"CREATING BUTTERFLY IDENTIFICATION APPLICATION USING ITS MORPHOLOGICAL FEATURES AND ANN"

A REPORT

SUBMITTED IN THE PARTIAL FULLFILLMENT OF THE REQUIREMENT FOR

THE AWARD OF BACHELOR OF TECHNOLOGY IN

CIVIL ENGG.

BY

UJJAWAL KHARE

ROLL NO. 16065077

UNDER THE GUIDANCE OF

DR. SAMEER SARAN
SF & Head, Geoinformatics Department
IIRS, ISRO, Dehradun

DEPARTMENT OF CIVIL ENGINEERING.
INDIAN INSTITUTE OF TECHNOLOGY
BANARAS HINDU UNIVERSITY
VARANASI – 221005.
JULY 2018

ACKNOWLEDGEMENT

I take this opportunity with much pleasure to thank all the people who have helped me through the course of my journey towards producing this thesis.

First and foremost, my utmost gratitude to my estimable guide Dr. Sameer Saran, whose sincerity and encouragement I will never forget. I sincerely thank him for his guidance, help and motivation. Apart from the subject of my research, I learnt a lot from him, which I am sure will be useful in different stages of my life. I am grateful to him for his tireless and consistent supervision and help that he rendered to me throughout the course and subsequently in preparation of this dissertation.

I acknowledge the co-operation of all faculty and staff members of Civil Engineering Department.

At last, I would like to thank my parents and my sister for their constant support to me in my every work. I would like to extend my thanks to all the people who were directly or indirectly connected with this project. Without their wishes and support it would never have been possible.

Place:

Date: July 28th, 2018 Ujjawal Khare

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Abstract

Throughout its history Earth has witnessed five major extinctions, the last one being of dinosaurs, 65 million years ago. According to the UN environment programme, earth is going through a phase of mass extinction right now with 150- 200 species of plants, insects, birds and mammals getting extinct every 24 hours. Scientists are calling this the sixth mass extinction. This fortifies the necessity for the development of a system for the conservation of biodiversity. Biodiversity forms rudimentary building blocks for the many goods and services and is intrinsic for a healthy environment and sustainable development of society.

Conservation of biodiversity is a matter of great concern for biologists and naturalists. It demands a dynamic, robust and consistent system for the accurate recognition of biological objects.

Among the plausible identification systems, identification of butterfly species has been considered in this research as it forms an intrinsic part of the ecosystem. Butterflies symbolize a healthy environment. It helps in pollination, eradication of weeds and forms an important part of the food chain.

Recognition of butterfly is an arduous task as butterflies are multi- coloured, have varying textures and easily blend with the background. Even butterflies of same species can have varying colours. Thus, shape feature detection was employed to extract the shape of its wing. Contour of its wing was extracted from the shape feature using Canny-Edge detection and finally, Branch Length Similarity (BLS) entropy has been calculated for the shape identification of butterfly. This data is then fed to a feed forward backpropagation neural network for training purposes.

The model has been integrated with an application built using kivy. The application provides an easy to use graphical user interface which can be deployed on multiple operating systems like Linux, Windows, Android and IOS.

2. Introduction

Extinction of species is becoming a matter of concern for naturalist and conversationalist nowadays as it directly affects our ecosystem [1]. On account of rapidly declining biodiversity count and vanishing species on a diurnal basis, identification of species has become very important for conservationists and naturalist to identify their hotspots, study their habitat and changes occurring in them in order to find a way to conserve them. Regular monitoring of identified species helps to study species behavior and improve management decisions [2].

With recent development in the field of image processing, computer vision and machine learning it has now become possible to identify species on the basis of feature extraction. Similar techniques have been used for text detection and recognition using artificial neural networks [3]. Neural Networks are computer systems modelled on human brain, which are capable of learning to recognize certain patterns based on supervised training. Thus, neural networks can be trained to identify species which will help in dealing with problem so called "taxonomy crisis" [4]. Until the advent of ANN technology in this field, taxonomic keys were used to identify species which required skilled taxonomist with vast knowledge of species characteristics. Each step in a taxonomic key generator gives user contrasting options to choose from which leads to the culmination of species identification after a series of time consuming iterations [5]. Neural networks can effectively supplant skilled person reducing the identification time as well. While it can be trained to identify various occurrences in our ecological systems, we have selected butterflies for their considerable role in food chains, pollination process and adding aesthetics to a healthy environment. But due to a multitude of variety in its color, texture and shape butterfly is one of the most difficult organisms to identify on any basis.

