An Automated Approach to Bee Identification from Wing Venation

By

Christopher Jonathan Hall

A thesis submitted in partial fulfillment of the

requirements for the degree of

Master of Science

(Electrical & Computer Engineering)

at the

UNIVERSITY OF WISCONSIN – MADISON

2011

To my parents

Abstract

Colony collapse disorder is killing off the world's honeybees at alarming rates. Since honeybees are the primary pollinators of much of mankind's food supply, we must look to other sources of pollination. There are thousands of different species of bees, some of which may be helpful in solving this problem. Yet many of these bees go unnoticed and remain unmonitored.

One obstacle that has slowed research of bee population dynamics among melittologists is the difficulty of identifying bees down to the species level. Modern pattern recognition techniques using images of the forewing have been used to successfully classify bee genus, species, subspecies, and even gender.

The MelittO Biotaxis System (MOBS), is described which is used to collect forewing images of living or deceased specimens. Images are preprocessed using common techniques so that features may be extracted. Using a set of training data, a feature selection algorithm selects a small set of features to use in the classification problem.

Specimens may be misclassified, because they are confused with other known classes, or they belong to an unknown class. Measures of reliability to help detect such occurrences are set forth for a Maximum Likelihood classifier and for a k-nearest neighbor classifier. Reliability measures are shown to be helpful in identifying misclassified bees under both scenarios.

A dataset of over one thousand bees representing over five genera and over twenty species of bees was tested with MOBS. Over 90% of specimens were correctly segmented. Species were correctly identified in over 90% of the tests.

Acknowledgements

I would like to thank the many people that helped me with this project. Specifically I would like to thank Claudio Gratton for establishing a relationship between the entomology department and the engineering department so that this project could take place, and then supporting the project with his time and tools. I would like to thank his students Rachel Mallinger and Hannah Gaines for sharing their expertise, collecting bee samples, and providing me access to them. I would like to thank Julia Tiede for creating our first test dataset.

I would like to thank William Sethares for initially guiding the project, his Spring 2011 Pattern Recognition class members, of which I was a part, for starting the project and testing a number of ideas. I would like to thank Chuck Hatt III and Mark Lenz for their excellent work and their continued support and feedback throughout the project.

I would like the thank the group that worked on the Automatic Bee Identification System (ABIS) whose work and articles inspired and guided my initial ideas in this project [1]. Particularly I would like to thank Volker Steinhage and Martin Drauschke for providing information on their previous work and a database of their images.

I would also like to acknowledge Sandia National Laboratories, who provided financial support without which I doubtless would not have had the opportunity to pursue the work found in this thesis. I would be remiss if I were not to thank my parents and family for the continual encouragement and support that they have provided me throughout my life, and throughout this project. I would also like to thank Almighty God for the guidance, support, and love that He has so freely given me.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Background on Bees

Pollination is a big industry; it is estimated that the annual value of pollination is approximately $65 billion [3]. Historically, our main tool for pollinating our food crops has been honeybees, in the last ten years Colony Collapse Disorder (CCD) has been killing off the world's honeybees at alarming rates [4]. The incredible diversity among bees suggests that pollination efforts may be enhanced by relying on different species of bees to help with the pollination. Unfortunately correctly differentiating between species or even genera of bees can be difficult and time-consuming for experts and nearly impossible for non-experts. The difficulty of quickly identifying bee species has hampered studies of trends in bee populations. There are approximately sixteen thousand different known species of bees contained in approximately four hundred genera [5].

Bees belong in the Kingdom Animalia, are part of the Phylum Arthropoda, are members of the Class Insecta, and are contained within the Order Hymenoptera, along with wasps and ants. Below the Order bees are grouped into a number of Families, Tribes, Genera, and Species.

Perhaps we are most familiar with honeybees, which belong to the Family Apidae, the Tribe Apini, and the Genus *Apis*. Another well-known bee is the bumblebee, which

belongs to the Family Apidae, the Tribe Bombini, and the Genus *Bombus*. Both of these bees are well known for producing and storing honey as well as being social insects, although bumblebee honey is not used by humans. Most bees, however are solitary. Bees nest in a variety of places including in the ground, in hollow twigs, and in other nooks and crannies.

There is incredible diversity among the group we know as bees. It is easy to overlook the value that these bees have to our economy, and the role they may have in pollinating our world's food crops may go unnoticed. It is challenging to monitor these bee populations. Organizing bees into phylogenetic branches such as genera and species is a large task alone, due to the incredible diversity in bee populations. Training individuals to reliably recognize and differentiate specimens at the genera and species level is likewise a difficult task.

