A Comprehensive Survey On Vision-Based Insect Species Identification and Classification

Preprint · April 2020

DOI: 10.13140/RG.2.2.10083.50720

3 authors:

Chinmaya Hs
Cambridge Institute of Technology

Manoj Balaji Jagadeeshan
Rakuten

Ganesh N Sharma
University Visvesvaraya College of Engineering

A Comprehensive Survey On Vision-Based Insect Species Identification and Classification

H S, Chinmaya * J, Manoj Balaji * Sharma, Ganesh N *

Abstract

The research on automation of identification and classification, supported by studies of insect ecology and related domains was marked by the introduction of DAISY - digital automated identification system. The framework embarked multiple approaches to solve the problem, varying from PCA to NNC (nearest neighbour clustering) algorithms. Efforts involving statistical methods are also evident in the domain. Evolution of artificial neural networks, associative memory networks also marked major changes in the direction of effort towards the problem statement. Other machine-learning based classification techniques such as SVM, PCALC have also been used as solution methods. Clustering techniques combined with image features were revived with Correspondence filters. Various transforms such as Fourier, SIFT and wavelet as features also based some studies. With the dawn of deep learning, more advanced techniques such as pose estimation have also become a base to solution framing. Image-based techniques, involving pattern extraction have prevailed with exceptional results. Approaches towards automation of the process to the solution have been decorated ever since.

1. Introduction

"Insects" is the colloquial term for Hexapods; biologically classified into Orthopods. These creatures appear in various sizes, shapes and colours; often intriguing physical appearances. In the biological category, they are members of kingdom Animalia, phylum Arthro- poda, and class Insecta, They contribute 90 per cent of the animal life forms in the world and is estimated that there are around 5.5 million different insect species on earth. Their impact, even though most of them are relatively smaller in size to the creatures around them, on the environment is as widespread as their presence too. With this large distribution, they play a major role in human life and the ecosystem. 80% of the countries in the world consume insects as food. They play a major role in pollination and hence contributing to plant reproduction.

Insects in ecology: Insects are morphologically identifiable with soft body texture and are found to flaunt various defence mechanisms such as camouflage, secretions to counter the predetors, often in vain due to the difference in size compared to other organisms. Insects being the primal in the list of pollinators, most of the flowering plants depend on them for reproduction, while some variants for source of nutrition as in the case of insectivourous plants.

Interaction with humans: A prominent group of insects, which are called pests, can be characterized by having harmful or destructive effects on their environment. They contribute to unintended outcomes through various means such as the spread of diseases, crop destruction. Some of the most commonly found illustrations include a wide variety of mosquito, common fly and caterpillar species. On the other hand, some species of insects, such as dragonflies and mantises act as predators to many others of their own, thus lending a helping hand to agriculturists. A wide varietal range of species of insects is consumed as food all over the globe, illustrated by gross hoppers and ants. Multiple technological inspirations have been drawn by insects and their behavioural mechanisms, as in, compound eyes, and hi-resolution cameras.

Owing to their populous nature, primal tasks in studies around insects are identification and classification. While identification involves diving into the details and features that uniquely define them, classification involves gathering commonality across the groups formed of similar ones. This literature exploration involves the investigation of 22 manuscripts published from 1996 to 2019, chronologically. The research in this field gained importance during the grate crisis but eventually, the hype went down with time. Even though this survey hints of summarisations of topics across various related domains, our major focus is on the image-based identification and classification efforts which numerous technical aspects and procedures. A detailed discussion to provide worthy insights and in-depth understanding, about the limitations and improvements is thus unfolded.

2. Insect identification and classification

2.1. Classical Methods

Insects are one of the prominent members of Animalia kingdom. They belong to Phylum Arthropoda and Class Insecta following the Linnaean taxonomy. The process of identification revolves around the parameters used in the categorisation. Any species that belong to these categories are later classified into a different order, family, genus, and species. Often, the family name is enough for identification but manually, an expertise in understanding their juvenile stage, adult stage, and immature stage becomes a must. A deeper understanding of insect characteristics such as 3 pairs of legs, three body regions namely, head, body and thorax, zero, one and two pairs of wings and a pair of antenna, as highlighted in [9], plays a key role in the process too. To conclude, an in-depth understanding and knowledge regarding the morphology and life cyle of instects are requried to orchestrate a simplest effort of identification or classification.