Due to its great ability to recognize patterns, a single layer neural network [5] was implemented to recognize species of butterfly. Recognition was based on shape and color features extracted from their images to create an input vector which was fed to feed forward-backpropagation neural network. BLS (Branch Length Similarity) entropy was used as shape descriptor and a normalized color variation descriptor to hold color information. BLS entropy has been implemented earlier for shape recognition and results have been quite successful in other areas [6] [7]. Color variation descriptor is a new technique which has been explained later in this text. Similar identification work on butterfly has been done previously [8] [9] but owing to less number of species involved and their inability to provide any application to biologist and naturalist this research was done. FSIM and SSIM features [10] used to identify the species of butterfly is very accurate but due to its inefficiency and more manual work it is not preferred by naturalist and biologist. This project was aimed towards developing an application based on mobile or web, which can be used by them for species identification purpose.

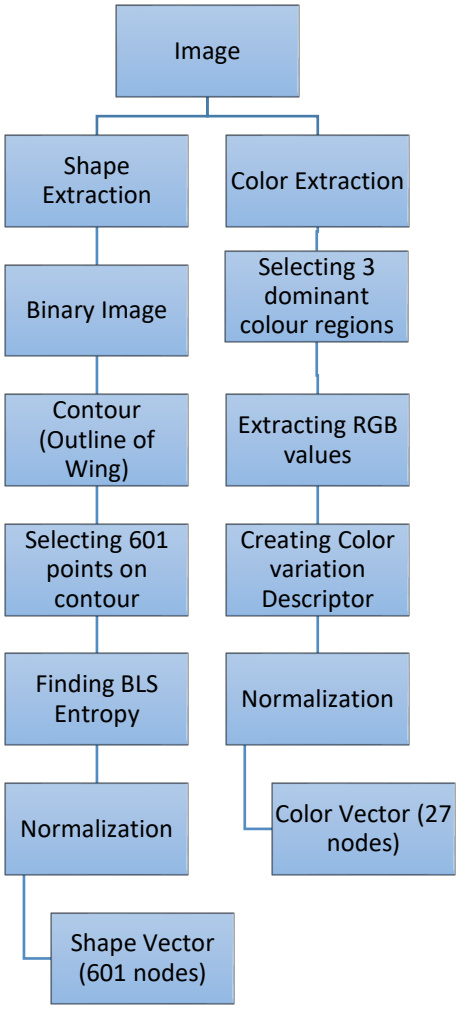
3. Methodology

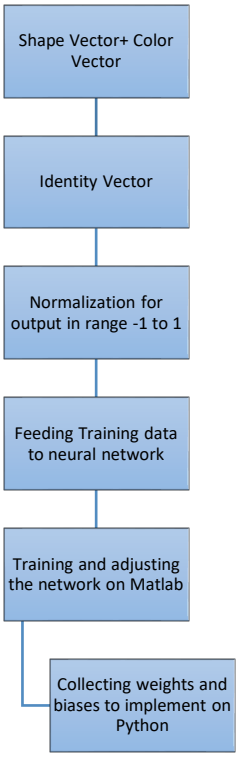
The methodology includes following steps, image processing, feature extraction, creating input data for neural network, training and adjusting the network, testing the network and creating usable application for it. Python programming language was used for its extensively useful libraries which provide functionalities for data analysis and scientific operations. Open Computer Visions (OpenCV) library of Python was employed for image processing and feature extraction. The mobile version for the application has also been built using, although the work is still under progress. Neural network was built and trained using Matlab 2010 and its weights and biases were used to create a network on python. A list of software facilities and libraries with their description is given in Table 1.

Table 1: List of Software Resources

Python 2.7	Programming Language
Anaconda 4.5.8	IDE for Python and mobile application
OpenCV (cv2)	Library for image processing
Numpy	Library for handling arrays, matrices and images
Math	For mathematical calculations and function
Matplotlib	For graphical interpretations
Kivy 1.10	For application development

The following flowchart summarizes the complete methodology adopted.





