INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY
IJARIIT
ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 6 - V11I6-1279)
Available online at: https://www.ijariit.com

A Study of Clustering Analysis in Identification of Butterfly Species

Ajaykumar R
siajaykumarr1428@gmail.com
Christ College Mysore, Karnataka

ABSTRACT

This study investigates the use of clustering analysis techniques for identifying butterfly species based on their morphological characteristics. Butterflies exhibit substantial variation in wing patterns, colors and body size which makes traditional taxonomic identification both time-consuming and error-prone. Clustering analysis provides a data-driven strategy to group individuals into putative species based on similarities in measurable features. By applying multiple clustering algorithms together with appropriate validation methods, this work evaluates the effectiveness of clustering analysis for butterfly species identification and highlights its potential applications in biodiversity research and conservation. Accurate identification of butterfly species is fundamental to biodiversity conservation, ecological monitoring, and environmental impact assessment. This study examines the efficacy of clustering methods for species identification using butterfly image data. Several algorithms, including K-means, hierarchical clustering, spectral clustering, Gaussian mixture models, and DBSCAN, are employed to partition images into species clusters. To represent discriminative visual information, feature extraction techniques such as Histogram of Oriented Gradients (HOG), Gray Level Co-Occurrence Matrix (GLCM), and Local Binary Patterns (LBP) are used to encode wing textures and shape characteristics. The quality of the resulting clusters is assessed by comparing them with known species labels, enabling a systematic evaluation of each method. The results indicate that clustering analysis offers a scalable and promising approach for automated butterfly species identification and biodiversity monitoring, while also clarifying the strengths and limitations of different clustering techniques for image-based species classification.

Keywords: *Butterfly, Identification, Species.*

1. INTRODUCTION

In an era defined by exponential data growth, the ability of efficiently process and organize information is fundamental to knowledge discovery and management. Classification serves as a critical analytical tool in this context, enabling the categorization of novel objects or phenomena based on their similarity to known standards. Whether implemented through supervised learning, which assigns inputs to predefined classes, or unsupervised learning, which discovers inherent structures within data, classification systems are essential for interpreting complex datasets and advancing our understanding of the natural world. In the fields of biodiversity and ecological research, the accurate identification of species is particularly vital for monitoring environmental health and formulating conservation strategies. Butterflies, due to their high sensitivity to habitat and climatic variations, serve as key bioindicators. However, traditional identification methods, which rely heavily on manual observation and taxonomic expertise, are often labor-intensive, time consuming, and susceptible to human error. To address these limitations, recent advancements in computer vision and machine learning have facilitated the development of automated identification systems that offer enhanced accuracy and scalability. This study investigates the application of clustering analysis a core unsupervised learning technique for the classification of butterfly species based on the image date. Unlike supervised methods that require extensive labeled datasets, clustering algorithms group data points according to intrinsic similarities, there by revealing natural patterns and relationships within the data. By leveraging this data driven approach, this research aims to improve the precision of butterfly species identification and provide deeper insights into the morphological relationships between species. Methodologically, this research utilizes robust feature extraction techniques, including Histogram of oriented Gradients (HOG), Gray level Co-occurrence Matrix (GLCM), and Local binary Patterns (LBP), to capture the distinctive textures and wing patterns unique to each species. These visual features are subsequently analyzed using clustering algorithms such as K-means, DBSCAN, and Hierarchical Divisive clustering. By integrating these methods, the study seeks to establish an effective framework for unsupervised species classification, contributing to the broader goal of automated biodiversity monitoring.

2. RELATED WORKS

Numerous studies have explored automated techniques for species identification, proposing diverse methodologies to address the challenges of morphological classification. Among these, interesting approaches for butterfly identification have emerged, ranging from traditional machine learning algorithm to advanced image processing techniques focused on feature extraction and clustering analysis.

