A fully automatic classification of bee species from wing images

Allan Rodrigues Rebelo, Joao M. G. Fagundes, Luciano A. Digiampietri, Tiago M. Francoy, Helton Hideraldo Bíscaro

Apidologie (2021) 52:1060–1074

DOI: 10.1007/s13592-021-00887-1

Original article

A fully automatic classification of bee species from wing images

Allan Rodrigues Rebelo, Joao M. G. Fagundes, Luciano A. Digiampietri, Tiago M. Francoy, Helton Hideraldo Bíscaro

School of Arts Science and Humanities, University of São Paulo, São Paulo, Brazil

Received 18 February 2021 – Revised 14 May 2021 – Accepted 2 August 2021

Abstract – Since bees are the main pollinators of natural and agricultural ecosystems, they play a fundamental role in the preservation of the environment and food production. However, species identification is one of the bottlenecks for bee conservation due to its complex taxonomy, a large number of existing species, and the scarcity of professional taxonomists. In this sense, the automatic identification of such species can present a good alternative for non-taxonomist scientists and to the general public. In this work, we propose a fully automatic bee identification system based on the patterns of forewing venation. Our system was based on a combination of image segmentation techniques followed by a simple classification method. We achieved an accuracy of 99% in the genus and 96% in the species in a dataset composed of 48 species and 23 genera. This result represents an advance compared to previous works in the literature and there are plans to make the system online available for the general public.

bees identification / computer vision / image segmentation / classification / taxonomy / conservation

1. INTRODUCTION

Ecosystem services, which can be defined as the benefits people obtain from the normal functioning of ecosystems, are directly linked to the well-being of human populations around the globe (World Commission on Environment and Development 1987). Among these services, pollination is essential for food security and is affected by the enormous loss of biodiversity the world has been experiencing since the Industrial Revolution (Rockstrom et al. 2009). The annual market value of the pollination services was estimated at US$ 235 billion—577 billion worldwide (on Biodiversity, I.S.P.P., Ecosystem Services, I 2016). As primary pollinators of crops and natural ecosystems, bees are responsible for the reproduction of the vast majority of the plants with flowers (on Biodiversity, I.S.P.P., Ecosystem Services, I 2016), which makes bees essential organisms in both economic and ecological terms, besides being vital for conservation purposes. The decrease in the number of pollinators can effectively impact the maintenance of the diversity of vegetation species, as well as the stability of some ecosystems (Potts et al. 2010). It is estimated that there are more than 20,000 bee species in the world (Michener 2007; Ascher and Pickering 2020) and the identification of these species can be a challenge due to the great similarity between species, which requires a great deal of experience of taxonomists. In addition, one must take into account the human error existing in such identifications (de Carvalho et al. 2007). This area is growing in recent years and is expected to increase due to two factors: first due to

Corresponding author: H. H. Bíscaro, heltonhb@usp.br
Handling Editor: Peter Rosenkranz
Contributions: HHB, TMF and LAD conceived this research and designed experiments; ARR and JMGF performed experiments and analysis; TMF provided the data set; All authors read and approved the final manuscript.

the increasing demand for the identification and classification of insects, and second due to the decrease in the number of taxonomic specialists (Gaston and O'Neill 2004; Houle et al. 2003).

Based on these difficulties, several methodologies are being proposed to minimize this problem, such as DNA barcode (Hebert et al. 2003) and automatic recognition of species based on wing morphology (Steinhage et al. 2001; Francoy and Imperatriz-Fonseca2010). These methodologies are beneficial since they allow access to species identification to a broader audience, leaving taxonomists to spend time in more urgent tasks, like species description. It is worth mentioning that, for some groups, automatic identification can be a challenge, especially stingless bees (Francoy and Imperatriz-Fonseca2010).

In the area of computer vision, there is the application of techniques for the identification of insects (Zhong et al. 2018; Lu et al. 2010). The use of image processing and pattern recognition algorithms for the automatic classification of insect species has changed the traditional manual descriptive model of morphological characteristics provided by taxonomic studies for their identification.

According to (Martineau et al. 2017), with the increase in the capacity of mobile devices, an increase in the open field capture system is expected, which would provide the possibility of non-specialists using identification/classification systems, increasing the number of users.

Using computational models with automated artificial intelligence techniques, the identification of insect species can be done by more laypersons and in less time than traditional models. And with the advantage that the classification, in addition to having a higher percentage of accuracy than if done manually, can be easily measured, tested, and replicated (Martineau et al. 2017).

Automatic insect classification is an application of computer vision, and when compared to traditional manual classification, it is relatively inexpensive and time and money efficient (Houle et al. 2003; Gaston and O'Neill 2004).

Several systems that classify bee species can be found in the literature (Francoy et al. 2008; Rojas et al. 2016); the most famous probably is the Automatic Bee Identification System (ABIS)(Steinhage et al. 2001), but it was discontinued in 2005 (Rojas et al. 2016) and it had downsides that most studies have today: Only a few species are generally analyzed and the systems are not fully automatic, requiring user input at some point in the middle of the process (Rebelo et al. 2020).

To classify bee species, the junctions of the wing venations are used by several studies (Steinhage et al. 2001; Santana et al. 2014b; Rojas et al. 2016; Silva et al. 2015; Strauss and Houck 1994) because they are excellent descriptors for this purpose (Santana et al. 2014b; Francoy et al. 2008; Silva et al. 2015). Thus, the challenge of building a system that automates the classification process can be summed up to the problem of identifying these junctions and their use in classification algorithms.

Image-based recognition depends on well-defined image acquisition and processing processes. The acquisition step is usually done by selecting the insects that are objects of research and photographing them individually, preferably under controlled conditions to avoid noise and different background and lighting conditions. The processing stage, on the other hand, consists of several processes depending on the objectives. In general, color images are converted to gray-scale and then binarized. The region of interest (ROI) is separated from the background. Then, some feature extraction technique must be applied and these features will be sent to a classification algorithm. Of course, different techniques can produce different results.

The process of automated extraction of bee wing characteristics is a computational challenge that involves first treating the image to remove or reduce possible noise, segmenting the wing so that only the area of interest is used, and then extracting numerical data (Santana et al. 2014a).

In addition to the fact that there are different species with very similar characteristics, the images are subject to variations in position, scale, resolution, lighting, noise level, among others. These facts make the problem quite challenging from a computational point of view.

