Classification of color features in butterflies using the Support Vector Machine (SVM)

Dhian Satria Yudha Kartika
Study Program of Information System
Faculty of Computer Science
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Surabaya, Indonesia
dhian.satria@upnjatim.ac.id

Hendra Maulana
Study Program of Informatic
Faculty of Computer Science
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Surabaya, Indonesia
hendra.maulana.if@upnjatim.ac.id

***Abstract*— Research in digital images is expanding widely and includes several sectors. One sector currently being carried out research is insects; specifically, butterflies are used as a dataset. Eight hundred ninety types of butterflies divided into ten classes were used as a dataset and classified based on color. Ten types of butterflies include Danaus plexippus, Heliconius charities, Heliconius erato, Junonia coenia, Lycaena phlaeas, Nymphalis antiopa, Papilio cresphontes, Pieris rapae, Vanessa atalanta, Vanessa cardui. The process of extracting color features on butterfly wings uses the Red, Green, Blue (RGB) method to become Hue, Saturation, Value (HSV) color space with color quantization (CQ). The purpose of adding CQ is that the computation process is carried out faster without reducing the image's information. The image will be converted into a size of 3 pixels and then normalized during the extraction process of color features. Normalizing the dataset has the aim that the value ranges in the dataset have the same value. The 890 butterfly dataset was classified using the Support Vector Machine (SVM) method. Based on this research process, the accuracy of the 256x160 pixel size is 72%, the 420x315 pixel is 75%, and the 768x576 pixel is 75%. The test results on a system with a 768x576 pixel get the highest marks with a precision value of 74.6%, a recall of 72%, and an f-measure of 73.2%.**

Keywords—image processing; classification; butterflies; color features; features extraction

I. INTRODUCTION

Research in the field of digital images is overgrowing. Digital era 4.0 is currently being used for several studies, for example, in the field of soil science, to determine the qualifications and characteristics of soil based on digital image processing [1]. In another study, digital images were used for taxonomy processing based on insects [2]. The implementation of research on digital images is very broad. A number of scientists use image processing to identify insect species [3]; digital images in our lives are developed to determine the location of vehicle license plates [4] [5]. This implementation of determining vehicle number plates can be used for automatic parking applications; digital images are also used for face detection in humans [6] and detection of motor vehicle license plates [7].

In digital image processing used in the field of insect identification, researchers have carried out more than 170,000 species. The types of insects discovered by various scientists have unique and different colors, various patterns on their wings, and wing shapes that have differences between species. Butterflies are a type of insect that is unique in terms of color, pattern, and texture on the part of its body, and this uniqueness is the value of a butterfly so that it can be distinguished by species. The uniqueness of these butterfly insects is the basis for some researchers taking the butterfly dataset in their research, as was done by Wang [8]. Butterflies have striking variations in wing shape, especially in terms of texture and color. Different shapes, some are sharp, blunt, elliptical, and round on the wing, are advantages to be able to do research. Butterflies are a type of Lepidoptera species. Butterflies distinguish types of butterflies based on their active time and physical characteristics. Butterflies that are active during the day have wings that are wider and brighter. On the other hand, butterflies that are active at night will tend to have smaller wings and are more dominant in one color [9].

Image processing is also used for automatic identification of leaf types based on texture, shape, and color features using the Kernel-based Particle Swarm Optimization (PSO) and Fuzzy-Relevance Vector Machine (FRVM) method for classification to get high accuracy of 99.87%, sensitivity, and the specifications for the image are 99.5% and 99.9%. Preprocessing is done with cellular automata (CA), which is useful for reducing noise, histogram equalization, and ROI segmentation and is used to increase the contrast and quality values in the image. Each part will perform feature extraction, including color features using RGB to HSV and texture feature extraction using Gabor-based Haralick. The feature extraction that has been produced is all processed using the selection feature using the PSO method. The resulting values are then classified using FRVM to produce high accuracy [10].