1.2 Background on Bee Classification

Traditionally, an expert has classified bees. Today there are a number of bee classification guides [2, 6], or dichotomous keys, which give step-by-step instructions for identifying a given bee. Usually, these classification keys only go down to the genus level. For a bee to be properly categorized down to the species or subspecies level in practice requires a sample to be sent to a melittologist trained in species identification. Many melittologists trained in species identification have retired and are retiring in the next few years, which is making studies of bees increasingly difficult. Furthermore, it is not uncommon for these experts to disagree on particularly difficult-to-classify samples. To have an accurate and consistent "golden standard" for classifying bees would revolutionize the

methodology and consistency of results of numerous bee studies the world over [7]. For a standard to be reached the methodology should be quick, require little effort, be reliable, and be repeatable. To rapidly distinguish between bees, the experts use a number of features that are usually easy to see with the naked eye or the aid of a magnifying glass. They look at features such as the tongue, the genitalia, coloration, number of stripes, antennae length, wing venation, wing features, etc. Naturally, it is best to use as few measurements as possible while reliably identifying the species. Ideally, the measurements and identification would be performed by an automated identification system, which would require little or no user interaction. Furthermore, it would be best if the measurements were easy to gather.

Steinhage, et al. have shown that among many species of bees, wing venation features and morphometrics serve as a sufficient statistic to determine not only the genus but also the species, gender, and sometimes even the subspecies [1] . This means a single photograph of a wing can provide all the measurements required to automatically identify the family, genus, species, subspecies, and gender.

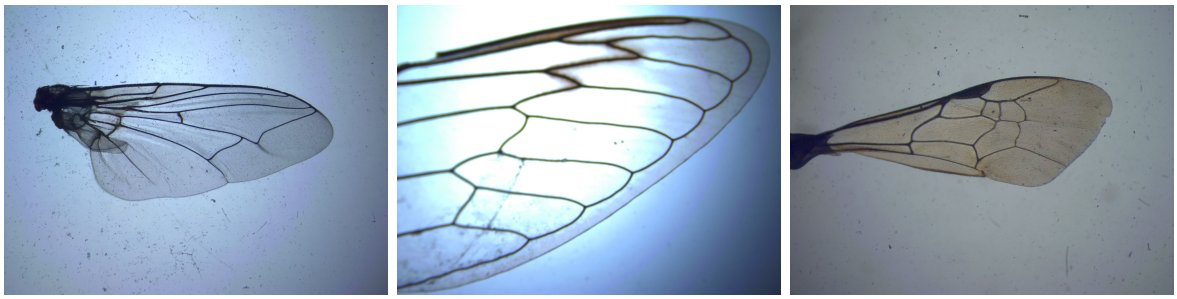
The work of this thesis builds on the ideas of Steinhage, et al. for phylogenetic classification of bee taxa (genera, species, subspecies) and for gender discrimination. This thesis works specifically with wings of species that are known to have hard-to-see venation or have other characteristics that make them particularly difficult for software to correctly process and identify.

1.3 Background on Bee Classification by Wing Venation

Characteristics in wing venation have long been used to help discriminate species of bees [5]. However, bee classification guides generally seek to use features, other than wing venation, that are easier to see at a glance. Typical wing features used are lengths between different points on the bee's wing and ratios between such distances, as well as shapes of cells.

Figures 2 through 5 illustrate both the sensibility and the difficulty of identifying insects based on wing venation. The figures start with insects distantly related to the *Lasioglossum coriaceum* (a type of sweat bee) and continue to specimens more closely related to the *Lasioglossum coriaceum*. As one might expect, the wing venation of wasps and bees are more similar than the wing venation of bees and ants, see Figure 2. An example specimen with only two submarginal cells is included, although most of the specimens included have three submarginal cells, see Figure 3. It also may become apparent why bee classification guides usually only go down to the genus level, since it becomes increasingly difficult to find distinguishing features that are readily identifiable as one goes down to the species level.