There are 26 orders belonging to the Class Insecta of which some of them require profound efforts to be identified [9], due to their size and habitat. Experienced taxonomists use features such as habitat, food habit, shape, body armor structure, wing patterns if any, size of the antenna, larval stage behavior, reproduction, and other visual aspects to classify them and is undoubtedly a challenging task.

Techniques apart from, classical observation, used to identify the insects are mentioned in [19] and [7]. The authors highlight DNA based techniques and internal morphology. While, the DNA based techniques implement fingerprinting technology to identify and classify the insect species procedures pertaining to internal morphology involves, dissecting the insect and later study the internal parts to identify and understand the food habits and other related information.

2.2. Visual Identification

Identification of different organisms involves the first step being towards understanding the behaviour and other details about it. A prominent of ways is by using images. One of the noteworthy identification systems, spanning across biological species is DAISY[8] - digital automated identification system. DAISY in its original form was implemented using a PCA based approach. For insects it mainly focused on wing pattern as an important feature. One of the intense shortcomings of the PCA technique in the linearity of the morphospace containing the pattern classes. Combined with other computationally intensive operations required such as, the need to recompute the transforms from a pattern space of higher-dimensions to a low-dimensionality component space each time material is added or removed from the system, PCA for DAISY resulted inadequate for the task. As an improvement, the introduction of nearest neighbour classifiers for DAISY were welcomed. The robustness and simplistic nature of the technique assisted in wide-spread applications ranging from insect identification to plant leaf and human face identification. Although, DAISY as a pattern recognition engine is limited by background noises, pose scale, cluster geometry in pattern space.

More focused efforts on particular species identification also bloomed. P.J.D. Weeks et al outlined in their work[12], a semi-automated technique, to identify the individuals belonging to 5 closely related species based on the difference in wing patterns. The system is able to extract the information and consider the variation in individuals of the same species, morphological features to identify individuals, the key differences between patterns of different species. The model works in different modules among these Gaussian smoothed images give higher accuracy. Gaussian smoothing highlights the features better than others. Hence making it successful than previous work. The same authors proposed a newer technique involving associative memory networks to identify wasp species based on wing pattern and venation in [13]. Using Kendal's T, a statistical method, the authors classified the insect and found out that accuracy is affected by number of images given for training.

Contemporarily, improvementative reseach on DAISY still yeilded a more evolved technique[14], in which digital automated identification system, first digitizes the image and converts it to a vector, which is then subjected to principal component analysis. The PCA gives a set of eigenfunctions that describe the variations accurately. The variations may be rotation and scaling. Then, Kendal's T statistics, is used to compare the original and reconstructed images. P.J. D.Weeks et al. outlined the benchmarks to explore the limitations of the prototype system. 60% of the dataset was classified as the training set and other 30% for the testing set. The classifier with the highest correlation will be considered correct. The highest accuracy was achieved using a higher count of the dataset. However, other parameters like, SD of correct identification, impart accuracy in the lesser dataset.

To overcome the drawbacks of previous approaches, Kevin J. Gaston et al discussed in [5], the technical obstacles and issues that hinders the process. There are many algorithms and techniques discussed, which helps in identifying the patterns in the training data. The advancement in artificial neural networks brought ease in identifying more complex, hidden patterns. Advancement in more sophisticated and faster algorithms increase the potential in extracting more complex patterns. It reduces the burden of routine identification of diversity.

With the previous works, it is clear that it is important to extract features that determine the individual that belongs to a species. Extracting features from the images is a difficult task. This plays a major role in the case of identifying different species of insects. In the paper [4], Yuefang Gao discussed a hybrid method that is insensitive to blurring, scaling, rotation and other affine transformations. Insect wings are unique to species. It is proposed that the network will identify the wing patterns. This technique uses a nearest neighbor correlation and plastic self-organizing maps for identification. The problem arises when the quality of the image is less. To overcome this problem the paper is provided with Hu's invariant and blur-affine invariant to identify insect species. This hybrid method implemented along with SVMs to classify given greyscale wing images into one of the 15 species. Butterflies, some of the beautiful and amazing insects, which are 18,500 species[18], and distributed all over the globe, except Antarctica.

