matrix with two categories or classes is shown in Table 1.

Table 1. Confusion Matrix for Classification of Two Categories

f_{ij}		Prediction category (j)	
		category=1	category=2
Real category (i)	category=1	f_{11}	f_{12}
	category=2	f_{21}	f_{22}

Each f_{ij} cell contains the number of objects i, which is categorized as j. The total number of objects correctly categorized is $f_{11} + f_{22}$ and the total number of objects incorrectly categorized is $f_{12} + f_{21}$.

Generally, the accuracy of the system is the comparison between the number of objects correctly categorized and the total number of predictions, thus it can be written in the formula as follows:

$$accuracy = \frac{f_{11}+f_{22}}{f_{11}+f_{12}+f_{21}+f_{22}} \quad (1)$$

3. Literature Review

The literature review contains theories and articles related to the concept of classification and a brief review of researches related to the k-NN algorithm and its application.

3.1. Classification

Classification is the process of grouping objects or patterns into certain class labels that have been previously defined, based on their characteristics or attributes [7]. The task of classification is to predict the categorical or discrete target variable. Pattern classification is an important area in learning machine and artificial intelligence. This area has become an integral part of most intelligent engine systems or automated machines built for decision making.

The input of the classification system is the pattern of unknown objects and the output is the category of unknown objects as shown in Figure 3. Pattern classification has been used for predictions and decision making [8].

Figure 3. The Block Diagram of Classification System

A classifier is a function that maps a pattern or object that can be represented as a feature vector to one of the class labels. In other words, a classifier is an algorithm used to perform classification tasks. There are several approaches in classification: (1) based on similarity; (2) based on probabilistic approach; (3) constructing decision boundaries; and (4) combining classifiers.

3.2. K-Nearest Neighbor

K-Nearest Neighbor (k-NN) is one of the classification algorithms based on the similarity approach. K-NN is a commonly used method in classification problems. This method is effective and has been widely used in classification problems. The advantages of this method are reasonably simple, popular, effective, and efficient. This method is often applied and gives good results [9]. Similar objects will be classified in the same category. The similarity is obtained based on the closest distance between the sample data and the object. Objects are classified based on the majority of nearest neighbors, where the parameter k shows the number of nearest neighbors.

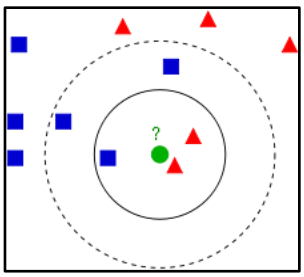

Figure 4. *k-NN* Illustration

Figure 4 is an illustration of the k-NN method. The question formulated in the figure is to determine the category of the green circle, whether it is a blue square or red triangle. If k = 3, then the green circle is categorized as a red triangle, because there are 2 red triangles and only 1 blue square inside the inner circle. If k = 5, then the green circle is categorized as a blue square, because there are 3 blue squares versus 2 red triangles in the outer circle.

The k-NN algorithm consists of two main steps: (1) find the number of k objects in the sample that are closest to the unknown object by using the feature vector distance metric; and (2) make a vote of the k number of the closest object to determine the class of the unknown object. The accuracy of k-NN depends on the distance metric and the value of k. Generally, the distance metric used is the Euclidean distance [10] as shown by Equation (2). If two vectors are known: x = [x_1, x_2, x_3, ... x_n] and y = [y_1, y_2, y_3, ... y_n], then the distance of the two vectors is:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

4. Results and Discussions

The system consisted of two subsystems, namely class builder subsystem that is intended to form a knowledge database, and classification subsystem that is used to classify the unknown shrimp categories.

4.1. Class Builder Subsystem

There are several buttons in the class builder subsystem that can be used by the developer to build a knowledge database. Sample images were used to build a database consisting of feature vectors and classes. The interface of the class builder subsystem can be seen in Figure 5.