The number of wing cells and relative positions are similar among species of bees, which is why the cells in a wing have been named by specialists. One difference that can be found is that some bees have only two submarginal cells whereas other bees have three. Also, the stingless bees have notably reduced wing venation, some have as few as two cells. More diagrams of genera bee wings can be found in Mitchell's work and others [6]. Names for the different cells and veins differ based on the author [2, 3, 6]. The

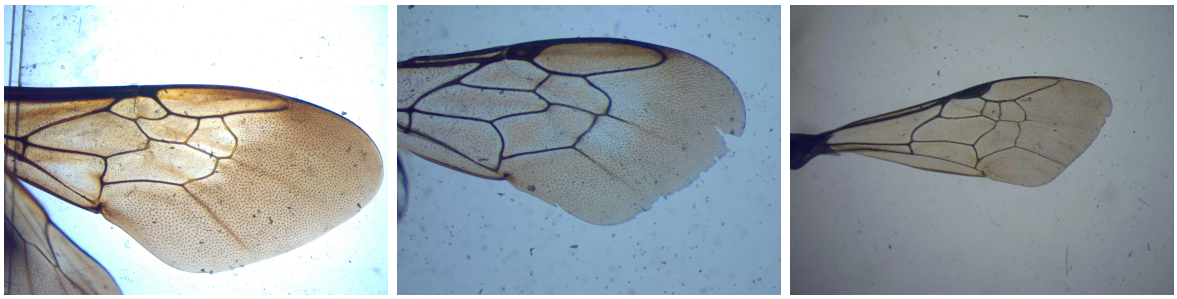


(a) Fly (b) Cicada (c) Bee

Figure 1: Various wing specimens within the Class Insecta (Insects)

(a) Wasp (b) Ant (c) Bee

Figure 2: Various wing specimens within the Order Hymenoptera

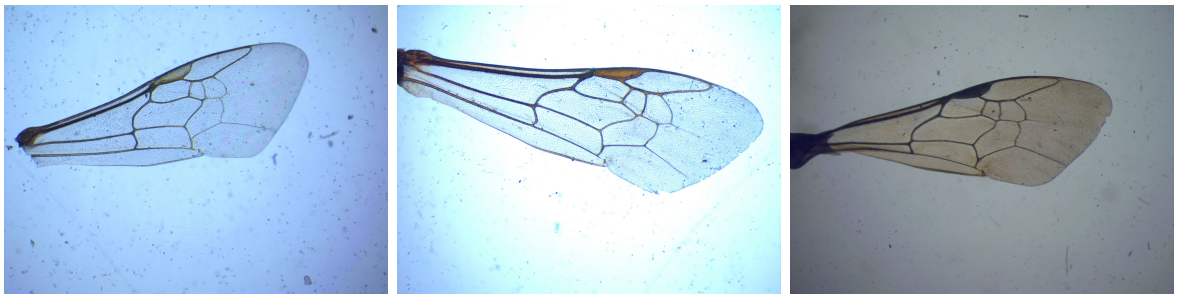


(a) Apidae (bumblebee) (b) Megachilidae (mason bee) (c) Halictidae (sweat bee)

Figure 3: Various wing specimens within the Superfamily Apoidea (Bees)

(a) *Agapostemon* (b) *Halictus* (c) *Lasioglossum*

Figure 4: Various wing specimens within the Family Halictidae (Sweat Bees)



(a) *nymphaearum* (b) *leucozonium* (c) *coriaceum*

Figure 5: Various wing specimens within the Genus *Lasioglossum*

notation given by Michener is followed in this thesis, which is depicted for convenience in Figures 6 and 7.

In the past few years, there have been several systems developed for classification of biological systems based on characteristics found in images, for insects, using wing venation has been a particularly successful area [8, 9, 10, 11, 12, 13, 3]. Among these, are a couple that have been specifically designed to work with the forewings of bees most notably the Automatic Bee Identification System (ABIS) and DrawWing [3, 14]. These systems can accurately discriminate bee species, reporting over 95% accuracy giving not only the phylogenetic taxonomy classification, but also the gender of the bee [15]. Recently, they have even been found to discriminate between bees from different locations [16, 17]. They can even be used to discriminate between difficult species of the genera *Andrena*, *Bombus*, and *Colletes* which even experts have a hard time differentiating [18, 17]. Taking measurements of the wing features by hand would be tedious, which is another reason this problem lends itself well to pattern classification techniques implemented on a computer.