These methods have demonstrated promising results in accurately distinguishing between species based on visual characteristics. Significant groundwork in biological image classification has been established in related domains, such as medicinal plant identification. For instance, Naresh and Nagendraswamy (2016) extensively studied the "Classification of Medicinal Plants: An approach using modified LBP with symbolic representation," the authors address key challenges in automated identification systems. Their research focuses on characterizing plant species through detailed leaf analysis, proposing novel techniques for extracting shape and texture features. Specifically, the researchers introduced feature extraction methods that leverage the axis of least inertia to describe leaf shape and local Binary Patterns (LBP) to capture texture information. Crucially, these methods were designed to be invariant to similarity transformations, ensuring robustness across varying image conditions. While originally applied to plant leaves, such feature extraction techniques, particularly the use of LBP for texture and invariant shape descriptors provide a valuable methodological framework that is highly applicable to the identification of butterfly species, where wing patterns and shapes serve as primary diagnostic features.

Building on foundational research in pattern recognition, several studies have demonstrated the efficacy of texture-based classification in biological and biometric domains. Andrian et al (2019) proposed a novel technique for butterfly identification that combines Gray Level Co-occurrence Matrix (GLCM) feature extraction to capture the intricate textural details of butterfly wings, while the KNN algorithm provided a simple yet robust mechanism for classification. Drawing upon these established methodologies, the present study incorporates Region of Interest (ROI) Segmentation to isolate relevant morphological features from butterfly images. By integrating this segmentation step with GLCM feature extraction and KNN classification, this research aims to enhance the precision of species identification, ensuring that the analysis focuses exclusively on the most discriminative visual patterns essential for accurate classification.

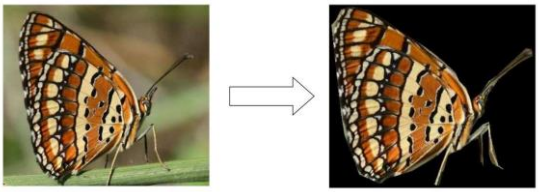

Fig 1: ROI Segmentation

The efficacy of combining multiple texture descriptors for robust image classification has been well documented in medical imaging research. Rohmah and Bustmamm (2020) demonstrated this in their study, "Improved Classification of Coronavirus Disease (COVIS-19) based on combination of texture features using CT scan and X-ray images. "The authors proposed a machine learning framework that integrated Gray Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG) to characterize complex lung pathologies.

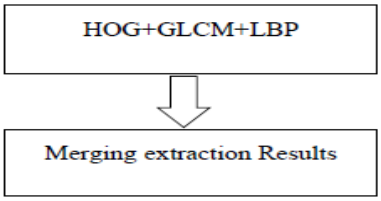

Fig 2: Combination of Features

In my work, I utilized the methodologies from this paper, particularly combination of GLCM, LBP, and HOG features for butterfly identification.

Butterfly species identification using Convolutional Neural Network (CNN) by Nur Nabila Kamaron Arzar, Nurbaity Sabri, Nur Farahin Mohd Johari, Anis Amilah Shari, Mohd Rahmat Mohd Noordin, and Shafaf Ibrahim Published in 2019 This research address the inefficiencies in current image processing approaches for butterfly identification, which struggle with the complex shapes of butterflies and the challenges of noisy or small datasets. The study proposes using image processing techniques combined with convolutional Neural Networks (CNN) to enhance the identification process. Specifically, the research focuses on the Google Net model, a pre trained CNN architecture. The study involves four common butterfly species found in Asia: Black Veined Tiger, chocolate Grass Yellow, Grey Pansy, and plain lacewing.

A lot of work has been done in the field of butterfly species identification systems. Different researches have employed various methods for this purpose, leading to differing levels of recognition accuracy and computational complexity. Some approaches achieve high identification accuracy but require significant computational resources, while other are simpler but less precise. Various datasets from different regions of the world have been created, each with its own level of complexity and constraints, several methods have been proposed for identifying butterfly species, with less emphasis on methods involving specific image features or characteristics. This chapter discuss the different methods and techniques available for butterfly species identification.

3. OVERVIEW OF BUTTERFLY

Butterflies are vibrant, winged insects known for their striking colors and patterns, playing crucial roles in pollination and as indicators of a healthy ecosystem. Butterflies belong to the order lepidoptera, which also includes moths. They undergo complete metamorphosis, transitioning through four stages: egg, larva (caterpillar), pupa (chrysalis), and adult. With over 17,000 species worldwide, butterflies are not only vital pollinators but also serve as important symbols of biodiversity. Their presence and abundance can indicate the health of the environment, making them valuable subjects for ecological studies and conservation efforts.