3.1 Dataset

As this project was aimed at developing usable application wherein a user can query images in real time to detect its species, the images were not retrieved using any special technique or high quality camera. Instead Google images and images from Butterflies of India website containing butterflies in their natural environment were used to create training dataset for the neural network. Due to inadequacy in openly available data, about 5-10 images for 57 species each were used. These images were used to extract shape and color features of the butterfly's wing to create a 628 –dimensional identity vector. While the color peculiarly identifies a common butterfly to an observer, shape of its wing is rather a more unique way to distinguish between higher number of species' classes. Thus to retain as much shape data as possible 601 nodes were assigned to the shape feature while 27 nodes to the color feature. Weights assigned to each feature are automatically normalized in the training phase according to the targeted outputs provided.

Figure 1: Images from ifoundbutterflies.org used for feature extraction

3.2 Feature Extraction

3.2.1 Shape Feature

Lee (2010) and Lee et al., 2010, Lee et al., 2011[11] proposed a novel method based on branch-length-similarity (BLS) entropy to recognize shape based on their outlines. A simple network, called Unit Branch Network (UBN), consisting of a single node with its branches defines BLS entropy for the node. Mathematically, BLS entropy, s, of a node can be defined as,

$$s = \frac{\sum_{i=0}^{n} p_i \log p_i}{\log n}$$

where p_i is the length probability of the i^{th} branch of the UBN which is further defined as,

$$p_i = \frac{l_i}{\sum l_i}$$

Here, l_i refers to the i^{th} branch length of the Unit Branch Network.

The image containing butterfly is sent for a preprocessing stage where we need to mask its left wing in order to retrieve BLS entropies for the species image. Masking is done using fillPoly function of OpenCV python library which is used to create a white mesh over the wing attributed to left mouse-click and this image results in a binary-greyscale image with values 255 at the wing area and 0 elsewhere. This image known as mask of the wing is sent to Canny edge function of OpenCV which identifies the outline from the two-valued image based on gradient difference.

Figure 2: Masking image using fillPoly (OpenCV)

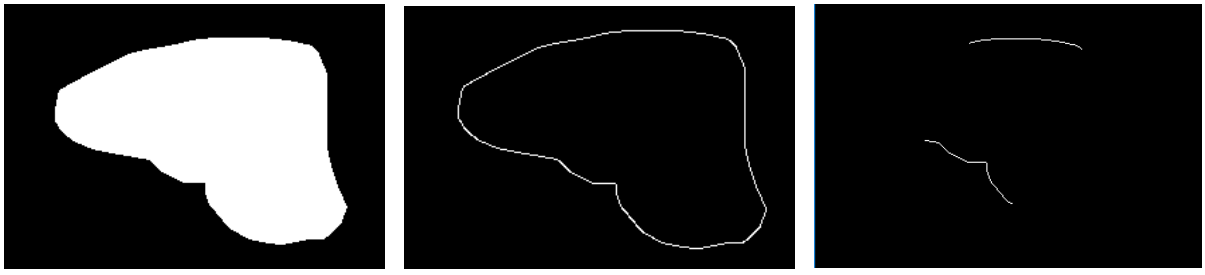


Figure 3: Binary Images, Outline and Contours near the middle region of Wing

Once outline of the wing is retrieved, we need to choose a point which will serve as the first point node for UBN, which must be common in almost all the species. Seung et. al 2012 [13], gave a method to select such a point. As we require entropies for 601 pixels on the boundary of the wing, also known as contour, we span iterator over all the points on the contour with varying step sizes calculated as,

$$step\ size = \frac{no.of\ pixels\ remaining\ to\ capture}{601}$$

for ongoing iteration. Thus after one iterations if **n** pixels are captured at step size as calculated above, no. of remaining pixels become **N**-n where **N** is the total number of pixels present on the contour initially. Hence, for the next iteration, step size will be given as,