1.3.1 Automated Bee Identification System (ABIS)

One of the earliest and most generalized bee classification systems is ABIS. Publications on the ABIS project range from 1997 to 2008, however the ABIS project is no longer being actively developed or supported. ABIS has been used to successfully discriminate between genera, species, subspecies, gender, and location of specimen collection. They also generalized the system to work with bees having reduced wing venation as in the case of the stingless bee [19]. The system consists of an electronic notebook with camera

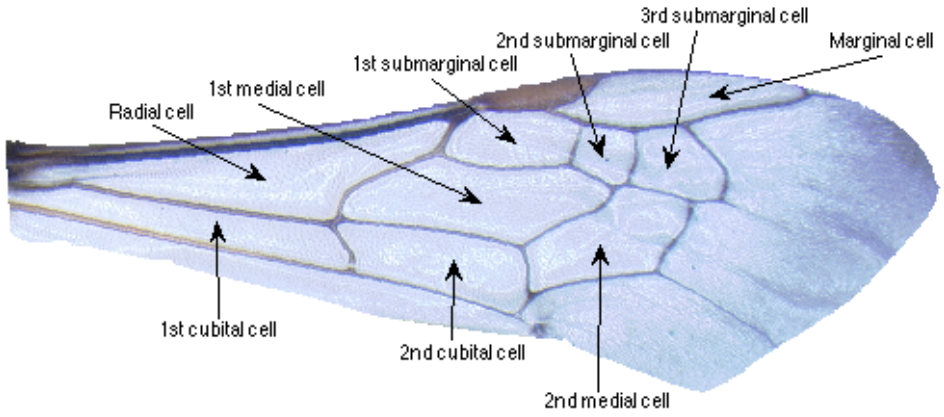


Figure 6: Names of cells in a bee's forewing, adapted from [2]

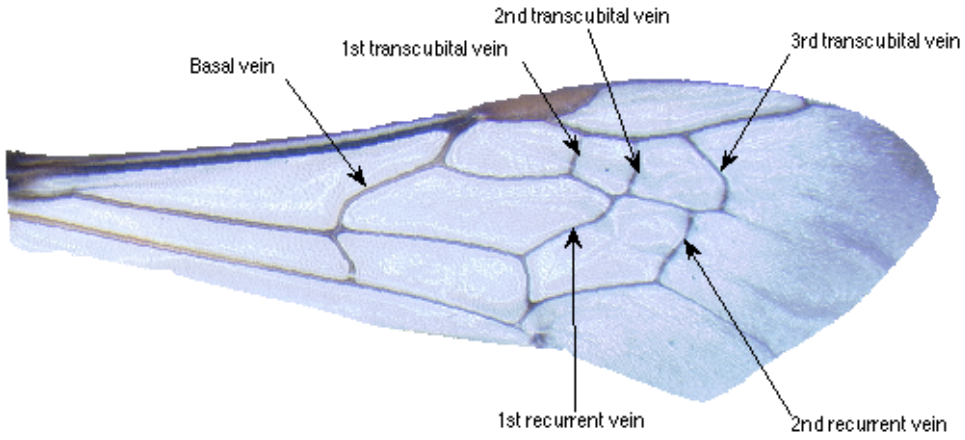
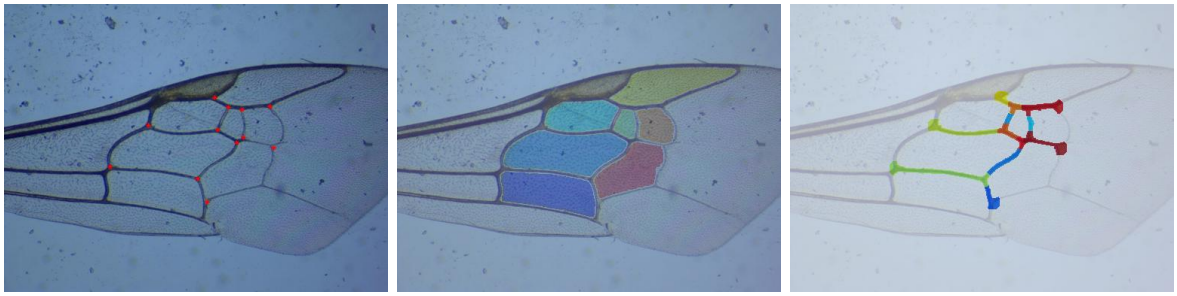


Figure 7: Names of veins in a bee's forewing, adapted from [2]

mounted on a stereomicroscope and can be used by taxonomists working in museums as well as taxonomists catching live bees in the field. The hope was that one day some institutions would provide a service where they would accept digital images of bee forewings and return the classification results to the layman. Alternatively, they hoped to create a website to which images of bee forewings could be uploaded and the identification would take place.