The authors in [16], pursues the automated identification and classification exercises of orchard moths based on local features rather than global features, which are, when used for inter-species classification, disadvantageous, attributing to causants such as total or partial rotational invariance, dependence on robust segmentation techniques. Also, other features such as illumination, edge and scale, considered as global features also extrapolate the woes of ROI extraction and usage of important features in classification. They use scale invariant feature transform (SIFT) descriptors for local regional features, image representation using a visual technique similar to bag of words where codewords represents image objects or regions, normalised as feature vector inputs to classifiers. The classification is evaluated using various methods such as KNNC, PCALC, SVM. Results have been obtained using the 10-fold cross-validation test, to show SVM and PCALC performing the best.

More recent studies involving content-based image retrieval was proposed in [10], capable of mass processing and able to operate with large amounts of data proposed by Jiangning Wang et al. Recent improvements and interests in image-based techniques for identifying insect species has given fuel to increase the advancement of the field. Butterflies have a large degree of diversity in color, wing pattern, shape, size lifespan and other features in morphological level. The images were cleared from the background and provided with a new background that is uncommon to the butterfly species such as yellow. Then using template matching and ANNs the best suitable template for given images is identified. The model is flexible for adding any new species without rebuilding the whole. New templates can be added to the database anytime. The similarity between template and a new image is calculated and the image is categorized to the family with the highest FS score. Precession recall values are used to denote the efficiency of the CBIR systems, but for identification task recall value alone played a major role as every image will have a template with some FS value.

Dedicated to orchards, the tree pests which infect and destroy trees from inside by consuming tree sap, Chenglu Wen et al proposed an image-based identification technique in [15]. The automated method of identifying the insects overcomes these problems by learning these local and global features. The model considers 54 global features and a hierarchical model considers the local features. The model works in a manner in which, if the probability threshold value is higher after considering global features and only if the probability values are lesser than the threshold then it will consider the local features for classification. The automated insect identifier is made with three feature models, local feature models, global feature model, and a hierarchical model for combining both the model. The model acts flexibly with lower quality images by considering all the features of both the model. The model acts flexibly with lower quality images by considering all the features. They overcome the drawback of manual identification. For a human to identify insects, the global feature like color, shape, size plays a major role but for an identification system, local features like the variation in brightness, rotation, blurring will affect it in an unpleasant way. In the era of decreasing the number of taxonomists, it is important to introduce new automated methods to identify species.

Similarly, focused on fruit flies, F.A. Faria proposed an automated identification technique in [2], that uses wings, aculei as key features for identification. They used three different approaches for the purpose, SVM for wings, SVM for aculei and a combined model which gave them a proper understanding of the classification of these fruit flies. It is concluded that wings, aculei and combined dataset of these two can be used for classification. With 5 fold cross-validation technique, the best outputs are taken for consideration. Classifier made with combining both the datasets, considerably reduced the error and gave an optimal result.

With the popularisation of neural network based approaches, Jiangning Wang et al. discussed an ANN-based method to classify insects at the order level in [11]. ANNs are used to capture the features provided during training. The backpropagation neural nets and sigmoid activation function during training to tune the network towards optimum. They have proposed a technique, which takes grayscale images as input and the network will learn the parameters and variations among different datasets. Similar to ANNs, SVMs, risk-minimizing classification technique is used to extract features from dataset during the training process. C-SVM, a popular type of SVM showed the results and classified the images. Using 10-fold cross-validation to all the images it is shown that these are better than ANNs. Here, both the systems considered major morphological features like shape, size, patterns, and structure of the body. It did not capture features like colors and color patterns.