The basic components commonly used for research in the field of the image include color, shape, and texture features [11]. This is what underlies this study using the butterfly dataset and is a continuation of Josiah Wang's research. Josiah Wang conducted research to recognize objects with natural

language through writing or object descriptions. The output generated from the object description can determine the type of butterfly species as well as classify them based on species [8]. The reason researchers use butterflies as a dataset is that they represent two components in image processing, namely color and texture, and between species of butterflies that have unique textures and various color patterns.

The dataset in this study continues previous research conducted by Wang et al., The focus of this study is to carry out the extraction process of color features consisting of Red, Green, Blue (RGB) on the butterfly dataset. The process of extracting color features using the RGB method moves into the Hue, Saturation, Value (HSV) color space. The proposed classification process is to use a Support Vector Machine (SVM) as in research conducted by previous authors [13]. This classification method is able to show optimal results used in previous studies [12]. The system testing process will show the performance of the proposed application to calculate the precision, recall, and f-measure values. Hopefully, this research is able to make a contribution, especially in the field of digital images with a butterfly dataset.

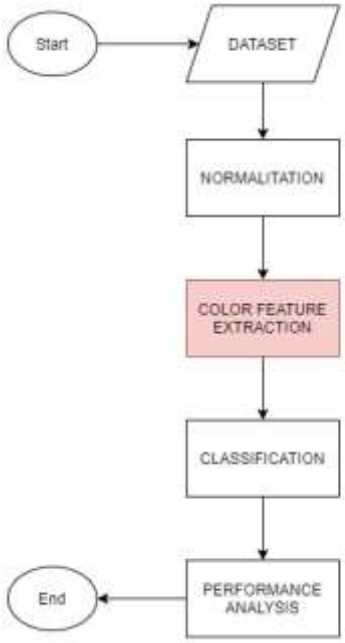


Fig. 1. Research Methodology

II. METHOD

The research methodology process in Figure 1 starts from collecting datasets, normalizing, extracting color features, classifying, and testing the performance of the whole system.

A. Dataset

The dataset used is 890 butterflies which are divided into ten classes. As in Figure 1 of the dataset, these butterfly species include Danaus plexippus, Heliconius charities, Heliconius erato, Junonia coenia, Lycaena phlaeas, Nymphalis antiopa, Papilio cresphontes, Pieris rapae, Vanessa atalanta, Vanessa cardui.

Each class has 89 data with various backgrounds and sizes. In the preprocessing stage, the dataset will clean the amount of noise, including leaves, flowers, and logs on which the butterfly rests. The image size will also be adjusted to fit the frame. The butterfly will be cut according to the outline of the object. As in Figure 2, it is mentioned the original image/photo of a butterfly in the open, then combined with masking to remove noise.

Josiah Wang, in his research, provided an image of the segmentation results that had been carried out in the form of an image mask. The segmentation result mask in Josiah Wang's research will be used for preprocessing, namely combining two images between the original image and the segmentation result (Mask) so as to produce a butterfly image according to its shape without a background..

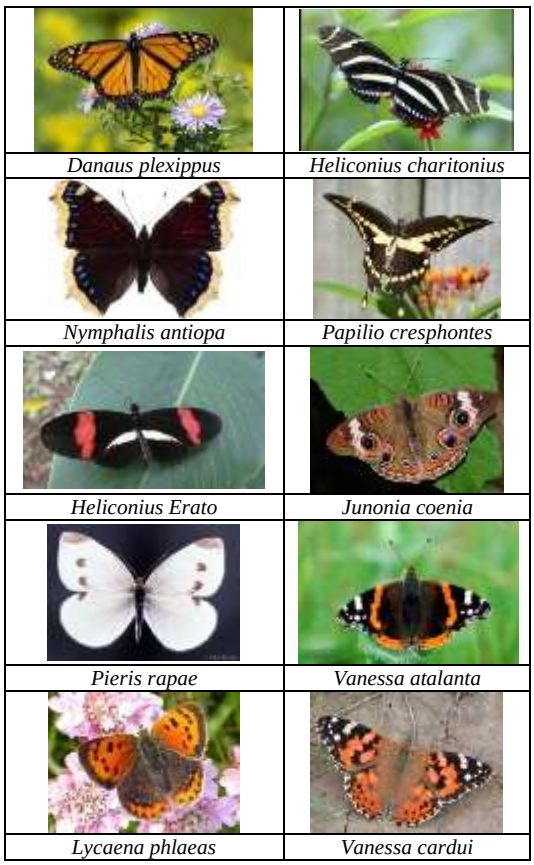
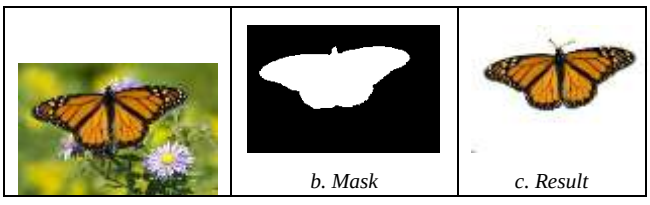