Our research hypothesis is that the accuracy of methods for automatic classification of bee species based on wing morphology can be improved

with a combination of image segmentation techniques and artificial intelligence.

2. BASIC CONCEPTS

This section presents the concepts of graphic processing and classification that were used in the elaboration of our algorithm and also necessary to understand section 3 of related works. We highlight the image filters described in this section, the modified Hausdorff distance, and the classification based on a decision tree.

2.1. Computer vision and image processing

Computer vision can be described as a sub-area of computer graphics where it seeks to extract relevant information from images. In other words, the data entry is an image and the output is information such as "there is or is not a certain object, feature or activity in the input image."

Image processing is a subarea of computer graphics that deals with techniques and algorithms for treatment and image manipulation.

Several filters were used in order to prepare the images for the classification algorithms. In the following, we will briefly describe each of them:

Thresholding is an image processing method to convert an image into a binary image. This is a basic thresholding algorithm. Replace each pixel to white value if the original pixel intensity is higher than a fixed chosen constant, or black otherwise.

The *Difference of Gaussian* is an edge detection method; the algorithm performs a Gaussian filter and creates a blurred version of an image, then the algorithm performs a second Gaussian filter and creates a second blurred version, but less blurred. Finally, the output is the difference between the two versions.

Expansion is one of the basic operations in morphological image processing. While most commonly applied to a binary image, the expansion operation usually uses a structuring element to investigate and expand the edges of the image (Silva 2015). In our proposed approach, expansion is mainly used to reconnect members of the wings that were improperly disconnected in the thresholding, filtering, or noise removal operations.

Erosion is the opposite of expansion. This operation removes details at the edges of objects. It is commonly used to reduce the size of an image.

Skeletonization is a method for drawing a one-pixel-wide skeleton of a binary image, maintaining the shape and structure of the complete image. The skeletonization algorithm of (Zhang and Suen 1984) is probably the most used in the category. It works in two steps, which means that for each iteration, it performs two sets of operations to remove pixels from the image. These operations are designed so that the first set is removed from the southeast corner (bottom right) of the image and the second set is removed from the northwest corner (top left).

Hit-and-Miss is a binary morphological operation that can be used to look for specific pixel patterns. As with other binary morphological operators, it takes a binary image and the specific pattern as input and then produces another binary image, where one value corresponds to locations where the chosen pattern is placed. In our project, it is used to detect features of the image.

Noise reduction and smoothing filters. Smoothing filters are used to smooth an image; it reduces the amount of intensity variation between one pixel and its neighbors. Since noise is an irregular variation of brightness or color information, it might be used to reduce those variations, therefore reducing noise. In this work, we use several filters, such as Gaussian, Median, and Bilateral, known as edge preservation filters, in order to remove as much noise from the input images as possible without losing important information (Kaehler and Bradski 2016). Additionally, it was also used as noise reduction techniques based on object size and closeness to the main object.

2.2. Image classification

Modified Hausdorff distance (Marinov 2012) is experimented by (Dubuisson and Jain 1994), and determined to be better than the other examined distance measures for object matching. It was used to measure the accuracy of the segmentation of our algorithm compared to a ground truth segmentation.

Given two finite sets $A = \{a_1, \ldots, a_p\}$ and $B = \{b_1, \ldots, b_q\}$ in a metric space, the Hausdorff distance (H) (Gao et al. 2014) between the both sets is defined as:

$$H(A,B) = \max\left(h(A,B), h(B,A)\right)$$

$$\text{where } h(A,B) = \max_{a \in A}\left(\min_{b \in B} \|a-b\|\right).$$

However, Hausdorff distance is very sensitive to outlier points (Dubuisson and Jain 1994), to reduce this sensitivity we used the modified Hausdorff distance (*MHD*), which corresponds to the maximum value between the arithmetic mean of the minimum distances from all points of the first set A to the second set B, and the arithmetic mean of the minimum distances from all points of the second set B to the first set A. It can be defined as:

$$MHD(A,B) = \max\left(h'(A,B), h'(B,A)\right)$$

$$\text{where } h'(A,B) = \frac{1}{|A|}\sum_{a \in A}\min_{b \in B}\|a-b\|$$

2.3. Artificial intelligence

One of the definitions of artificial intelligence is the writing of any program that can learn to perform a task that has not been previously programmed to do so. Within artificial intelligence, there are the classification algorithms that are used to identify the class of a sample or group of data samples (Russell and Norvig 2010).

They can be separated into two groups: supervised classification algorithms, in which the algorithm learns from labeled examples; and unsupervised algorithms, used to group/cluster the data samples in problems where the examples are not labeled.

Typically, the automatic identification of species from images is treated as an artificial intelligence supervised classification problem.

In general, these algorithms work as follows: the training data is presented to a classifier; it adjusts its internal parameters to try to predict the class of a future sample. Generally, together with the training set, there is a validation and a test set, both previously labeled. The first is used to improve the classifier's performance during training, and the second to test the classifier's accuracy.

There are several supervised classification algorithms, the most common are as follows: Bayesian network; artificial neural network (ANN); multilayer perceptron (MLP); decision tree (DT); support vector machine (SVM); and K-nearest neighbors (KNN).

Bayesian networks are statistical models of probability that are assembled during the training phase. If all the information in a universe is known, the model correctly matches the classification of any sample with 100% accuracy, but, in practice, only a percentage of the universe is known.

Artificial neural networks (ANNs) are computational models inspired by the biological neural networks that make up the brain of animals. It is formed by artificial neurons called perceptrons that have their values adjusted during the training phase. A perceptron is a mathematical model of representation of a biological neuron. During the training phase, the weights of the connections between neurons are updated in order to learn from the data training samples.

The *multilayer perceptron* (MLP) is one of the most used ANNs, formed by several layers of perceptrons, in order to solve more complex problems than the ones that can be solved by an ANN with only one or two layers.

The *decision trees* are binary trees built from the training set, where each internal node represents a condition that will be tested on an attribute, each branch represents the outcome of the tested condition, and each leaf represents the resulting label (the classification performed after the tests of the conditions). The paths from the root to a leaf represent the classification rules. A tree produced from training samples can be used to classify new ones.

