

BUTTERFLY CLASSIFICATION

CRISTIAN DA COSTA ROCHA
LINGCONG ZHAO
NTAMBAAZI TONNY
(Project for Introduction of MATLAB)



ABSTRACT

1.INTRODUCTION	2
2. METHODOLOGY	3
Methodology	3
Processing Steps	4
3. RESULTS	5
Overall Result	6
Worst 2 Analysis	7
4. MATLAB FUNCTIONS	11
5. REFERENCES	13

1. INTRODUCTION

The project seeks to classify individual butterflies in a provided study set of different butterflies. Classification in this case was done basing on results of color segmentation done by the K-means clustering pattern recognition. According to Nijhout (2001), the color patterns on the wings of butterflies are unique among animal color patterns in that the elements that make up the overall pattern are used for their classification. Results from the color clusters in the different butterfly species were therefore used for grouping the butterflies basing on differences in the observed colors as well as the volumes of a given color in the different butterfly species.

There were ten data sets with differences in color, shape, and texture provided for study. Similar data sets such as set 01 and set 10 had similar color and thus could differently be classified basing on their shape or texture. Color was used to classify the butterflies and below is their respective color characterization.

Data set	01	02	03	04	05	06	07	08	09	10
Example Pictures										
Major color	Orange	Black	Black	Grey	Orange	Reddish brown	Black	White	Black	Orange
Middle color	Black	White	Red/white	Orange	Grey	white	Yellow/white	Black	Orange	Black
Minor color	White	brown/reddish	White/black	White	white	Black	N/A	grey	white	white

Table 1: Main colors of butterflies

In this study, a machine learning algorithm was designed to quantify color and classify the butterflies after clustering; however, results obtained were greatly affected by input factors such as unclear blurred images, butterfly data sets having similar colors such as in dataset 01 and 10, and also incomplete images. The similarity in the data set colors can be explained by a similarity in the illumination condition and also the angle at which the picture of the butterfly was taken.

2. METHODOLOGY

2.1 Methodology

There are many ways of classifying the butterflies with methods which could work pretty well. Techniques such as SIFT or SURF descriptors could have a really high accuracy on this dataset. However, our main goal is to recognize the butterflies using only color features. We are doing that for two reasons; 1) firstly, in this case, each kind of butterfly has a lot of information based on color, and we can use this information for recognising the patterns for each class of butterfly, and, 2) secondly, we would like to learn more about color features and how to use them properly. In fact, we tried to use other features such as shape descriptors but we did not have good results with it, given that the dataset has many differences in size and in position for the butterflies in the same class.