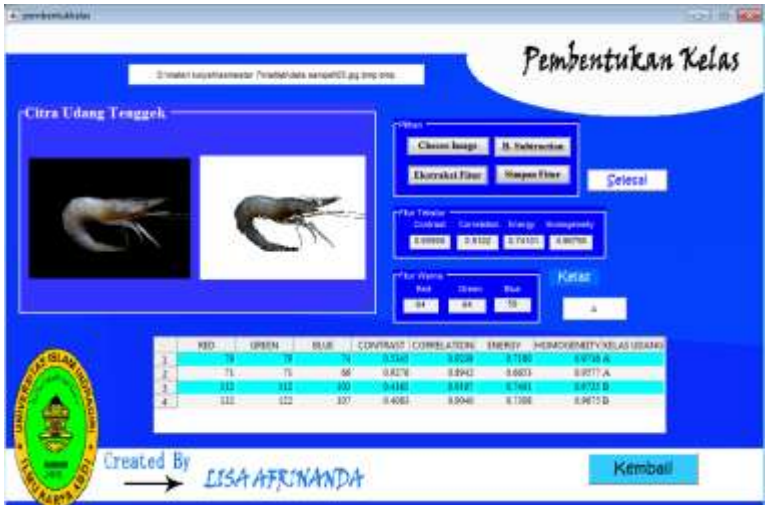

Figure 5. Class Builder Subsystem Interface

4.2. Classification Subsystem

Classification subsystem aims to make a classification or detection of the image of white shrimps. There are several buttons in the classification subsystem that can be used by the user to perform the classification of shrimps. The interface of the classification subsystem can be seen in Figure 6.

Figure 6. Classification Subsystem Interface

4.3. Results and Discussion

A system evaluation was executed to measure the performance of the detection system. The evaluation was performed by several different parameters: k = 1, k = 3, k = 5, and k = 7. The validation test was carried out by the holdout method, where 60 images were used as training data and 30 images as test data.

Table 2. Confusion Matrix for k = 1

f_{ij}		Predicted Class	
		Class A	Class B
Original Class	Class A	10	5
	Class B	3	12

The confusion matrix can be used to measure the performance of classification. The configuration of confusion matrix at k = 1 is shown in Table 2. Based on Table 2, it is obtained that from 30 test images, there are 22 images correctly classified, while the remaining 8 images are misclassified. Thus, it can be calculated that the accuracy is 73.33%.

Table 3. Confusion Matrix for k = 3

f_{ij}		Predicted Class	
		Class A	Class B
Original Class	Class A	10	5
	Class B	4	11

Based on Table 3, the classification results using k-NN for k = 3 with the test data of 30 images show the test images that are correctly classified by the system are 21 test images, while the remaining 9 test images are classified wrongly by the system. Thus, it can be seen that the accuracy obtained is 70%.

Table 4. Confusion Matrix for k = 5

f_{ij}		Predicted Class	
		Class A	Class B
Original Class	Class A	10	5
	Class B	4	11

Based on Table 4, the classification results using k-NN for k = 5 with the test data of 30 images demonstrate that the testing images correctly classified by the system are 21 test images, while the remaining 9 testing images are misclassified. Thus, it can be concluded that the accuracy obtained is 70%.

Table 5. Confusion Matrix for k = 7

f_{ij}		Predicted Class	
		Class A	Class B
Original Class	Class A	9	6
	Class B	6	9

Based on Table 5, the classification results using k-NN for k = 7 with the test data of 30 images indicate the test images that are correctly classified by the system are 18 test images, while the remaining 12 test images are misclassified. Thus, it can be calculated that the accuracy obtained is 60%.

Some examples of correctly or successfully classified white shrimps by the system can be seen in Figure 7.

Category : Class A (Natural)
Distance : 2.83
Similarity : 96.98%

Category : Class B (Poisonous)
Distance : 2,45
Similarity : 95.84%

Figure 7. White shrimps that are correctly classified

The name of the detected class of shrimps is given by the system, which consist of two possibilities, namely natural white shrimp and poisonous white shrimp. The distance value is calculated by using the Euclidean distance metric. The similarity is given by the system to indicate the percentage of similarity between the target white shrimps and sample images in the knowledge database.