Support vector machines (SVM) are supervised learning models that use regression during the training phase in order to identify the best hyperplanes (in the attributes space) to separate the training instances according to their classes.

KNN is a method based on distance, in which the class of a new instance is assigned as the most common class among its k nearest neighbors (the

k instances with the lowest distance to the new instance) (Altman 1992). There are different approaches to calculate the distance among the data samples. One of the advantages of the application of k-NN in image-related problems is the possibility of using a domain-specific distance, e.g., the modified Hausdorff distance.

2.4. Evaluation

Models that use supervised classification algorithms can be validated by cross-validation techniques. This type of validation works by separating the data into training and validation sets. As soon as the model is trained with the training set, the validation set is used to check its accuracy. This process is repeated several times, selecting different training and validation sets from all samples.

Overfit is a problem that occurs when training a model intensively over a set of samples, as the model becomes too adjusted for that set of samples and loses generalization, which makes it difficult to predict the class of a future sample.

In order to avoid overfit, the classifier can be evaluated with a third set of data, that of tests. In this case, all samples are divided into these three sets and then the model is created normally, after that, the test set is used to evaluate the classifier.

When a sample is classified by a classification algorithm, it can be said to belong to one of the following four types: true positive (T_P); true negative (T_N); false positive (F_P); and false negative (F_N).

True positive corresponds to the a sample that belongs to class A that was correctly assigned as class A. True negative represents the opposite, when the sample does not belong to class A and the classifier has identified it as such. There are two types of misclassification: a false positive, when a sample is classified as belonging to class A but it belongs to another class; and false negative, when a sample belongs to class A but it is classified as not belonging to this class.

These types of classified samples are used by classifier evaluation functions to measure the quality of the classification. There are several measures that can be used to evaluate a classifier, and they generally use the concepts of T_P, T_N, F_P, and F_N. Among them, the most common are precision (P_r); recall (R_c); F-measure (F_1); and accuracy (A_c).

Precision, or P_r, is the fraction of samples that have been identified and are correctly classified as true positives.

Recall, or R_c, is the fraction of samples that have been correctly identified as belonging to a particular class among all samples that are in that class.

The F-score or F_1 score is the measure that combines precision and recall, that is, it is the harmonic mean between them. The closer to 1 the better the measurement, and the closer to 0, the worse.

Accuracy, or A_c, is the fraction of samples that were correctly classified as belonging and not belonging to a class among all samples. That is, of all samples, what percentage is correctly classified.

Figure 1 represents a summary of the measures presented in this section.

3. RELATED WORK

Winged insects can be divided into two groups according to their wing opacity: insects with opaque wings, and those with translucent ones (Rebelo et al. 2020). Opaque wings are those that the light can not pass through, and in translucent wings the opposite occurs, the light can pass through, but only partially, making the wings appear semi-transparent(Sciences 1999).

Bees are in the group of translucent wing insects, together with wasps, fruit flies, and mosquitoes, all being part of the orders Hymenoptera (bees and wasps) and Diptera (fruit flies and mosquitoes) (Stork 2018). Studies found in the literature of this group tend to use shape features to extract data from the wings and it is rare to find some study using texture and color features due to the wings' translucent nature (Sonnenschein et al. 2015; Brkljač et al. 2012; Faria et al. 2014; Neto et al. 2017). Insects with opaque wings, e.g., butterflies and moths, benefit more from color and texture features; thus, these techniques are found with more frequency when we study insects of the Lepidoptera order (Kaya and Kayci 2014; Zhu and Zhang 2010; Li and Xiong 2018).

In the early 1980s, the first automated system was proposed to solve the problem of bee species

classification; it was the Africanized Bee Identification System (FABIS). Unfortunately, it could take several hours to train the model and process all the data (Francoy et al. 2008).

In the year 1991 (Schröder et al. 1995), the Automatic Bee Identification System (ABIS) was proposed and used by the scientific community until 2005, when it was discontinued (Francoy et al. 2008; Rojas et al. 2016). Although ABIS successfully classified the bee species with more precision and faster than its predecessors, it was not fully automatic, requiring some manual input by the user (Steinhage et al. 2001).

After the ABIS was discontinued, several studies were conducted, and some systems proposed, but generally, only a few species were analyzed, and the systems tend not to be fully automatic.

A detailed review of automatic methods for the classification of winged insects can be found in the work of Rebelo et al. (Rebelo et al. 2020).

Expanding the scope of the study to winged insects in general and not just bees, according to (Rebelo et al. 2020), the automation of the classification processes of winged insect species using computers started to attract more attention around 2009. Since then, it has been growing and continues with an upward trend as presented by Rebelo et al. (Rebelo et al. 2020).

The junctions of the wing venations are used by several studies (Steinhage et al. 2001; Santana et al. 2014b; Rojas et al. 2016; Silva et al. 2015) because they are excellent descriptors for this purpose (Santana et al. 2014b; Francoy et al. 2008; Silva et al. 2015). Thus, the challenge of building a system that automates the classification process can be summed up to the problem of identifying these junctions and their use in classification algorithms, and that is exactly what we propose to do in the next section.

4. MATERIALS AND METHODS

The dataset used in this paper is composed of 904 images of wings, from 45 bee species, from 22 genera, and from 3 wasp species from the same genus (Table I). We used 8–20 individuals per species, with preference of larger numbers when available. For this experiment, the right forewing of each individual was mounted between microscope slides, as described in (Francoy et al. 2008). The images were taken in different conditions regarding lighting, background, resolution, and zoom, and were selected for a bee specialist by presenting different challenges in their segmentation.

Therefore, there are also some explicit challenges in a few images, such as too much brightness, containing "salt and pepper" noise, two wings captured in one image, dirty wings, files of different sizes, and zoomed out images. Hence, this dataset provided the evaluation of a comprehensive method that tackles these different challenges.