Fig. 2. Dataset



a. *Original*		

Fig. 3. Preprocessing

B. Normalization

The normalization process is a process used to prepare data so that it can be used to meet user needs and support the next process to get better results. The importance of the normalization process so that the input has a uniform value. The value in question is the image size with the same pixel.

Because the input of data that is not the same will affect the resulting process and also affect the final result in the application, several scenarios of the dataset normalization process include removing the image background (white color), converting the image to black-white, reconstructing the image by adding a function, cropping and resizing the image according to the desired pixel size. There are several scenarios that are done when resizing an image. Namely, the image size becomes 256 x 160 pixels, resizes it to 420 x 315 pixels, and changes the image with a size of 768 x 576 pixels. Each pixel size will then perform color feature extraction, classification, and result testing.

C. Color Feature Extraction

The color feature extraction process in this study uses the image value conversion process from RGB to HSV color space. In digital images, to be able to distinguish objects with certain colors, you can use the Hue value, which represents the colors red, orange, yellow, blue, green, and purple. Hue values can be combined with saturation and value values, which represent the brightness of a color. The color space, which was originally a cube in the RGB color space, will be converted into a cone-shaped color space (HSV). The first step before extracting is to find the maximum and minimum values of R, G, B. After finding the maximum and minimum values, then calculate the Hue, Saturation, and Value values.

The results from RGB extraction into HSV color space will then be reduced to reduce computation without reducing image quality (Youssef, 2012). One of the quantization techniques is by separating unused numbers. The proposed method divides the color feature extraction process into 72 parts. This extraction process shows a good and suitable combination to do. The results show results that are in line with expectations. The result of feature extraction from 72 parts is divided into eight parts with a value of Hue, three parts with a value of saturation, and three parts with a value of saturation. The process of color feature extraction is as in equation (1) (2) (3).

$$H = \begin{cases} 0\ if\ h \in [316,20] \\ 1if\ h \in [21,40] \\ 2if\ h \in [41,75] \\ 3if\ h \in [76,155] \\ 4if\ h \in [156,190] \\ 5if\ h \in [191,270] \\ 6if\ h \in [271,295] \\ 7if\ h \in [296,315] \end{cases} \quad (1)$$

$$S = \begin{cases} 0if\ s \in [0,0,2] \\ 1if\ s \in [0.2,0.7] \\ 2if\ s \in [0.7,1] \end{cases} \quad (2)$$

$$V = \begin{cases} 0if\ v \in [0,0,2] \\ 1if\ v \in [0.2,0.7] \\ 2if\ v \in [0.7,1] \end{cases} \quad (3)$$

D. Support Vector Machine

Before the testing and analysis process is carried out in this study, a classification process will be carried out. The classification process on the results of color feature extraction was carried out using the 2015 Matlab application. A total of 890 data from the extraction of color features had been normalized and divided into 3-pixel sizes. In the classification process, 890 data already had labels or classes for each type of butterfly. Each class consists of 89 data. Before the classification process is carried out, the entire data will be divided into two, namely training data and testing data. Training data is 790 data, and testing data is 100 data. The training data is used as a reference for the testing data testing process.

Data that is ready for classification uses a Support Vector Machine (SVM). The testing process with the SVM method in previous studies mentioned in reference, the results provide the maximum value. So that in this study using the same method in the classification process.