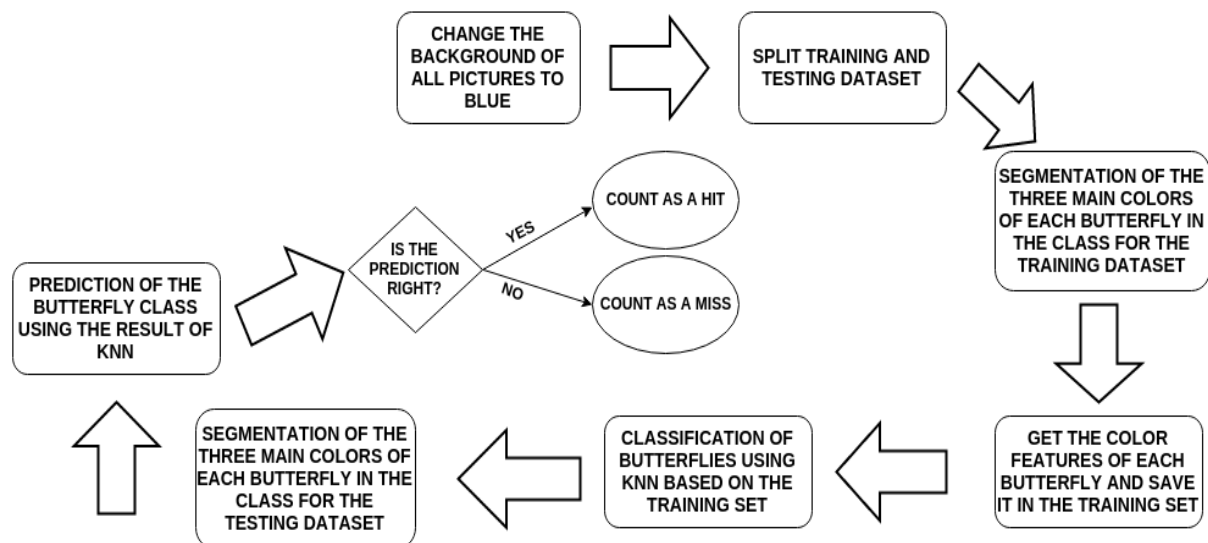


Figure 1: Butterfly Classification flow

As the classification of the butterflies is a complex problems with a lot of data involved, we decided to use Machine Learning due to its convenience and agility on software development. Machine learning can be a bit tricky sometimes if you do not use a training set which reflects the reality of the whole dataset or if the wrong method for classification is chosen. Probably we could increase our accuracy only doing data analysis and trying to improve the machine learning part of the project, but as it is not our main goal, we decided to do not try to increase the accuracy of the machine learning algorithm in order to get better results on our classification. Therefore, we did not develop all the machine learning algorithms, we just used the

built-in functions from Matlab for K-nearest neighbor classification and prediction, which work pretty well for our need.

2.2 Processing Steps

The color space chosen for color segmentation was the $L^*a^*b^*$ because it had the best results for the classification in our case. In figure 1 it is possible to see the flow of the butterfly classification.

- ❑ Firstly, we changed the background of all the butterflies to blue as the black parts of the butterflies were being clustered as part of background as well.
- ❑ Then, we splitted the dataset, thirty pictures of each class for training and the remaining photos of each class for testing. For each butterfly we get the three main colors which were clustered using k-means, and for each of the colors we get its mean and its standard deviation for the L^* , a^* and b^* channels. The mean is a way of getting the central tendency estimation of the color, which could be also done using the median, but as finding the median of the color can be a complex problem we just use the mean. And with the standard deviation we can have some information on how much this color changes in the cluster. We tried to not take the L^* channel in consideration as there is a lot of different illumination settings for each butterfly, but our clustering algorithm was having problems for identifying the difference between black and white in the butterflies, which have really close values if you do not take L^* in consideration.
- ❑ After classifying the butterflies using the KNN algorithm for the training set, we calculate the same color features for each butterfly in all the classes in the testing part.
- ❑ Then, we use the prediction function from Matlab to make a smart guess of what is the class of the butterfly being analysed. If the prediction function gives us the right result, we compute it as a hit, otherwise we compute it as a miss.

With all the misses and the hits we can calculate what is the accuracy of our classification method.

3. RESULTS

Butterfly class	Number of hits	Number of misses	Accuracy	False positive rate ranking
01	37	14	73%	7% Class 9, 7% Class 5
02	45	18	71%	20% Class 7, 4% Class 6
03	26	5	84%	10% Class 9, 3% Class 4
04	51	9	85%	7% Class 10, 3% Class 9
05	40	18	69%	13% Class 10, 8% Class 1
06	56	14	80%	10% Class 4, 7% Class 2
07	36	23	61%	34% Class 2, 5% Class 6
08	22	3	88%	12% Class 6
09	49	11	82%	8% Class 1, 3% Class 5
10	29	25	54%	24% Class 4, 6% Class 5
SUM	391	140	73.6%	

Table 2: Accuracy for Each Class

3.1 Overall Result

In table 2 is possible to see the accuracy for all the ten classes. The classes which got the best accuracy are the classes 04 and 08 due to the colors particularity of them. For class two, even though there is not any other class with the same colors, it did not have a really good accuracy because we are taking the L^* in consideration in the features for training and in different illuminant situations the yellow part of the butterfly 07 can be confused with the white part of the butterfly 02. The class which got the worst results for accuracy is the class 10. This is one of the hardest butterflies for clustering because the colors in some parts of the butterfly are not well defined. Also, the colors of this butterfly can be easily confused with the colors from butterflies 04 and 05.

Accuracy for the whole dataset	73.6%
Total number of hits	391
Total number of misses	140

Table 3: Overall accuracy

The classes with the worst results for accuracy have the same color patterns. This could be improved with a better clustering part, which could classify colors which are similar without taking in consideration the illumination of each picture. Unfortunately we had to consider illumination for the clustering part, as it had been already mentioned, the black and white colors are close if we not consider the L^* channel.