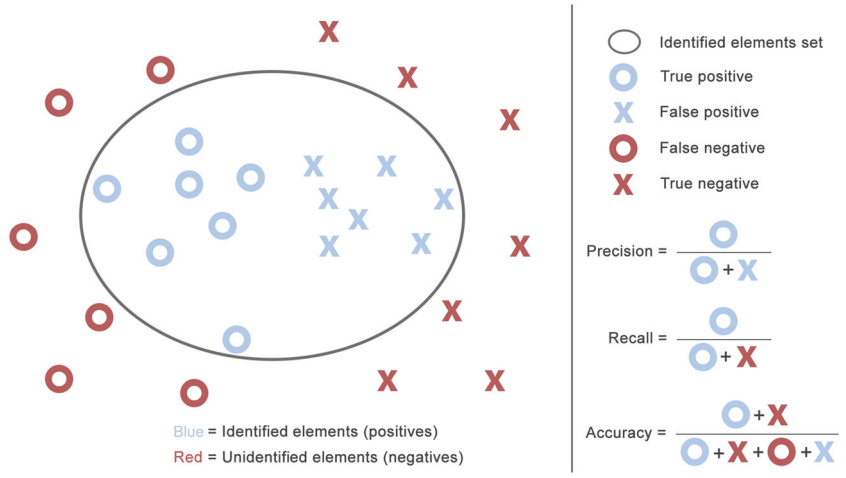


Figure 1 Precision, recall, and accuracy representation

Table I. Dataset description—45 bee species, from 22 genera , and 3 wasp species from the same genus

Tribe	Species (samples)
Euglossini	*Eufriesea violacea (20)*
	Euglossa (Euglossa) mixta (20)
	Euglossa annectans (19)
	Euglossa truncata (15)
	Eulaema nigrita (10)
	Exaerete smaragdina (08)
Apini	*Apis cerana (20)*
	Apis dorsata (20)
	Apis florea (20)
	Apis mellifera (19)
Bombini	*Bombus (Fervidobombus) brasiliensis (20)*
	Bombus (Fervidobombus) pauloensis (11)
3Meliponini	*Austroplebeia australis (18)*
	Austroplebeia cincta (20)
	Austroplebeia essinatoni (20)
	Austroplebeia striped (20)
	Austroplebeia symei (20)
	Axestotrigona ferruginea (20)
	Dactylurina staudingeri (20)
	Geotrigona sp. (16)
	Lestrimelitta limao (19)
	Melipona (Eomelipona) bicolor (14)
	Melipona (Melipona) mandacaia (20)
	Melipona (Melipona) quadrifasciata (20)
	Melipona (Melipona) subnitida (19)
	Melipona (Michmelia) flavolineata (20)
	Melipona (Michmelia) scutellaris (20)
	Melipona seminigra seminigra (20)
	Meliponula bocandei (20)
	Mourella caerulea (20)
	Nannotrigona testaceicornis (20)
	Paratrigona subnuda (20)
	Partamona helleri (20)
	Plebeia droryana (20)
	Plebeia flavocincta (19)
	Plebeia nigriceps (20)
	Plebeia pugnax (20)
	Plebeia remota (20)
	Plebeia sp (20)
	Scaptotrigona bipunctata (20)
	Scaptotrigona depilis (20)
	Scaptotrigona tubiba (19)
	Scaura latitarsis (20)
	Schwarziana quadripunctata (19)
	Trigona spinipes (20)
	Trypoxylon aurifrons (20)
	Trypoxylon lactitarse (20)
	Trypoxylon rogenhoferi (20)

4.1. Classification algorithm

The developed approach to classify bee species can be divided into three main steps: image-processing, feature extraction, and classification.

The developed algorithm for image-processing takes as input a bee wing image and extracts landmarks (vein junctions in the wing) and, to properly perform this extraction, the segmentation needs to be reliable and accurate to deal with noisy images.

The knowledge used for the specification and construction of this approach corresponds to a combination of several approaches used in the related literature on image processing, the expertise obtained studying bees' wings, and empirical experiments.

Figure 2 represents, graphically, a summary of our segmentation algorithm. The input image (Figure 2a) has good quality overall, and it was purposely chosen to be an example since bad-quality images can still be used on our algorithm, but with worse outcomes.

Right in the beginning (Figure 2b), the image is converted into grayscale, and the pixel size is reduced. Both measures reduce the complexity of the next steps. Moreover, our approach applies two smoothing filters in the image: bilateral filter and median filter, both to reduce noises.

In the second step (Figure 2c), the difference of Gaussian followed by thresholding is used to binarize the image, reduce complexity, and highlight regions of interest. The idea here is to have all pixels related to the main wing white, and black to the background.

The third step is performed to remove noise and dilate the image (Figure 2d). The goal is to remove connected components (based on their size), i.e., the removal of small objects (noises) from the image, and perform small dilations to assure that all the vein junctions are connected. It is important to alternate between these two operations, so unwanted components do not connect to the image, and parts of the wing do not get removed.

In the fourth step (Figure 2e), the most centralized wing is cropped out of the image. It is important for images with more than one wing or with large-sized noises. Besides, it makes it