Some of the advantages of their system are that it was small, mobile, and handy. It works with live bees as well as mounted specimens, requires little interaction by the user, and little knowledge of taxonomy. They noted that different origins will generally yield greater intraspecific variations due to the adaptation of the bees to the specific territorial conditions, which ultimately may be a stumbling block in creating a reliable system that will continue to work as bees are gathered from around the world. ABIS reports upwards of 90% accuracy as results from leave-one-out cross-validation. Genera are distinguished using linear discriminant analysis and species are discriminated using kernel discriminant analysis [3]. A list of nearly 300 features are used which include things such as junction-position coordinates, number of incoming veins, angles between adjacent incoming veins, cell area, cell perimeter, cell compactness, cell eccentricity, veins, vein junctions, skin cells, lengths, widths, curvatures, angles, distances, coordinates, area descriptions, and pixel values from a downscaled image. Figure 8 shows some of these representative features.

Features from the first and second medial cells and second cubital cell are used for genera discrimination and features from all seven cells are used for species identification. Features are reduced using principle component analysis and the resulting features that contribute little are removed. Dendrograms and projections into a two-dimensional



(a) Vein junctions (b) Cells (c) Veins

Figure 8: Relationships between vein junctions, cell perimeters and areas, and veins serve as features

subspace are further used to visually analyze results. ABIS is not yet ready for general public use, as one of the authors noted:

- "ABIS was developed for users that are experienced enough to deal with taxonomical, morphometrical and statistical terms, since these will form the interface (i.e. input, output, parameter setting) of ABIS"

- "The project is not funded since 2004 and we are not a company with hotline or any other support. Thus, we must relay on the technical, taxonomical, morphometrical and statistical skills of the users [20]."

Perhaps the largest stumbling block to widespread use is that for a specific application the user must first complete the training phase, which includes gathering a large dataset, correctly labeling samples in the set manually, running the resulting images through a number of training steps in ABIS, manually labeling all junctions (which as Tolfilski noted is prone to inconsistencies, errors, and problems with reproducibility [3]), and verifying and validating the results. Some of these training steps are subjective, such as determining which samples to keep in the training set and which to discard, the

actual placement of the junctions, and how many specimens should be used. This process, at best, takes several weeks if the researcher already has collected a set of bees, and has correctly labeled them which means the researcher need only photograph each specimen's forewing and run the results through the training phase. This discourages new studies from groups that do not already have a collection of labeled bees. It requires an investment in time and technical training for those that have such resources. It seems that ultimately the loss of funding has caused the group to fall short of designing a system, which is easy enough to use that even a none-expert can get results. Once the training is complete, identifying a new specimen requires little specialized knowledge and is relatively quick, (one to three minutes per specimen) [15].

1.3.2 DrawWing

DrawWing is an alternative bee identification system [14, 3]. It requires that the wing first be removed from the specimen and scanned; several specimens can be scanned at once to streamline the process. From the scanned image, it determines two thresholds to produce a drawing of the wing venation similar to those found in textbooks and scientific keys. Historically, such sketches have been produced using a camera lucida, which essentially projects the image onto a screen where a piece of paper can be placed on top and the image traced. In DrawWing, the automatic thresholds are not always chosen accurately, and may need to be adjusted by the user. The sketch for some may be the final output as they may publish the image in a scientific paper, a classification guide, or a textbook. However, using this sketch, vein junctions can be found which are natural landmarks in the images and can be used for discrimination. Using landmarks

over a downscaled image has an advantage in that the features are readily interpretable and provides easy verification of results as well as an intuitive understanding of how the classification is accomplished [14, 3].

Tolfilski showed that landmarks could be found more reproducibly using an automatic system over manually labeling the key points, especially with wide veins. Another advantage is that since there are fewer features, the classification is faster. Another motivating principle is that measurements of wing length and width have proven to not be repeatable. DrawWing has also been shown to correctly identify species over 90% of the time using leave-one-out cross-validation. In a test of six hundred images it was able to find all of the landmarks over 90% of the time, a test other groups do not appear to have done. The results from the leave-one-out cross-validation seem to be from images where all landmarks/features have been successfully found. Another advantage of DrawWing is that it is open source and so the code can be freely downloaded and modified. DrawWing may fail when the wing cells contain dark areas [14].