$$step\ size = \frac{N - n}{601}$$

and the iterations continue till all 601 points are acquired. This ensures a uniformly distributed selection of points on the contour. Moreover, to avoid redundancy of pixels in the subsequent iterations, initial point of start of an iteration is incremented by one. Still remaining 2 to 3 pairs of redundancy can be ignored as they account for insignificant change in the calculated BLS entropy values. Further the entropy values are normalized using the formula

$$S_i = \frac{s_i - S_{min_i}}{S_{max_i} - S_{min_i}} \quad (0 < i < 602\ )$$

3.2.2 Colour Feature

Colour from three most dominant regions on the butterfly's wing were extracted as a colour identity for a specific species. Three colour samples for each dominant regions were taken in the decreasing order of their area coverage on the wing. Thus, 9 nodes for each dominant regions were created. In case the wing contained only two colours, first 18 nodes represented the more dominant colour and the remaining 9 the lesser dominant and in case of a single coloured butterfly, all 27 nodes contained the same colour. In this way, we could devise a certain norm for selecting dominant colours on the wings of butterfly. Once BGR values were recorded, they were normalized as,

$$Normalized\ Colour\ variation = \frac{R\ or\ B\ or\ G\ value\ (in\ range\ 0 - 255\ )}{\sum R + B + G}$$

This is a novel, effective and easy approach to identify color differences, that is, contrast in color on the wings of a butterfly. Graphical representation elucidates the difference between extracting just the color and normalized color variation vector.

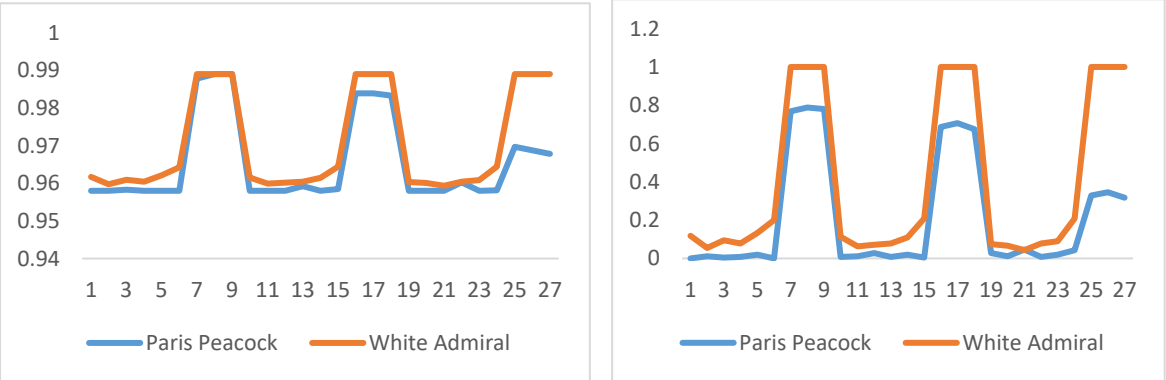


Figure 4: Graph 1: Colour Vector, Graph 2: Colour variation vector for Black-Blue (Paris Peacock) and Black-white (White Admiral) butterfly

The difference in the BGR values of the colours in case of ambiguous colour patterns such as a Black-blue and a Black-white butterfly with black as the dominant colour remains negligible up to first 18 nodes and thus, becomes difficult to distinguish as shown in the above graph 1. When we normalize the colour vector using the summation of all BGR values, the contrast is magnified and both the black and blue/white region starts showing distinction. Thus, graphs can now be easily distinguished as is clear from graph 2.