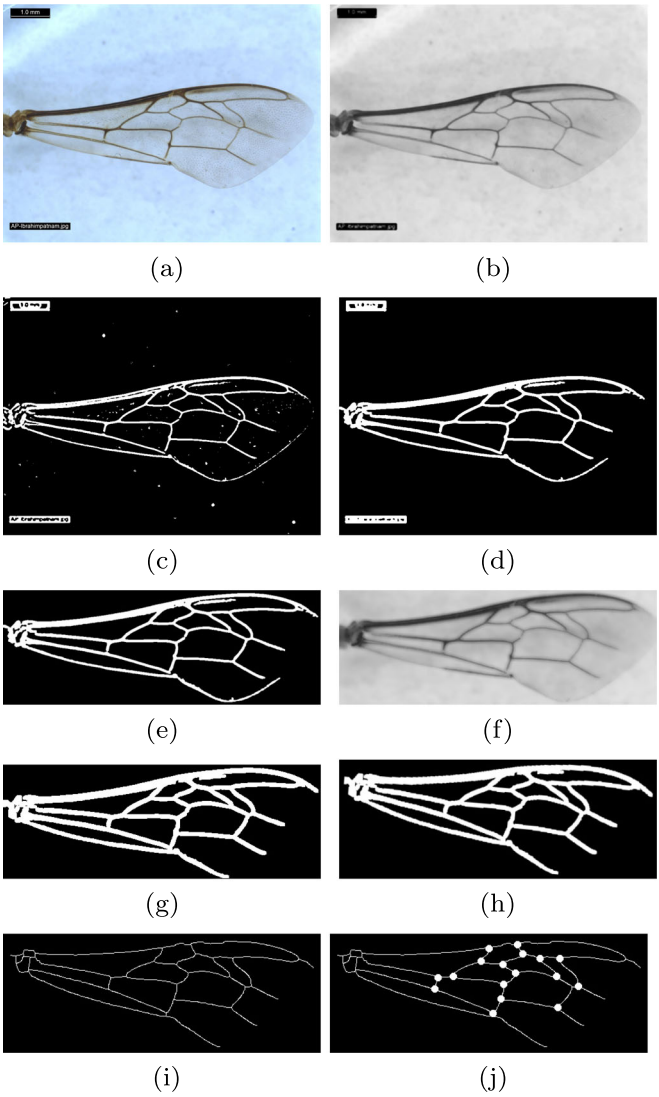


Figure 2 Sequence of steps of the segmentation algorithm. **a** Original image. **b** Grayscale image. **c** Binary image. **d** Noise reduction. **e** Wing cropped out. **f** Sharpened grayscale image. **g** Remade binary image. **h** Image rotated. **i** Skeleton extraction. **j** Features identified

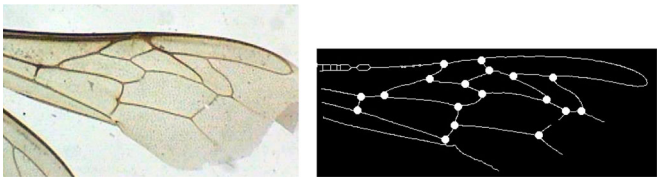
possible to reduce the image size even more because it is easier to work with a cropped out object.

In the fifth step (Figure 2f), the white pixels are converted to the original grayscale value, and the black pixels are changed to a blurred pixel of the grayscale value. In summary, the output here is the grayscale image but sharpened and with less noise.

The sixth step, illustrated in Figure 2g, fundamentally, repeats the steps 2 and 3 to get better results, and then applies a Gaussian filter. The past two steps are used to improve the results; it is especially necessary with problematic and challenging images.

To standardize the image results, so all the data become comparable, the image is resized to a specific dimension and properly rotated in the seventh step (Figure 2h).

In the penultimate step, we used the Zhang-Suen thinning algorithm to skeletonize the image.



(a) Original image (b) Segmented image

Figure 3 Two wings, wing partially photographed

It facilitates the task of detecting vein junctions. The result can be seen in Figure 2i.

The last step (Figure 2j) regards extracting the landmarks; to do that we adopted the hit-or-miss morphology operation to identify all shapes of line junction. Furthermore, it removes some of these landmarks, based on the size of the line junction and closeness with other junctions, because these are presumably not a real landmark.

Furthermore, the HSV color model was also used to improve the classification accuracy. Once we have the wing skeleton extracted, it is possible to identify the wing color by using the positions of the skeleton so that the background does not influence this extraction. And this operation results in a new feature to the classifier.

This model is formed by the components hue, saturation, and value, which represent, respectively, the color, the intensity of the color, and the illumination of the image. It is indicated for image classifications, as it isolates the lighting in a single component, which allows the disposal of that component since the lighting may not be related to the color of the wing.

The *color information* is sensitive to the type of camera, lens, and light present in the photo. It is not recommended for all applications, especially using a dataset without a clear specification, so we did tests with and without the use of this information.

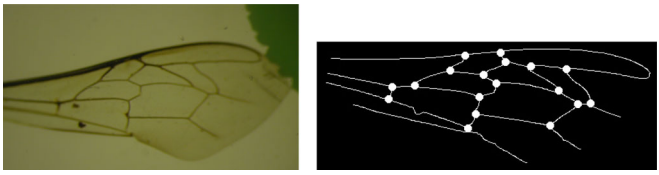
This procedure ends the segmentation step. The geometric data, that is, the coordinates of the skeleton's pixels, illustrated by Figure 2i, and the corresponding color from the original images, joined with the morphometric characteristics extracted from each wing (Figure 2j), will be used in a classifier to identify the species and the genus of each bee.

Our classification strategy uses different types and combinations of data. It also uses the KNN algorithm, with distinct similarity metrics, based on modified Hausdorff distance, which will be detailed later.

This whole dataset is the result of the segmentation algorithm. However, to increase the robustness of the classification, in addition to the standard segmentation, two other variants of the segmentation were created. Through parameterization in our algorithm, we created a second segmentation more tolerant to noise than the original, and a third less tolerant to noise compared to the original. With this, we were able to combine the data from these three different segmentation strategies to increase the robustness of the data, which consequently increased the accuracy of the classification. On the other hand, the execution time has increased, since all stages are tripled to achieve a combination of results in the end. This strategy was called *combination of segmentations*.

(a) Original image (b) Segmented image

Figure 4 Two wings, contain noise, zoomed out wing



(a) Original image (b) Segmented image

Figure 5 Broken wing, different color effect

In the classification, we use the supervised machine learning KNN algorithm. As a measure of similarity, the modified Hausdorff distance was chosen. Since different wing information was extracted, within the same algorithm, it was possible to obtain three different approaches, with different accuracy for genus and species.

To classify a wing, it starts by computing the modified Hausdorff distance of that wing in relation to all other training wings in the dataset. Therefore, the prediction will be the wing belongs to the same group as the training wing that obtained the shortest distance, that is, a classification approach based on "the nearest neighbor."

The modified Hausdorff distance will determine which approach will be used, within the possibilities. Three of them were tested, as follows.

In the *skeleton approach* , the information used is from the step represented by Figure 2i, that is, the entire skeleton of the wings. This approach is the most robust and time-consuming; for this reason, the skeleton has been reduced (resized), achieving a similar accuracy result and reducing the execution time.

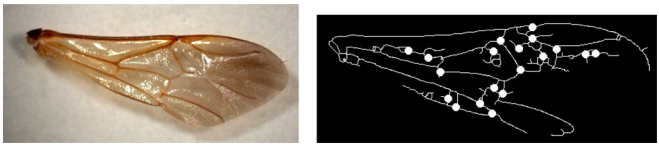
The *intersection approach* uses the features extracted in Figure 2j. This approach is the fastest because it contains a small amount of data, using only the information from the joint intersections.

Finally, the *hybrid approach* performs the distance of the features extracted in Figure 2j(intersections) in relation to the information of the skeleton represented by Figure 2i (complete skeletons). Although the approach requires different information (the skeletons and the extracted intersections), the processing cost is reduced because each skeleton is associated to a small amount of data (intersections).