E. Performance Analysis

The system testing process is to test the system that has been made. The testing process of this system will measure the accuracy value of the classification results. In another study, the measurement process also calculates the precision, recall, and f-measure values. The first testing process is carried out by dividing the training data and testing data, then calculating the similarity distance between the data (euclidean distance). The data is divided into ten classes, each with 79 training data and ten testing data. After the data class group is found, then sorting (sorting) the data from the smallest value to the largest. Sorting data aims to make it easier to perform checks and also calculations between correct data and wrong data, whether the data should fall into the intended class or not.

The next process is to calculate the similarity value between testing data and training data. Calculation of similarity by calculating the correct data and should enter the class in question (true positive), will be labeled with the number 1, while the wrong data in a class (true negative) will be labeled with the number 0. true positive. In addition to calculating the true positive value, all data that the system has successfully retrieved (true positive + false positive) is also counted. After knowing the correct amount of data in each class (true

positive), the next step can be used to calculate the accuracy, precision, recall, and f-measure values.

III. Result and Discussion

The test results and classification based on the previous explanation will be presented in more detail in this section. The results of the color feature extraction are as described in Table 1 below.

TABLE I. RESULT OF COLOR FEATURE EXTRACTION

No	*f1*	...	*f10*	...	*f66*	...	*Jenis Kupu*
1	44	...	22	...	17	...	1
2	19	...	20	...	165	...	1
3	5	...	5	...	66	...	1
4	12	...	15	...	44	...	1
5	421	...	100	...	70	...	1
...	...	...	...	...	...	...	...
250	0	...	0	...	30	...	3
251	832	...	825	...	123	...	3
252	573	...	863	...	82	...	3
253	288	...	75	...	15	...	3
254	490	...	888	...	195	...	3
...	...	...	...	...	...	...	...
887	512	...	85	...	13	...	10
888	53	...	63	...	182	...	10
889	48	...	60	...	37	...	10
890	10	...	15	...	72	...	10

TABLE II. RESULT OF CONFUSION MATRIX FEATURE

Class	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Acc (%)*
1	**7**	0	2	1	0	0	0	0	0	0	70
2	3	**0**	0	0	2	1	0	1	1	2	0
3	1	0	**9**	0	0	0	0	0	0	0	90
4	5	1	0	**1**	2	0	0	0	1	0	10
5	0	3	4	0	**3**	0	0	0	0	0	30
6	2	1	0	0	1	**2**	1	1	0	2	20
7	1	1	1	0	1	0	**5**	0	0	1	50
8	0	2	0	1	0	1	1	**5**	0	0	50
9	1	1	0	0	0	1	1	0	**2**	4	20
10	1	0	2	0	1	1	0	0	0	**5**	50

TABLE III. NILAI PERFORMA

Fitur	**Precision**	**Recall**	**F-Measure**
256x160	76,4	75	75,7
420x315	73,4	70	71,7
768x576	73	71	72

The results of the author's color feature extraction appear at a size of 768x576 pixels. A total of 890 datasets were obtained as many as 72 color feature extraction results were divided into ten classes. The next process measures the performance of the classification process with a confusion matrix. A confusion matrix is also often called an error matrix. The confusion matrix provides information on the comparison of the classification results carried out by the system (model) with the actual classification results. The confusion matrix is in the form of a matrix table that describes the performance of the classification model on a series of test data whose true value is known.

The effectiveness of a system is measured by applying the concept of information retrieval systems. There are three types of basic measurements that are often used, namely precision, recall, and f-measure. Precision is the amount of relevant data taken by the system to be compared with the overall data. The recall process is the amount of relevant data taken by the system in comparison with all relevant data. In the precision and recall process carried out, the results will be entered into a confusion matrix called True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Based on table 3, the results of the performance analysis show that the features with a size of 256x160 pixels have a precision value of 76.6, a recall of 75, and an f-measure of 75.7. Features with a size of 420x315 pixels have a precision value of 73.4, a recall of 70, and an f-measure of 71.7. A feature with a size of 768x576 pixels has a precision value of 73, a recall of 71, and an f-measure of 72.