Insect identification and classification as a problem statement does not only address the real world challenges of protecting and enhancing harvests, but also protecting them. Insects being both friends and foes to humans, are worthwhile to be classified, as the authors, in [1], use K-means clustering algorithm as a classifier and correspondence filters. Correspondence filters being constructed from different angular rotations of the k-means output clusters aid in the recognition, owing to the feature of distortion invariance, they exhibit. The k-means clustering algorithm is backed by transformation techniques on the colour space. The results are presented to be successful in exact detection and identification among plant pests of various shapes, sizes, positions, and orientations using computationally efficient algorithms.More recent efforts towards the problem shift towards using neural network based approaches. The authors, in [6], exhibit the usage of classical digital image processing techniques such as edge detection, rotation of ROI and removal of noise with ANNs as classifiers for the classification of copepods. ANNs flaunting a great generalization on the input feature vectors, are used to classify and generalise on the selected feature vectors after the pre-processing. The authors have published a result of 97.3% accuracy on 240 samples after 137 epochs. Evaluation of the network output layer is done using MSE and confusion matrices.

Chenglu Wen et al proposed a pose estimation based technique using pyramidal stacked de-noising auto-encoder (IpSDAE) deep learning architecture in [17], to address an issue related to the dataset as each of the pictures had different pose of moths trapped and different backgrounds causing poor features extraction and thus creating hurdles for automatic identification. The authors have published the results achieving 96.9% accuracy using 10-Fold cross-validation on 762 moth image samples, thus showing the effectiveness of using pose-estimation. The authors in [3] have proposed an automated species identification and retrieval system (SPIR) which uses pattern cognition technique on moth species images. SPIR consists of probabilistic-models that predict Semantically Related Visual (SRV) features from moth images present in the training sample, where wings of moths are segmented to extract features locally. This was used on unlabeled data samples for test-set to identify SRV of patterns in moth's wings followed by image similarity technique to identify moth species. The proposed technique worked effectively on available entomology datasets, thus showing the effectiveness of the utilization of SRV for moth species detection.

He-Ping Yang et al in [20] have proposed a system called DAIIS which uses the wing outlines as a feature where the insect's wing outline is transformed using Fourier transformation after which the output for the Fourier transform is fed to a Support Vector Machine(SVM) classification model. The authors have shown that this proposed technique, on repeated sampling of the dataset, resulted in an average accuracy between 90-98% for each species class identification individually. They also showed that the resultant accuracy increased to 99% by segmenting the dataset using the features of insects compound eyes, thus showing the effectiveness of the proposed system.

Recently, Ankang Xue et al proposed a Gray-Level Co-occurrence Matrix Features for identifying butterfly species in [**?**]The paper specifies technique that uses, a weight-based K - Nearest Neighbour method(popularly referred as KNN), on gray-level co-occurrence matrix for classification. The pro- cess involves the conversion of an image into a gray-scale, dividing the greyed into a number of blocks and this will be given as input for the classifier. The weight-based KNN was able to classify them efficiently with less number of classes but the efficiency dropped with increase in number of classes.

3. Conclusion

In this survey, our main motif was to understand how computer vision is used in insect classification. We investigated image-based insect classification with the utilization of pattern recognition, machine learning. Our findings suggest that major visual features such as wing outlines, wing patterns, the color of the wings, pose of insects are some of the very important factors in the identification of insects. We also found that most of the efforts are on a minor subset of species in the class Insecta such as moths, mosquitoes, Butterflies whereas a small subset of works focus on the identification of less popular and important species, for example, Dragonflies, Bees, etc. We conclude that there's a major requirement for automated learning algorithms which can drastically reduce human efforts for feature construction and there's also a major requirement of research work in the identification of aforementioned species of class Insecta.

4. Future Work

After deep investigation, the findings of the survey show that there's a requirement of an end-to-end automated insect species identification algorithm, thus paving a path to the utilization of complex deep neural networks for extraction for visual features from images, which will be used to classify insects species. The findings also suggest requirement of research work for identifying other subsets of insect species, thus work needs to be done on aforementioned subset of insect species.