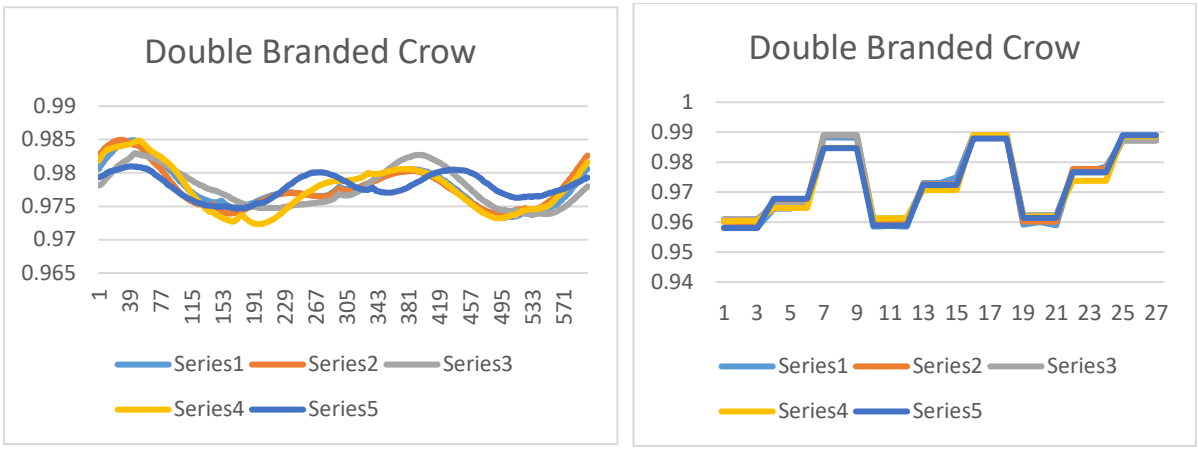
The colour variation vector, cvv, was further normalized on an approximate scale of 0.958 to 0.99 which encompasses the entropy values of all the species using the following linear relation,

$$cvv = color\ variation\ vector\ \times m + 0.958$$

And,

$$m = \frac{0.031 \times \sum color\ vectors}{Color\ vector_{max}}$$

Here $Color\ vector_{max}$ is the maximum Blue or Green or R value in the colour vector. Hence, this normalized colour vector of the species is appended to its BLS entropy to create a 628-dimentional identity vector which is fed to the forward feed-back propagation neural network for training. BLS Entropy Values and color variation vectors of 2 species are represented graphically as shown below.



Graph 1: 601 entropies and 27 colour variation nodes for Double Branded Crow (Euploea sylvester)

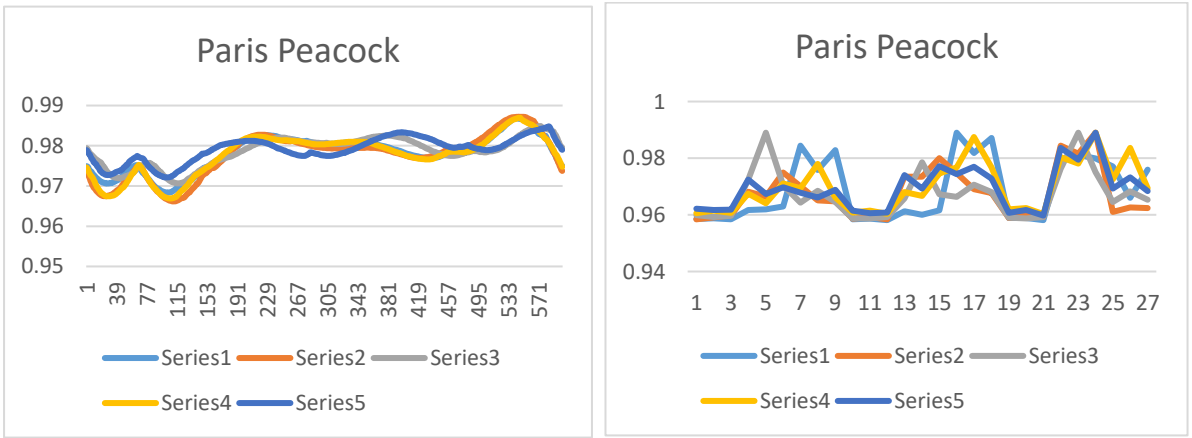


Figure 5: 601 Entropies and 27 colour variation vectors for Paris Peacock (Papilio paris)

3.3 Neural Network

The training data for neural network constitutes 57 species each having 5-10 samples. A 628-dimensional identity vector is fed as input for sample. Neural Network was drafted using MATLAB and the resulting weights and biases were imported into the python program. The Neural Network was constructed using MATLAB as it is ideal for the implementation, visualization and simulation on

account of its easy to use interface and functionalities. It can be constructed either using Neural Network Toolbox or Neural Pattern Recognition toolbox, the former being implemented for this research. Neural Network Toolbox has algorithms, pre-trained models and GUI to create, train, visualize and simulate artificial neural network. Graphical user interface of Neural Network Toolbox can be accessed through nprtool command of the MATLAB. In MATLAB input is normalized by mapminmax function so that it falls within a specified range which is by default -1 to 1 and can be modified by accessing script of the artificial neural network. Redundancy in the data is not allowed in order to improve the efficiency of neural network.