In conclusion, with this approach, just computing the color features of the butterflies we got 73.6% of accuracy for the whole dataset. This accuracy, in our opinion, is a good result given that the dataset has a lot of problems with butterflies with really different illumination situations, different positions, different sizes and for some classes there are some butterflies which have some missing parts. It is clear, however, that the approach to get the best results is the one which can compute not only color features, but other features such as texture, shape and descriptors such as SIFT and SURF.

3.2 Worst 2 Analysis

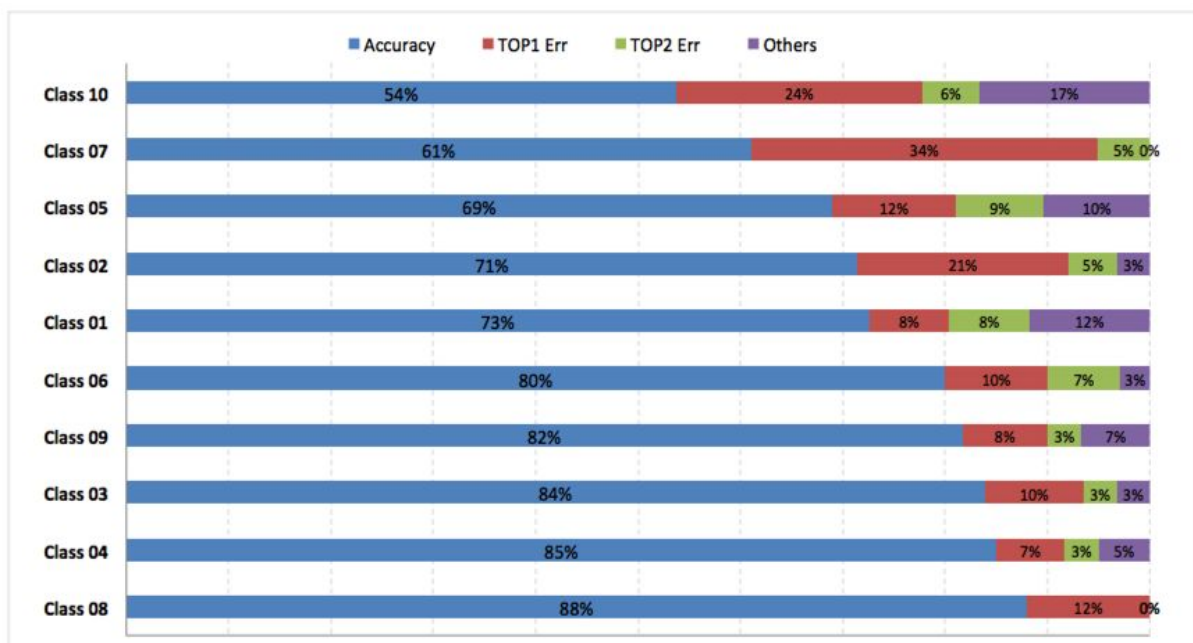


Chart1: Accuracy & Error Distribution

According to Chart 1, the worst 2 accuracy are **Class 10 (54%)**, **Class 07(61%)**.

3.2.1 Reasons of Errors & Other Trails

> Class10 Errors

24% Class10 are mistaken as Class 04 as the following reasons:

1) They share similar color information including Numbers of color, percentage of major color (orange)





	Original image	Clustered image
Class 10		
Class 04		

Chart2: An Example of Comparison Between Class 10 and 04

After Clustering their Mean Value of each components and Standard Deviation tend to be closed

2) 83% of error images has abnormal conditions which summarized as below. To find universal filters to centralize all the images need further explorations.