Acknowledgement

All Authors have equally contributed in this survey and writing of this manuscipt. We would also like to thank all the author's of previous work, for their contribution on the topic of image-based insect detection

References

[1] Fina Otosi Faithpraise, Philip Birch, Rupert Young, Ja Obu, Bassey Faithpraise, and Chris Chatwin. Automatic plant pest detection and recognition using k-means clustering algorithm and correspondence filters. 2013. 4

[2] Fabio Faria, Paula Perre, Roberto Zucchi, Leonardo Jorge, Thomas Lewinsohn, Anderson Rocha, and Ricardo Torres. Automatic identification of fruit flies (diptera: Tephritidae). *Journal of Visual Communication and Image Representation*, 25:1516–1527, 10 2014. 3

[3] Linan Feng, Bir Bhanu, and John Heraty. A software system for automated identification and retrieval of moth images based on wing attributes. *Pattern Recognition*, 51, 09 2015. 4

[4] Y. Gao, H. Song, X. Tian, and Y. Chen. Identification algorithm of winged insects based on hybrid moment invariants. In *2007 1st International Conference on Bioinformatics and Biomedical Engineering*, pages 531–534, 2007. 3

[5] H. C. J. Godfray, S. Knapp, Kevin J. Gaston, and Mark A. O'Neill. Automated species identification: why not? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444):655–667, 2004. 2

[6] Lee Kien Leow, Li-Lee Chew, Ving Chong, and Sarinder Kaur Dhillon. Automated identification of copepods using digital image processing and artificial neural network. *BMC Bioinformatics*, 16, 12 2015. 4

[7] Kelly B. Miller, Yves Alarie, G. William Wolfe, and Michael F. Whiting. Association of insect life stages using dna sequences: the larvae of philodytes umbrinus (motschulsky) (coleoptera: Dytiscidae). *Systematic Entomology*, 30(4):499–509, 2005. 2

[8] M. A. O'Neill. Daisy: a practical tool for semi-automated species identification. In N. MacLeod, editor, *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*, chapter 9, pages 131–152. CRC Press, Boca Raton, FL, 2008. 2

[9] Scott Schell and Dr. Alex Latchininsky. *Insect Identification*. CES – Entomology - Renewable Resources Univ. of Wyoming, 2007. 2

[10] Jiangning Wang, Liqiang Ji, Ai-Ping Liang, and Decheng Yuan. The identification of butterfly families using content-based image retrieval. *Biosystems Engineering - BIOSYST ENG*, 111, 01 2012. 3

[11] Jiangning Wang, Congtian Lin, Liqiang Ji, and Aiping Liang. A new automatic identification system of insect images at the order level. *Knowledge-Based Systems*, 33:102 – 110, 2012. 3

[12] P.J.D. Weeks, I.D. Gauld, K.J. Gaston, and M.A. O'Neill. Automating the identification of insects: a new solution to an old problem. *Bulletin of Entomological Research*, 87(2):203–211, 1997. 2

[13] P.J.D Weeks, M.A O'Neill, K.J Gaston, and I.D Gauld. Species–identification of wasps using principal component associative memories. *Image and Vision Computing*, 17(12):861 – 866, 1999. 2

[14] P. J. D. Weeks, M. A. O'Neill, K. J. Gaston, and I. D. Gauld. Automating insect identification: exploring the limitations of a prototype system. *Journal of Applied Entomology*, 123(1):1–8, 1999. 2

[15] Chenglu Wen and Daniel Guyer. Image-based orchard insect automated identification and classification method.*Computers and Electronics in Agriculture*, 89:110 – 115, 2012.3

[16] Chenglu Wen, Daniel Guyer, and Wei Li. Local feature-based identification and classification for orchard insects. *Biosystems Engineering*, 104:299–307, 11 2009.3

[17] Chenglu Wen, Daoxi Wu, Huosheng Hu, and Wei Pan.Pose estimation-dependent identification method for field moth images using deep learning architecture. 2015. 4

[18] Wikipedia contributors. Butterfly — Wikipedia, the free encyclopedia, 2020. [Online; accessed 01-April-2020]. 3

[19] Kipling W. Will and Daniel Rubinoff. Myth of the molecule: Dna barcodes for species cannot replace morphology for identification and classification. 2004. 2

[20] He-Ping Yang, Chun-Sen Ma, Hui Wen, Qing-Bin Zhan, and Xin-Li Wang. A tool for developing an automatic insect identification system based on wing outlines. *Scientific Reports*, 5(1), Jul 2015. 4