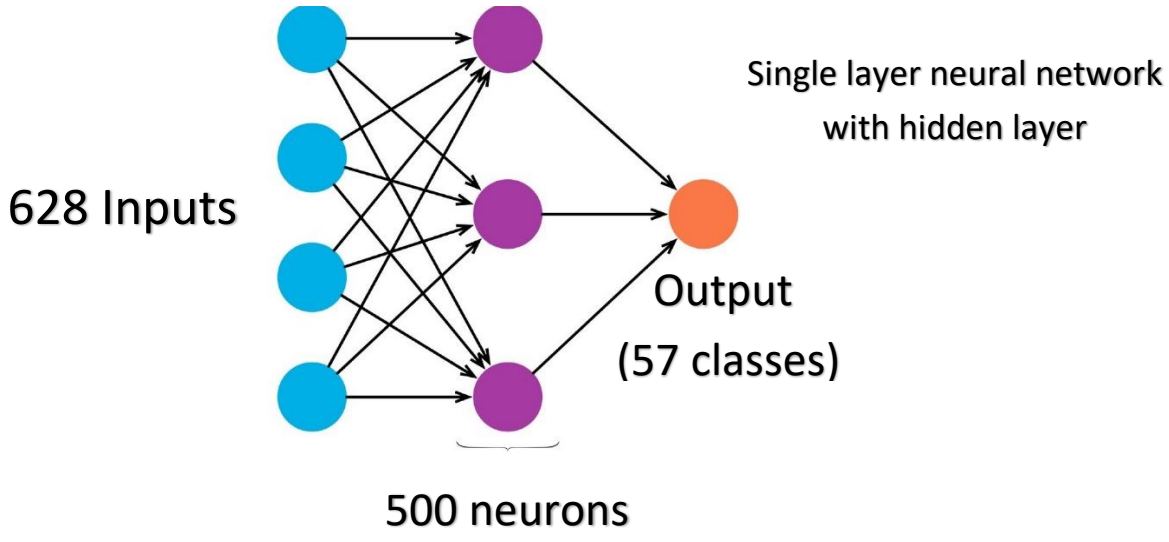
Training was based on the feed-forward back propagation model. In this type of model, each training data is progressively fed as input into the neural network and propagates forward. Output is then compared to target data to compute the error. The error from each target output is then used (back propagated) to adjust the weights and biases. TRAINRP is deployed as the training function. In this, weights and biases are updated on the basis of resilient backpropagation algorithm. Performance of the network was measured on the basis of the mean squared error. LEARNGDM (Learning through gradient descent with momentum) is used as the adaptation learning function which modifies weight for a given neuron on the basis of its input, error, weights, biases and its momentum. A layer comprising of 500 nodes is used as the hidden layer with log-sigmoid as the activation function which is given by,

$$logsgn(n) = \frac{1}{1 + e^{-x}}$$

Once all the weights and biases are assigned, all the outputs are calculated on the basis of these weights and biases. These outputs are normalized to get the required output using the following formula,

$$result = target_{max} - target_{min}\left(\frac{y - y_{min}}{y_{max} - y_{min}}\right) + target_{min}$$

The outputs of the hidden layers were multiplied by their respective weights and biases were added to them which were then encapsulated in SOTFMAX function which classifies an input vector on the basis of calculated probabilities.



4. Application

The butterfly recognition process has been integrated with an application and a simple graphical user interface has been developed. The application has been developed with the intent of making it accessible to the masses. The application has been build using kivy library which as an open source python library for developing mobile apps and other multi touch applications. It is a cross platform application and can be run on Linux, Android, IOS, Windows, OS X and Raspberry Pi. Furthermore, it is GPU accelerated.