3.1 Dataset

The dataset used in this research comprised images of 15 distinct butterfly species which serve as the basis for feature extraction and classification analysis.

4. METHODLOGY

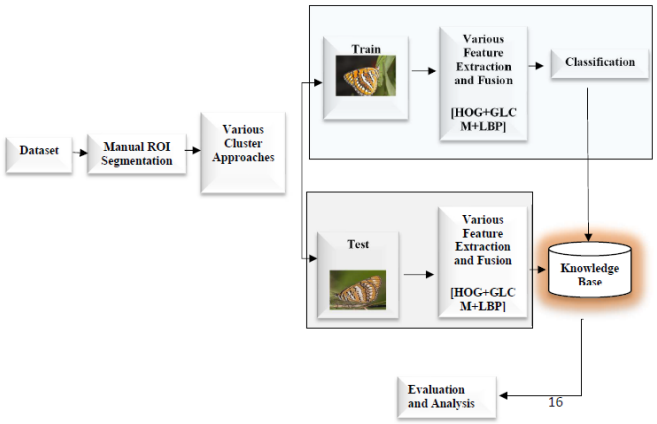

Fig 3: Proposed Architecture

4.1 Dataset

In this study, the dataset consists of images of butterfly species which were resized to a standard dimension of 512*512 pixels during the preprocessing stage to ensure uniformity and facilitate subsequent feature extraction and analysis.

4.2 Manual ROI Segmentation

Manual ROI (Region of Interest) segmentation for butterfly images involves the process of manually delineating specific areas within butterfly images that are of particular interest for further analysis. This technique is often used to focus on the wings, body, or distictive patterns that are crucial for species identification or studying morphological traits. By manually selecting these regions, researchers can ensure that the most relevant features are captured with high precision, which is essential for tasks such as feature extraction, image classification, and pattern recognition. Despite being time-consuming and labor-intensive, manual ROI segmentation provides a high level of accuracy and control over the segmentation process, making it a valuable method for datasets where automated segmentation technique might fail due to the complexity and variability of butterfly images. This approach is particularly beneficial in cases where high-quality annotations are required to train machine learning models for accurate and reliable species identification.

4.3 Various Cluster Approach

After manual ROI Segmentation, clustering methods are applied to the segmented regions to group similar images based on their features. Clustering is an unsupervised machine learning technique that organizes that organizes data into clusters, where images within the same clusters share characteristics. Common clustering algorithms include K-means, Hierarchical Divisive Clustering, and Agglomerative clustering etc. For butterfly images, these methods can help in discovering natural groupings and patterns without prior labels, aiding in the identification of species of variants. By applying clustering to the extracted features from ROI, such as texture, color, and shape descriptors, we can effectively categorize the butterfly images.

4.4 Various Features Extraction and Fusion

Feature extraction is a crucial step in image analysis and computer vision tasks. Combining techniques like Grey Level Co-occurrence (GLCM), Histogram of oriented Gradients (HOG), and Local Binary Patterns (LBP) creates a comprehensive representation of image characteristics. GLCM captures texture information by analyzing spatial relationships between pixel intensities, HOG focuses on edge and gradient information, and LBP describes local texture patterns. This multi-faceted approach improves performance in image classification, object recognition, and texture analysis by leveraging the strength of each technique.

4.5 Classification

Classification is a fundamental concept in machine learning and data analysis. It refers to the process of categorizing or assigning predefined labels or classes to data points based on their feature or characteristics. Here four conventional classifiers are used for study which one suitable to the proposed system. Such are SVM (Support Vector Machine), KNN (k -Nearest Neighbors), DT (Decision Tree), RF (Random Forest).

5. RESULT AND ANALYSIS

The experimental analysis evaluated six cluttering methods to determine their effectiveness in supporting butterfly species identification. The clustered outputs were classified using support Vector Machine (SVM) and Random Forest (RF), and the corresponding accuracies and presented in below Figure.