1.3.3 tpsdig2

Yet another alternative that has been used successfully to discriminate species of bees is tpsdig2. It is a more general system that can be used to label common landmarks across images manually. It has an interface that speeds up landmark labeling by using a template image whose landmarks are superimposed onto the image. In the case of bees, the vein junctions have been used as landmarks, and although this system takes more time than the previous two (approximately five minutes per bee) it also has been shown to correctly identify species of bees over 90% of the time using leave-one-out

cross-validation [21, 16]. One of the primary advantages of this system is its flexibility since it doesn't make any assumptions about the landmarks, thus it can be used to look at more features than junctions in wing venation. However it does require a lot of human interaction, both in selecting the landmarks and then in pinpointing the selected landmarks across a variety of images.

1.4 MelittO Biotaxis System (MOBS)

Each system has limitations, which leaves room for innovation and new ideas. Each system is unique and therefore has unique benefits and problems, some of the limitations are: a limited subset of bees that are currently classified, a rigid procedure for acquiring images that may include incapacitating the subject bee, manual location of key points on the images of the bee's wings, and acquiring a large feature vector to accurately classify the bees [18, 14]. It seems as though all automated systems require a very particular method for obtaining quality images, and all systems that are flexible with the quality of images they accept require manual intervention to help the system identify key points [22, 17, 14]. It also seems that although the creators have accurately used these systems, there is little use by other groups, which may indicate that a lot of technical expertise is needed to use these systems, they may not be easy to use, and the results may be hard to repeat for new users.

Ideas from these working systems have been generously used, in an effort to make a comparable system given a shorter time frame than these multi-year projects. The system is for biotaxis of melitto, or the classification of living bees according to their anatomical (wing venation) characteristics. In other words it is a MellittO Biotaxis

System (MOBS). It is hoped that this work will lead to a system, which is:

- Reliable
- Easy to use (little training required to use proficiently)
- Robust
- Catch and Release (works with live bees as well as museum specimens)
- Quick (less than one minute per bee)
- Works automatically when given an image
- One picture per bee
- Small/mobile
- No specialized knowledge of taxonomy, statistics, or computer science required
- Provide a measure of reliability for the classification
- Intelligently choose the best features

Technical aspects of this project different from previous efforts for bee identification include the ability to automatically choose a subset of features suited to discriminating between new groups. Other systems reuse the same large feature set, which could result in overfitting and makes the results harder to analyze, understand, and verify. Also a measure of reliability on a specimen-by-specimen basis is provided so that suspicious specimens can be examined more closely. Other unique aspects relate to the specifics of the image processing and classification which are described in detail in the chapters to follow.

1.4.1 Project Evolution and Milestones

This project started out with an entomologist (Claudio Gratton) with a dream of having a system for which his students could quickly and accurately identify bees to the species level, so that it would be more feasible for them to do their research regarding dynamics and diversity of bee populations. Several emerging technologies suggested that this was a possible dream. A visiting scholar Julia Tiede, from Germany, gathered over three hundred images of bees, which included four genera and sixteen different species. The original dataset of bees were dried and pinned specimens that were photographed with lighting from above the specimen, this dataset was given to a pattern classification course which worked with the data. Their initial results showed promise, but left large areas for improvement.

At the onset the bee lab provided a data set of sixteen species that were dispersed among four different genera. They also represented a variety of different families, but they were all of the order Hymenoptera. The bees in our study were classified by an expert into the categories shown in Table 1. Bees of particular interest are those belonging to the genus *Lasioglossum* as they are noted for having week venation [5], also the genus *Bombus* had very different wing tones from the remaining genera.

In our preliminary results the genera were correctly distinguished 98% of the time in the training data and 96% in the test data. The species were correctly identified 72% of the time in the training data 52% the test data. The species in *Bombus* were correctly identified 30% of the time, in *Lasioglossum* 50% of the time, in *Agapostemon* 70% of the time, and in *Ceratina* they were correctly identified 100% of the time. These results were taken from the k-nearest neighbor classifier with a variable set of features

Family	**Subfamily**	**Tribe**	**Genus**	**Species**
Halictidae	Halictinae	Halictini	*Agapostemon*	*sericeus*
Halictidae	Halictinae	Halictini	*Agapostemon*	*texanus*
Halictidae	Halictinae	Halictini	*Agapostemon*	*virescens*
Halictidae	Halictinae	Halictini	*Lasioglossum (Dialictus)*	*MA WI sp. B*
Halictidae	Halictinae	Halictini	*Lasioglossum (Dialictus)*	*rohweri*
Halictidae	Halictinae	Halictini	*Lasioglossum (Dialictus)*	*zephyrum*
Halictidae	Halictinae	Halictini	*Lasioglossum*	*acuminatum*
Halictidae	Halictinae	Halictini	*Lasioglossum*	*coriaceum*
Halictidae	Halictinae	Halictini	*Lasioglossum*	*leucozonium*
Halictidae	Halictinae	Halictini	*Lasioglossum*	*nymphaearum*
Halictidae	Halictinae	Halictini	*Lasioglossum*	*pilosum*
Halictidae	Halictinae	Halictini	*Lasioglossum*	*zonulum*
Apidae	Xylocopinae	Ceratinini	*Ceratina*	*calcarata / dupla*
Apidae	Apinae	Bombini	*Bombus*	*impatiens*
Apidae	Apinae	Bombini	*Bombus*	*griseocollis*
Apidae	Apinae	Bombini	*Bombus*	*bimaculata*