In addition, it is worth mentioning that all three of these approaches have two other variables, previously explained, that can be used during the data extraction: *color information* and *combination of segmentation approaches* .

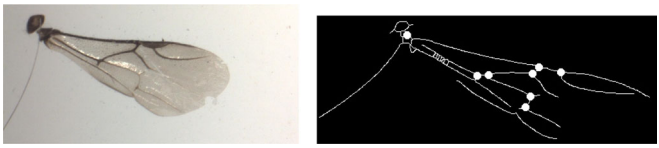
The first variable might improve the results of our approaches but it could put an unwanted bias on the equipment and lighting conditions used during the photograph of the pictures. The second consists in using different segmentation approaches and considering, as the skeletons, for example, the union or intersection of the segmented images, or, even, produce multiple skeletons to be used in the next steps. It can improve the results but it will increase the execution time of the classification.

Thus, we have three classification approaches with two optional data extraction variables, resulting in a total of 12 different possibilities strategies for classifying the species and genus of bees.



(a) Original image (b) Segmented image

Figure 6 Image too bright in certain parts



(a) Original image (b) Segmented image

Figure 7 A noise object connected to the wing

In the next section, we present a summary of the results achieved with these approaches.

5. RESULTS AND DISCUSSIONS

The identification of bee species to non-taxonomists scientists is a challenge, given the enormous variety of bee species around the world (Michener 2007) and the difficult taxonomy of the group. There are few specialists and their work should focus on the description of new species and taxonomic reviews of larger groups instead of identifying species to a larger audience (de Carvalho et al. 2007). This section shows the results of our segmentation approach, and also points out their strengths and weakness. Therefore, we selected three very positive segmentation results (Figures 3, 4, 5) and two negative ones (Figures 6, 7).

The wings of Figures 3 and 5 are similar in appearance but their images have different challenges in color, position, and noise. As can be seen, both skeleton results are good and capable of extracting all features correctly.

Additionally, Figure 4 had even more challenges. The wing is zoomed out; the image has excessive noise, and had almost two entire wings. Nevertheless, the final result was great; the algorithm could recognize the central wing and deliver a proper feature extraction, even with all the mentioned challenges.

Notwithstanding the good results, the proposed algorithm presented some limitations, especially with images very dark or very bright such as Figures 6 and 7. The problem with those images is that the noise reduction filters were not capable of dealing with that amount of light, which resulted in many improper features extracted. Interesting to note that in both cases the light is projected from a source that is above the wings, instead of a source under the wings, in which the light would pass through the membrane and contrast with the wing veins. This might be an interesting point to determine a series of procedures to increase identification efficiency.

Among the worst results we obtained, we highlight the image in Figure 7. In it, one can see a big noise very close to the wing. In this situation, our algorithm recognized the noise as part of the wing,

Table II. Classification results of all approaches to genus and species accuracy

	Acc.	Without CI		With CI		Time
		Genus	Sp.	Genus	Sp.	
With_	SA	0.9314	0.7798	0.9944	0.9601	*90X*
out	IA	0.9126	0.7411	0.8949	0.7732	*X*
CS	HA	0,907	0.7577	0.9712	0.8661	*3X*
	SA	0.9712	0.8329	0.9988	0.969	*300X*
With	IA	0.9524	0.7953	0.9524	0.8539	*4X*
CS	HA	0.9469	0.7953	0.9878	0.9048	*12X*

CS, combination of segmentation's strategies; *CI*, color information; *Acc.*, accuracy; *SA*, skeleton approach; *IA*, intersection approach; *HA*, hybrid approach; *Time*, *X* time to classify 904 images

which caused an error in the rotation and resizing steps. The result was an inadequate segmentation. The noise in question is probably some dirt that was mounted with the wing in the microscope slides used to store these organs.

All things considered, the segmentation approach fulfilled our expectations, even with the limitations. Our next time was to classify all the dataset, using a simple machine learning algorithm (KNN)(Table II).

As we can see in Table II, the results varied depending on the strategy used, with good accuracy overall. We obtained 97.1% accuracy for the genus, and 83.2% for species, which can be increased to 99.8% and 96.9% respectively if the wing color is used as a classification characteristic. The individuals that were misidentified were always confused within the same genus, except for 3 individuals out of 904. The other 25 misidentified individuals were all confused within the genus, most of it in genera that still lack taxonomic revision, such as *Austroplebeia* and *Plebeia*, which hold 17 out of 25 misidentifications. So, it is impossible to determine if problem is in the methodology we used or in the original identification of the individual. Although some of the analyzed genera can present a very difficult taxonomy, our analyses here are only related to corbiculate bees, but the initial results are very satisfactory. We are now planning to test in other groups of bees, with greater taxonomic difficulties.

The group with CS (combination of segmentation's approaches) of Table II presented superior results than the group without CS, due to the greater volume of data. However, the execution time was also higher, since the amount of data is tripled in this group.

In addition, the group with CI (color information) from Table II also showed better results than the group without CI, with even more evidence in species classifications, because the morphology of wings of the same genus can be very similar, which shows the importance of using other features. The only caveat is if the dataset has pictures taken by different cameras, lenses, or lighting, since in these situations, using color information may not be beneficial. It opens the possibility of using the different approaches on different situations. For example, one can use the color information only when analyzing specimens in the laboratory, under most controlled conditions of image acquisition, and leave the color information out of the analysis when using photographs taken in the field, with live specimens. Although we did not test, the latter situation is possible using CO_2 and a portable stereomicroscope, as stated by (Schroder et al. 2002).

According to Table II, the best approach, in terms of accuracy, was the skeleton approach, which was expected because a larger volume of data is used, so much so that the time of this approach is at least 8 times longer than that of any other approaches.

However, the approach of the intersections obtained very interesting results for such a low execution time. It is the fastest approach, and still achieved a solid accuracy of 91% regarding gender.

Furthermore, the hybrid approach presented a surprising result in the group with CS and CI in Table II, the accuracy of genus and species is quite high, compared to its low execution time.