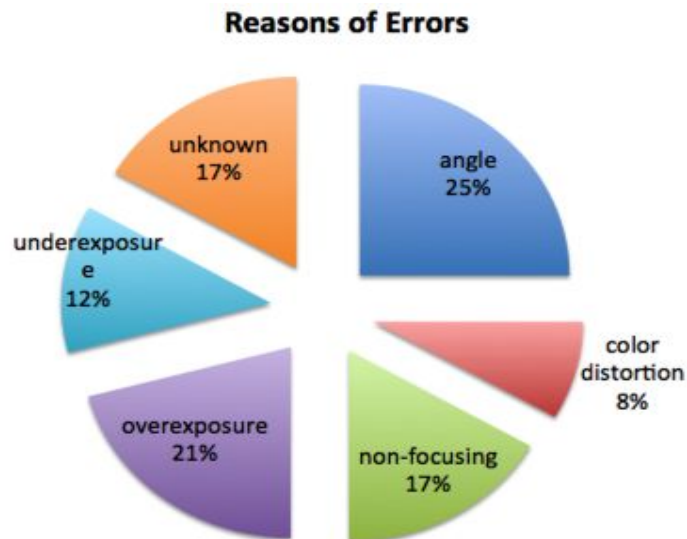


Chart3: Reasons of Errors for Class 10

Related trails to improve

Due to the big difference of lightness, focus and angle in images, the color of butterflies are not consistent but varied case by case. For seek of enhancing difference between Class 10 and 04, we used some image pre-processing method to stretch the dynamic (imadjust), remove noise (Median Filter) before converting color space to Lab, but none of them worked(accuracy 43%). The reason is big difference within class 10 itself exists too. When we stretch dynamic of all the images, the similarity within class 10 is weakened simultaneously. In order to improve this, we need to study more methodologies to calibrate images before machine learning, like white balance calibration, or consider shape, texture.

> Class 07 Errors

34% of Class 07 were mistaken as Class 02, while 20% of Class 02 were misjudged as Class 07 in our result. Reason is:

The main color of these two class are quite similar: black and white (Class 10 is yellow-white). And the two colors tend to have identical percentage. In many cases, their color turn to be the same because of illuminants, making perfect differentiation hard to reach.





	Original image	Clustered image
Class 07		
Class 02		

Chart4: An Example of Comparison Between Class 07 and 02

Related trails to improve

Several ways have been implemented to better distinguish these two classes. One of the method we tried is imtophat to extract circles in butterflies since class 2 contains more white line parts while class 10 has more circles. The accuracy is below 20% between these two butterflies due to different angle of each images. So it's difficult to choose a universal kernel to extract and thus hard to consider shape as another parameter in final classification part. This is what we want to learn more in the future (shape extraction, texture classification <function graycomatrix,graycoprops>).

4.MATLAB FUNCTIONS

Parts	Function	Description	Input	Output
Back-ground Changing	Convert black background to blue			
	Dir	List folder content	Imagepath	Struc Array
	Length	Length of array (Largest dimension)	Array	Length of Array
	Strcat	Concatenate strings horizontally	File Path, imagename	Combined string
	Size	Size(I,N) gives the number of the Nth dimension	Matrix or Array, Dimension Number	Size
	Cat	Concatenate arrays along specified dimension	Matrix or Arrays, Dimension Numbers	Merged matrix
	Imwrite	Create images under given path	Image matrixs, path	Image file
Training	Get Identified Model			
	ImageSet	Define collection of images	Image Folder, 'parameter' ('recursive' search)	A vector of image
	Read	Read data in Database Datastore	dataset,counter	Data of given count
	Makecform	Create color transformation structure	Parameter (E.x. srgb2lab)	Color transformation structure
	Applycform	Apply Color Space Transformation	Image,Transformation Structure	Color Space Changed Image
	Double	Convert data to double	other type of data (uint8)	Double

	Reshape	Reshape Array	Array, size of each dimension	New array
	Kmeans	K-means Clustering	Array, Clustering Number, 'Distance', parameter(SqEuclidean...), 'Repliate', Integer	Clustered Array, k cluster centroid locations
	Mode	Most frequent values in array (The background)	Clustered Array	The Most frequent values
	Unique	Unique values in array	Clustered Array	The Unique Values (Each Clustering name)
	Mean	Mean Value of an Array	Array	Mean Value, Indices of Min Values
Testing	Classification of Butterflies			
	Fitcknn	Fit k-nearest neighbor classifier	Training Data1, 2 /Neighbouring NO. 'Standardize' Value (Mean, Standard Deviation...)	Mdl: Classification Model
	Predict	Predict the class of butterfly using training data	Classification Model	Classification NO.

Table 4: Summary of matlab functions

REFERENCES:

NIJHOUT, H,F.,(2001). Elements of Butterfly Wing Patterns. Journal of Experimental Zology. P.1.

Color-based segmentaton using K-means:

<https://fr.mathworks.com/help/images/examples/color-based-segmentation-using-k-means-clustering.html>

K-nearest neighbor classification for matlab:

<https://fr.mathworks.com/help/stats/classificationknn-class.html>