Table 1: Initial Species Classified

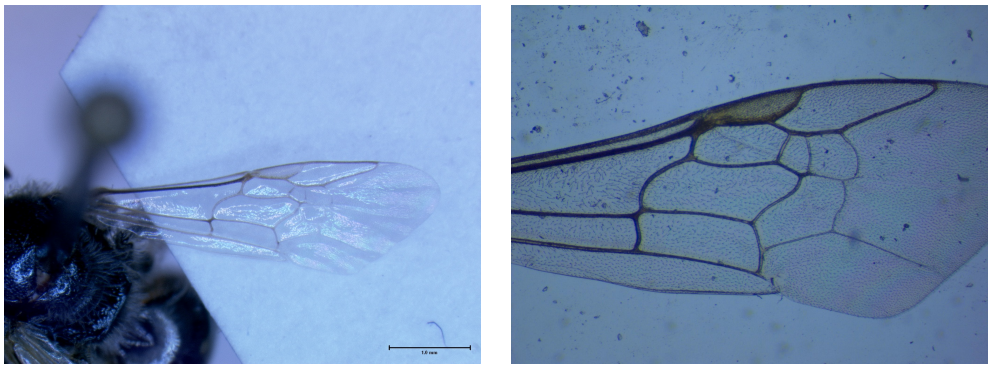
fewer than twenty, although there were instances above and below these results these were chosen as the worst result of k one to four nearest neighbors for the best number of features below twenty.

Meanwhile matching SURF and SIFT features were found in the images and RANSAC was used to remove outliers, these are common methods for comparing images [23, 24, 25]. Remaining key points could be used to create an affine homography so that a test image could be affine transformed so that key points, found by SIFT or SURF could be matched to a template image. Alternatively, when three or more junction points have been identified they can be used as key points to create the homography. Images within a genus usually matched quite well and the matches could be found relatively quickly, while intergenus images didn't match as well. Matching an image to a template using junctions seemed to be more reliable, since SIFT/SURF points would often be

mismatched.

Continued work motivated a system that would be more readily field deployable. A more transportable and economical 2D imaging system (a Leica MZ6 stereoscope with accompanying Jenoptik ProgRes digital microscope camera) replaced the previous imaging system and yet another dataset was collected. More details about the imaging system are discussed in Chapter .

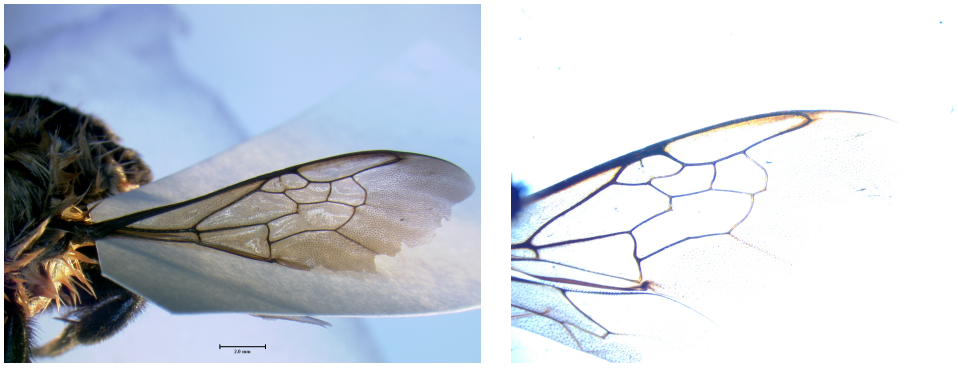
A key hope for the new imaging system was that it would eliminate reflections in the system by using substage lighting, so that faint wing venation would be more clearly seen and that the system would be field deployable and would be able to work with live bees. To keep the wing flat so that all wing venation would be in focus the wings were held between two glass slides. Faint venation and reflections were significantly reduced, see Figure 9. Also wing coloration was reduced which made the image processing easier, see Figure 10.