The interface is intelligible, user selects a query image from his device, selects 3 most dominant color points, crops the wing of the butterfly and clicks on the result button and the 5 most probable resulting butterfly species in order of their ranks are displayed on the next screen. The application is still under progress and can be extensively developed further to add functionalities to it.

5. Results and Discussion

The accuracy of the network is not quite high as was anticipated in the earlier stages of the project due deficiency in the training data. Seung-Ho et. al 2012, [12] did a similar research without any color vector on good quality images, with top facing butterfly wings, for 7 species with 40 specimens of each, hence a with total of 280 specimens. In contrast to it, we were classifying 57 species with shape and color features, but with just about 5 specimens for each species, with images containing butterfly in various orientations. Despite all these complications satisfactory results were achieved with maximum accuracy of about 70% suggesting that addition of new color variation vector can surely enhance the recognition ability of the network by multifold times. Further to enhance the results of the identification application, 5 most probable results are shown with a sample image for each class, which enables user to match the query image with 5 possible samples to get the desired result.

6. Conclusion

Butterfly occurs in multifarious forms, with a myriad of patterns, colors, shapes and textures. To develop a robust identification system which can identify such high number of species, proper sample data, about 100 specimens of each sample, must be provided for the training set. Moreover, to rectify problems due to orientation, several projection techniques must be deployed and trained with individual models for each species, i.e., 57 different models for 57 species, each with a good amount of sample data. Due to unavailability of sufficient time and deficit dataset, current project shows just above satisfactory results but open ways to a more efficient identification system which can actually identify such high number of species with accuracy ranging over 90%. Nevertheless, current research presents a new color variation vector whose graphical results are quite impressive and can be fruitfully exploited for further enhancement of the system. In addition to the color variation vector, specific patterns that are peculiar to a certain species can also be added as a compound input feature which can distinguish between species with similar shape and color patterns and differing only by a minimal amount.

References

1. Rao, M., & Larsen, T. (2010). Ecological consequences of extinction. *Lessons in conservation. American Museum of Natural History, New York*.

2. Martin, J., Kitchens, W. M., & Hines, J. E. (2007). Importance of well-designed monitoring programs for the conservation of endangered species: case study of the snail kite. *Conservation Biology*, *21*(2), 472-481.

3. Bartz, C., Yang, H., & Meinel, C. (2017). STN-OCR: A single Neural Network for Text Detection and Text Recognition. *arXiv preprint arXiv:1707.08831*.

4. Dayrat, B. (2005). Towards integrative taxonomy. *Biological journal of the Linnean society*, *85*(3), 407-417.

5. Jonathan Y Clark, **Artificial neural networks for species identification by taxonomists**, Biosystems, Volume 72, Issues 1–2, 2003, Pages 131-147, ISSN 0303-2647

6. Kwon, O., & Lee, S. H. (2017). Branch length similarity entropy-based descriptors for shape representation. *Journal of the Korean Physical Society*, *71*(10), 727-732.

7. Lee, S. H., & Kang, S. H. (2016). Performance enhancement of the branch length similarity entropy descriptor for shape recognition by introducing critical points. *Journal of the Korean Physical Society*, *69*(7), 1254-1262.

8. Faruk, E. Ö., Kaya Yılmaz, K. L., & Ramazan, T. (2015). Identification of Butterfly Species by Similarity Indexes Based on Prototypes. *Journal ISSN: TBA*, *1*.

9. Hernández-Serna, A., & Jiménez-Segura, L. F. (2014). Automatic identification of species with neural networks. *PeerJ*, *2*, e563.

10. Faruk, E. Ö., Kaya Yılmaz, K. L., & Ramazan, T. (2015). Identification of Butterfly Species by Similarity Indexes Based on Prototypes. *Journal ISSN: TBA*, *1*.

11. Lee et al., 2010, S.H. Lee, P. Bardunias, N.Y. Su, **A novel approach to shape recognition using the shape outline,** J. Korean Phys. Soc., 56 (2010), pp. 1016-1019

12. Seung-Ho Kang, Su-Hee Song, Sang-Hee Lee**, Identification of butterfly species with a single neural network system**, Journal of Asia-Pacific Entomology, Volume 15, Issue 3, 2012, Pages 431-435, ISSN 1226-8615