Thus, the three approaches proved to be effective in different situations in the tested sets. It is also worth mentioning that even the most time consuming is faster than the time spent in sending the unidentified individual to a taxonomist. Another important point here is the possibility of accurately identifying most of the tested species for a broader audience, which is an old demand (Francoy and Imperatriz-Fonseca2010), leaving only the most difficult cases for the taxonomist and freeing their time for more important works. It is worth mentioning that the term "difficult cases can be understood in two ways: first, those images with a lot of noise or with other acquisition problems that make segmentation process to difficult; and second, the one that could not be identified by our classifier mainly due to its similarity with other species.

A limitation of our tests is the lack of code optimization, so it is believed that it is possible to improve the execution times for each of these approaches.

The measure of time "*X*" of Table II was used just as a matter of comparison between each approach. It represents the execution time of the fastest approach.

6. CONCLUSIONS AND FUTURE WORKS

Our methods presented a great overall result in segmentation and classification steps, still with room for improvement. Despite the fact we developed our approaches to classify bee's species, the core of our work could also be used in similar projects.

Our experiments have shown that the proposed approach reached an accuracy of 96% for species and 99% for genus. The works found in our literature review that achieved greater accuracy than ours (only six) did not focus on bees.

It is possible to observe, in the related literature (Santana et al., 2014c), that the image acquisition procedures usually are very demanding; it is requested a specific background color, low level of glare, a large occupation area of the main object, and wings in a preserved state. Considering that, in real-life conditions, it is rare to attend to these requirements, we decided to develop a more robust algorithm able to classify bees, even without these optimal circumstances.

The strength of our approach is the combination of segmentation techniques that have been used. This approach allowed for a very high accuracy, in fact greater than the related works in the literature, even with the use of an extremely simple classifier, such as the one used in the present work. It is worth noting that our dataset (48 species from 23 genera) was larger than the sets used in related works. In addition, our approach is fully automated, requiring no user interference during all the processes.

The limitations found in our experiments are related to images containing a large amount of noise and images with poor lighting conditions.

We can proceed with this work in at least two aspects: (1) the use of more sophisticated classifiers in order to improve the accuracy of the results; (2) with the improvement of segmentation processes so that noisy images and/or images with poor light conditions can be better segmented.

A free application is on the list of future developments, aside from the addition of more genera and species, in order to create a tool for everyone interested on bees and their identification, since there is a growing number of beekeepers and bee enthusiasts worldwide.

Availability of data and material/data availability Confusion matrix without color information: https://bit.ly/2SlyG41;
Confusion matrix with color information: https://bit.ly/3b1SEYb

Code availability At the moment, the code is not sufficiently documented so that we can share it.

AUTHOR CONTRIBUTION

HHB, TMF, and LAD conceived this research and designed experiments; ARR and JMGF performed experiments and analysis; TMF provided the dataset. All the authors read and approved the final manuscript.Funding

University of São Paulo

DECLARATIONS

Ethics approval Not applicable

Conflicts of interest/Competing interests Not Applicable

REFERENCES

Altman, N.S.: An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. **46**(3), 175–185 (1992). DOI https://doi.org/10.1080/00031305.1992.10475879

Ascher, J.S., Pickering, J.: Discover Life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila) (2020)

Brkljač, B., Panić, M., Ulibrk, D.Ć., Crnojević, V., Ačanski, J., Vujíc, A.: Automatic hoverfly species discrimination. In: ICPRAM 2012 - Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods, vol. 2, pp. 108–115 (2012)

de Carvalho, M.R., Bockmann, F.A., Amorim, D.S., Brandão, C.R.F., de Vivo, M., de Figueiredo, J.L., Britski, H.A., de Pinna, M.C.C., Menezes, N.A., Marques, F.P.L., Papavero, N., Cancello, E.M., Crisci, J.V., McEachran, J.D., Schelly, R.C., Lundberg, J.G., Gill, A.C., Britz, R., Wheeler, Q.D., Stiassny, M.L.J., Parenti, L.R., Page, L.M., Wheeler, W.C., Faivovich, J., Vari, R.P., Grande, L., Humphries, C.J., De Salle, R., Ebach, M.C., Nelson, G.J.: Taxonomic impediment or impediment to taxonomy? a commentary on systematics and the cybertaxonomic-automation paradigm. Evol. Biol. **34**(3), 140–143 (2007)

Dubuisson, M.P., Jain, A.K.: A modified hausdorff distance for object matching. In: Proceedings of 12th international conference on pattern recognition, pp. 566–568. IEEE (1994)

Faria, F.A., Perre, P., Zucchi, R.A., Jorge, L.R., Lewinsohn, T.M., Rocha, A., Torres, R.D.S.: Automatic identification of fruit flies (Diptera: Tephritidae). J. Vis. Commun. Image Represent. **25**(7), 1516–1527 (2014). DOI https://doi.org/10.1016/j.jvcir.2014.06.014

Francoy, T., Imperatriz-Fonseca, V.: A morfometria geométrica de asas e a identificã‡ãfo automãtica de espécies de abelhas. Oecol. Australis. **14**(1), 317–321 (2010)

Francoy, T.M., Wittmann, D., Drauschke, M., Müller, S., Steinhage, V., Bezerra-Laure, M.A., Jong, D.D., Gonçalves, L.S.: Identification of Africanized honey bees through wing morphometrics: Two fast and efficient procedures. Apidologie (2008). DOI https://doi.org/10.1051/apido:2008028

Gao, Y., Wang, M., Ji, R., Wu, X., Dai, Q.: 3-d object retrieval with hausdorff distance learning. IEEE Trans. Ind. Electron. **61**(4), 2088–2098 (2014). DOI https://doi.org/10.1109/TIE.2013.2262760

Gaston, K.J., O'Neill, M.A.: Automated species identification: Why not? Philos. Trans. R. Soc. B: Biol Sci **359**(1444), 655–667 (2004). DOI https://doi.org/10.1098/rstb.2003.1442

Hebert, P.D.N., Cywinska, A., Ball, S.L., de Waard, J.R.: Biological identifications through dna barcodes. Proc. R. Soc. Lond. Ser. B Biol. Sci. **270**(1512), 313–321 (2003). DOI https://doi.org/10.1098/rspb.2002.2218

Houle, D., Mezey, J., Galpern, P., Carter, A.: Automated measurement of Drosophila wings. BMC Evol. Biol. **3**(1), 25 (2003). DOI https://doi.org/10.1186/1471-2148-3-25