If the percentage of similarity of the target objects is less than the threshold of 50%, it will be rejected by the system automatically. The testing for rejection ability of the system against other objects is shown in Figure 8.

Category : Rejected
Distance : 93,01
Similarity : 25%

Figure 8. Object that is rejected to be classified

Based on Figure 8, it is known that if the percentage of similarity between the test image and training image is less than 50%, then the classification system will perform rejection. Thus, it can be concluded that the developed system is able to resist foreign objects.

5. Conclusion

The poisonous white shrimp detector system has been successfully developed using the k-Nearest Neighbor method. The features used were RGB colors (red, green, and blue) and texture (energy, contrast, correlation, and homogeneity). The level of similarity was measured through the feature vector distance by using the Euclidean distance metric. The prediction was conducted by the system if the percentage of similarity is above 50% and rejected if otherwise. A system performance evaluation was executed by using the holdout validation method, where 60 images were used to build a knowledge database and 30 images were used for testing. The experiment was performed for several parameter values: k = 1, k = 3, k = 5, and k = 7. Based on the performance evaluation using confusion matrix, the best accuracy is 73.33% for k = 1. Nevertheless, the accuracy still needs to be improved. Optimal accuracy could not be achieved due to several factors: (1) the collection of shrimp samples was not done simultaneously; (2) the quality of the camera was considerably low; and (3) the image segmentation process was not excellent. Therefore, for better performance of the system in the future, it is suggested to overcome the abovementioned factors.

References

[1] D. Novita, T. R. Ferasyi, and Z. A. Muchlisin, “Intensitas dan Prevalensi Ektoparasit Pada Udang Pisang (Penaeus sp .) Yang Berasal dari Tambak Budidaya di Pantai Barat Aceh,” *Jurnal Ilmiah Mahasiswa Kelautan dan Perikanan Unsyiah*, vol. 1, no. 3, pp. 268–279, 2016.

[2] M. Prashanth and C. Indranil, “Journal of Medical and Health Sciences Food Poisoning : Illness Ranges from Relatively Mild Through To Life Threatening,” *Journal of Medical and Health Sciences Food*, vol. 5, no. 4, pp. 1–19, 2016.

[3] Abdullah, Usman, and M. Efendi, “Sistem Klasifikasi Kualitas Kopra berdasarkan Warna dan tekstur Menggunakan Metode Nearest Classifier (NMC),” *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 4, no. 4, pp. 297–303, 2017.

[4] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN Classification With Different Numbers of Nearest Neighbors,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2017.

[5] P. Galdi and R. Tagliaferri, “Data Mining: Accuracy and Error Measures for Classification and Prediction,” in *Reference Module in Life Sciences*, no. January, Elsevier, 2018, pp. 1–14.

[6] J. M. Kirimi and C. A. Moturi, “Application of Data Mining Classification in Employee Performance Prediction,” *International Journal of Computer Applications*, vol. 146, no. 7,

pp. 28–35, 2016.

[7] N. C. S. Reddy, K. S. Prasad, and A. Mounika, "Classification Algorithms on Datamining : A Study," *International Journal of Computer Intelligence Research*, vol. 13, no. 8, pp. 2135–2142, 2017.

[8] P. Sagar, Prinima, and Indu, "Analysis of Prediction Techniques based on Classification and Regression," *International Journal of Computer Applications*, vol. 163, no. 7, pp. 47–51, 2017.

[9] M. Kibanov, M. Becker, J. Mueller, M. Atzmueller, A. Hotho, and G. Stumme, "Adaptive kNN using Expected Accuracy for Classification of Geo-Spatial Data," in *Proceedings of Symposium on Applied Computing (SAC)*, 2017, pp. 1–9.

[10] E. López-Iñesta, F. Grimaldo, and M. Arevalillo-Herráez, "Classification similarity learning using feature-based and distance-based representations: A comparative study," *Applied Artificial Intelligence*, vol. 29, no. 5, pp. 445–458, 2015.