IV. Conclusion

The research was conducted using 890 butterfly datasets and then classified using a Support Vector Machine (SVM), and analytic testing obtained good results as shown in the test results, the highest precision value of 76.4, recall of 75, and f-measure of 75. , 7. The highest results are obtained in the dataset image size 256x160 pixels. The normalization process before color feature extraction is important so that the data input to be processed has the same standardization. So results of color feature extraction will have more valid results. This research can be developed by adding or using other feature extraction methods and methods for the classification process.

References

[1] Tang, C. S., Lin, L., Cheng, Q., Zhu, C., Wang, D. W., Lin, Z. Y., & Shi, B. (2020). Quantification and characterizing of soil microstructure features by image processing technique. Computers and Geotechnics, 128, 103817.

[2] Joshi, N., Ghate, H., & Padhye, S. (2020). Digital image post-processing techniques for taxonomic publications with reference to insects. Journal of Threatened Taxa, 12(1), 15173-15180.

[3] Kaya, Y., Kayci, L., & Uyar, M. (2015). Automatic identification of butterfly species based on local binary patterns and artificial neural networks. Applied Soft Computing Journal, 28, 132–137. https://doi.org/10.1016/j.asoc.2014.11.046

[4] Akbar, F. A., & Maulana, H. (2018, December). Detection of Indonesian Vehicle Plate Location using Harris Corner Feature Detector Method. In International Conference on Science and Technology (ICST 2018) (pp. 877-881). Atlantis Press.

[5] Maulana, H., Saputra, W. S., & Alit, R. (2020, October). We are combining Region-Based and Point-Based Algorithm to Detect Vehicle

Plate Location. In 2020 6th Information Technology International Seminar (ITIS) (pp. 183-187). IEEE.

[6] Yunanto, A. A., & Herumurti, D. (2016, October). Face recognition based on Extended Symmetric Local Graph Structure. In 2016 International Conference on Information & Communication Technology and Systems (ICTS) (pp. 80-84). IEEE.

[7] Maulana, H., Herumurti, D., & Yuniarti, A. (2018). Metode Maximally Stable Extremal Regions Dan Harris Corner Untuk Mendeteksi Lokasi Plat Nomor Kendaraan Bermotor. SCAN-Jurnal Teknologi Informasi dan Komunikasi, 13(1), 29-38

[8] Wang, J., Markert, K., & Everingham, M. (2009). They are learning models for object recognition from natural language descriptions. Learning, 2.1-2.11. Retrieved from http://eprints.pascal-network.org/archive/00006257/

[9] Kaya, Y., Kayci, L., & Uyar, M. (2015). Automatic identification of butterfly species based on local binary patterns and artificial neural networks. Applied Soft Computing Journal, 28, 132–137. https://doi.org/10.1016/j.asoc.2014.11.046

[10] VijayaLakshmi, B., & Mohan, V. (2016). Kernel-based PSO and FRVM: An automatic plant leaf type detection using texture, shape, and color features. Computers and Electronics in Agriculture, 125, 99–112. https://doi.org/10.1016/j.compag.2016.04.033

[11] Kartika, D. S. Y., Herumurti, D., Rahmat, B., Yuniarti, A., Maulana, H., & Anggraeny, F. T. (2020, October). We are combining Extraction Butterfly Image using Color, Texture, and Form Features. In 2020 6th Information Technology International Seminar (ITIS) (pp. 98-102). IEEE.

[12] Junhua, C., & Jing, L. (2012). Research on Color Image Classification Based on HSV Color Space. 2012 Second International Conference on Instrumentation, Measurement, Computer, Communication, and Control, 255(3), 944–947. https://doi.org/10.1109/IMCCC.2012.226

[13] Kartika, D. S. Y., Herumurti, D., Rahmat, B., Yuniarti, A., Maulana, H., & Anggraeny, F. T. (2020, October). We are combining Extraction Butterfly Image using Color, Texture, and Form Features. In 2020 6th Information Technology International Seminar (ITIS) (pp. 98-102). IEEE.