Kaehler, A., Bradski, G.: Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library, 1st edn. O'Reilly Media, Inc. (2016)

Kaya, Y., Kayci, L.: Application of artificial neural network for automatic detection of butterfly species using color and texture features. Vis. Comput. **30**(1), 71–79 (2014). DOI https://doi.org/10.1007/s00371-013-0782-8

Li, F., Xiong, Y.: Automatic identification of butterfly species based on HoMSC and GLCMoIB. Vis. Comput. **34**(11), 1525–1533 (2018). DOI https://doi.org/10.1007/s00371-017-1426-1

Lu, A., Hou, X., Lin, C., Liu, C.L.: Insect species recognition using sparse representation. In: Proceedings of the British Machine Vision Conference, pp. 1–10. BMVA Press (2010). DOI https://doi.org/10.5244/C.24.108

Marinov, E.: On the algorithmic aspect of the modified weighted hausdorff distance. Information Models & Analyses p. 126 (2012)

Martineau, M., Conte, D., Raveaux, R., Arnault, I., Munier, D., Venturini, G.: A survey on image-based insect classification. Pattern Recogn. **65**, 273–284 (2017). DOI https://doi.org/10.1016/j.patcog.2016.12.020

Michener, C.: The Bees of the World. Johns Hopkins University Press (2007)

Neto, F., Braga, I., Harber, M., Paula, I.: Drosophila melanogaster gender classification based on fractal dimension. In: Proceedings - 30th Conference on Graphics, Patterns and Images, SIBGRAPI 2017, pp. 193–200 (2017). DOI https://doi.org/10.1109/SIBGRAPI.2017.32

on Biodiversity, I.S.P.P., Ecosystem Services, I: Assessment Report on Pollinators, Pollination and Food Production (2016). DOI 10.5281/zenodo.3402857

Potts, S.G., Biesmeijer, J.C., Kremen, C., Neumann, P., Schweiger, O., Kunin, W.E.: Global pollinator declines: trends, impacts and drivers. Trends Ecol. Evol. **25**(6), 345–353 (2010)

Rebelo, A., Fagundes, J., Digiampietri, L., Bíscaro, H.: Methods for automatic image-based classification of winged insects using computational techniques: A systematic literature review. In: SBSI'20 (2020)

Rockstrom, J., Steffen, W., Noone, K., Persson, A., Chapin, F.S., Lambin, E.F., Lenton, T.M., Scheffer, M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sorlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J.A.: A safe operating space for humanity. Nature **461**(7263), 472–475 (2009)

Rojas, J.P.P., Bogantes, M.R., Monge, I.A., Mata, G.F., Gonzalez, C.T., Gonzalez, E.H.: Automatic discrimination of Costa Rican stingless bees based on modified SIFT of its wings. In: 2016 IEEE 36th Central American and Panama Convention, CONCAPAN 2016 (2017). DOI https://doi.org/10.1109/CONCAPAN.2016.7942339

Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3 edn. Prentice Hall (2010)

Santana, F., Costa, A., Truzzi, F., Silva, F., Leal, S., Francoy, T., Saraiva, A.: A reference process for automating bee species identification based on wing images and digital image processing. Ecol. Inform. **24** (2014a). https://doi.org/10.1016/j.ecoinf.2013.12.001

Santana, F.S., Costa, A.H., Truzzi, F.S., Silva, F.L., Santos, S.L., Francoy, T.M., Saraiva, A.M.: A reference process for automating bee species identification based on wing images and digital image processing. Eco. Inform. **24**, 248–260 (2014b). DOI https://doi.org/10.1016/j.ecoinf.2013.12.001

Schröder, S., Drescher, W., Steinhage, V., Kastenholz, B.: An automated method for the identification of bee species (hymenoptera: Apoidea). In: Proc. Int. Symp. on Conserving Europe's Bees, pp. 6–7 (1995)

Schroder, S., Wittmann, D., Drescher, W., Roth, V., Steinhage, V., Cremers, A.: The new key to bees: Automated identification by image analysis of wings. In: Pollinating Bees - The Conservation Link Between Agriculture and Nature, pp. 2009–2016. Ministry of Environment / Brasília (2002)

Sciences, T.: What determines whether a substance is transparent? For instance, why is silicon transparent when it

is glass but not when it is sand or a computer chip? (1999)

Silva, F.: Automated bee species identification through wing images. Ph.D. thesis, University of Sao Paulo (2015). https://doi.org/10.13140/RG.2.1.3987.9849

Silva, F., Grassi Sella, M., Francoy, T., Costa, A.: Evaluating classification and feature selection techniques for honeybee subspecies identification using wing images. Comput. Electron. Agric. **114**, 68–77 (2015). DOI https://doi.org/10.1016/j.compag.2015.03.012

Sonnenschein, A., VanderZee, D., Pitchers, W.R., Chari, S., Dworkin, I.: An image database of Drosophila melanogaster wings for phenomic and biometric analysis. GigaScience **4**(1), 25 (2015). https://doi.org/10.1186/s13742-015-0065-6

Steinhage, V., Arbuckle, T., Schröder, S., Cremers, A., Wittmann, D.: Abis: automated identification of bee species. In: BIOLOG Workshop (2001)

Stork, N.E.: How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? Annu. Rev. Entomol. **63**(1), 31–45 (2018). DOI https://doi.org/10.1146/annurev-ento-020117-043348

Strauss, R.E., Houck, M.A.: Identification of Africanized honeybees via nonlinear multilayer perceptrons. In: IEEE International Conference on Neural Networks - Conference Proceedings, vol. 5, pp. 3261–3264 (1994)

World Commission on Environment and Development: Our common future. Oxford University Press, Oxford (1987)

Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Commun. ACM **27**(3), 236–239 (1984). DOI https://doi.org/10.1145/357994.358023

Zhong, Y., Gao, J., Lei, Q., Zhou, Y.: A vision-based counting and recognition system for flying insects in intelligent agriculture. Sensors (Basel, Switzerland) **18**(5), 1489 (2018). https://doi.org/10.3390/s18051489

Zhu, L.Q., Zhang, Z.: Auto-classification of insect images based on color histogram and GLCM. In: Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2010, vol. 6, pp. 2589–2593 (2010). DOI https://doi.org/10.1109/FSKD.2010.5569848