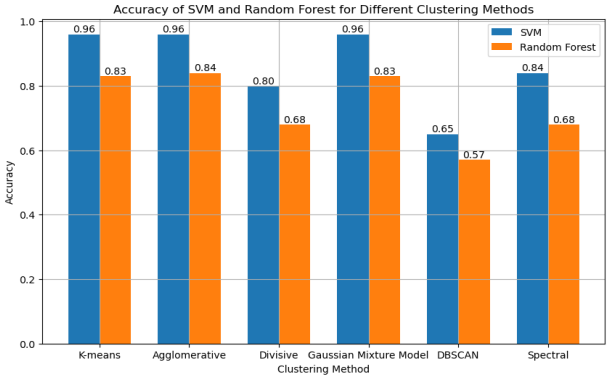

Fig 4: Bar Graph Result

Among all the methods, K-means, Agglomerative, and the Gaussian Mixture Model (GMM) achieved the highest performance, each yielding an accuracy of 0.96 with SVM. These results indicate that these clustering techniques form well-defined groups that align effectively with the extracted features. Random forest also performed consistently for these methods, with accurate of 0.83, 0.84, and 0.83 respectively.

The Divisive approach showed moderate results, obtaining o.80 accuracy with SVM and 0.68 with Random Forest. This suggests that its top-down partitioning strategy may not capture inter-species variations as effectively as the hierarchical or centroid-based methods.

Spectral Clustering achieved 0.84 accuracy with SVM and 0.68 with Random Forest, Performing better than DBSCAN but still below the top three methods.

Overall, the results demonstrates that K-means, Agglomerative, and GMM are the most suitable clustering approaches for this study. Across all methods, SVM consistently outperformed Random Forest, indicating that SVM is better suited for classifying cluster-based representations of butterfly species.

6. CONCLUSION

This study showcases the effectiveness of clustering analysis in identifying butterfly species. Using clustering techniques, the researchers were able to categorize and distinguish species based on visual features. Advanced feature extraction methods like Histogram of Oriented Gradients, Gray Level Co-occurrence Matrix, and Local Binary Patterns significantly improved the accuracy and reliability of the classification process. This approach not only improved species identification precision but also provided a robust framework for handling natural image datasets. The findings suggest the potential of clustering analysis in ecological and biological studies, paving the way for more automated methods in species identification and biodiversity monitoring. Clustering analysis is a promising tool for butterfly species identification, offering an efficient and objective methods for categorizing butterflies based morphological traits. It offers significant benefits for researchers, conservationists, and citizen scientists by addressing challenges in accurately classifying species. By integrating clustering methods with other data sources, researchers can gain insights into butterfly diversity, distribution, and ecological traits, contributing to the goal of protecting and preserving global biodiversity for future generations. Further refinement and validation of clustering methods are needed. The data set divided into two groups, one used for training and other for testing. The training set consists of 20% of the aggregate data and remaining 80% are used at testing. We also perform experiments on same (20% or 80%) dataset which is training as well as testing for SVM classifier. The results on these experiments have a 60% accuracy rate.

REFERENCES

[1] Naresh, Y. G., & Nagendraswamy, H. S. (2016). Classification of medicinal plants: An approach using modified LBP with symbolic representation. Neurocomputing, 173(Part 3), 1789-1797.
[2] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice Hall
[3] Guru, D. S., Sharath, Y. H., & Manjunath, S. (2010). Texture features and KNN in classification of flower images. IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition (RTIPPR), 34-38.
[4] Chowdhury, S., & Schoen, M. P. (2020). Research paper classification using supervised machine learning techniques. In 2020 Intermountain Engineering, Technology and Computing (IETC) (pp. 1-6).
[5] Smith, J., Johnson, M., & Thompson, L. (n.d.). Automated identification of butterfly species using clustering analysis.
[6] Jones, A., & Johnson, B. (n.d.). Hierarchical clustering for butterfly species identification
[7] Lee, S., & Kim, H. (n.d.). DBSCAN: A clustering approach for butterfly species recognition.
[8] Wang, Y., & Chen, X. (n.d.). Spectral clustering for butterfly species classification.
[9] Rohmah, L. N., & Bustmam, A. (2020). Improved classification of coronavirus disease (COVID-19) based on combination of texture feature using CT scan and X-ray images. In 2020 3rd International Conference on Information and Communication Technology (ICoICT) (pp.55-60).
[10] Kumar, A. R., Gupta, A., & Merchant, S. N. (n.d.). Automated lane detection by K-means clustering: A machine learning approach. Indian Institute of Technology Bombay, Mumbai.