(a) Image with above stage lighting (b) Image with substage lighting

Figure 9: Wing venation is clearer with a different camera system for this *Lasioglossum nymphaerum* specimen

Changes in the images used by the system made feasible a different image processing



(a) Image with above stage lighting (b) 2D image with substage lighting

Figure 10: Wing venation contrast with wing cells is much greater with a different camera system for this *Bombus impatiens* specimen

algorithm, which worked more quickly, found and followed wing venation more reliably, and simplified the entire process. The results also saw improvement: species were correctly identified 90% of the time in the training data and 89% the test data initially.

Volker Steinhage, et al found that the inner cells of the wing, those cells that are named in Figure 6, are sufficient to identify the species [3]. Reducing an image to only these cells has several advantages, oftentimes the wing edges are tattered and torn so their wing cells would give less consistent results, a reduction in data is good so long as that which is necessary for classification is retained, and it is often hard to get images of the wings edge which is connected to the body, simply because a brittle bee specimen (such as those found in museums) can be resistant to such contortions. After some experimentation, it was determined that this sort of reduction in information was an appropriate one to make.

1.5 System Overview

1.5.1 System Overview

It was desirable for the program to handle bees from a variety of families and genera. Bee wing venation varies greatly between different genera which makes distinction between bee genera somewhat easier, however wing coloration also varies greatly between genera which increases the difficulty of having one image preprocessing algorithm reliably separate wing venation from the wing cells for bees of all genera, also some wings are hairy which can increase the difficulty of finding and separating the wing veins from the rest of the image.

A taxon is a group of any rank that a taxonomist judges to be a unit, such as species, genus, and family. In a very general sense taxonomic identifications follow the model shown in Figure 11, in which a sample of a taxon is observed by a sensor, the output from the sensor is then processed and features are extracted which can be used to classify the taxon, the identified species is reported from the classifier. The system described here takes the bee's wing as a taxon sample; a digital camera is used with some magnification as the sensor. A collected image is preprocessed to locate the wing venation and the wing cells from which characteristics from the wing can be easily extracted and calculated. The resulting feature vector is run through the classifier resulting in the bee species being identified.

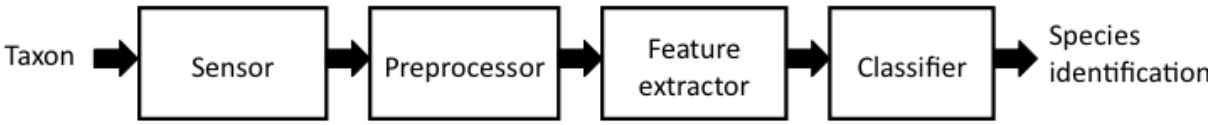


Figure 11: A general model for an automatic identification system, redrawn from [3]

Chapter 2

Data Acquisition

2.1 Data Acquisition

2.1.1 Collecting Bees

The bees used in this study were collected for a different project and were actually labeled by a melittologist with specialized knowledge in bee classification for bees found in the Midwest region of the US. The actual bee research could have gone more quickly if each bee did not first have to be labeled by an expert. Furthermore, if little specialized knowledge were required, research in areas requiring bee species identification would be accessible to a much larger group of researchers which would greatly facilitate worldwide knowledge of bees.

The bees in this project were captured using a standard procedure that involves three brightly colored cups each containing soapy water. Bees are attracted to the cups and die in the soapy water, later their bodies are collected [26, 27, 28, 29, 30]. They are cleaned and pinned. After the entomologists finished working with these specimens, they were used as training and test data for this project.

Alternatively, live bees could be captured, anesthetized, photographed, and released. This project's focus was in the training and testing stage which both require labeled

samples. These are a precursor to the implementation stage in which normal data gathering does not require a melittologist to label specimens and might consist only of live samples, especially important for endangered bees.

2.1.2 Image Collection

Images currently are gathered using a Leica MZ6 stereoscope with accompanying Jenoptik ProgRes digital microscope camera. In order for the entire bee's wing to be in focus, the wing is held between two microscope slides. Care must be taken that dirt, dust, and grime at best are first cleaned off of the slides or at worst that they do not intersect the bee's wing venation. Typical values used in the zoom of the stereoscope for this project are 1.25x, 1.6x, 2x, 2.5x, and 4x. It is important that the pictures clearly show the wing venation, the light in the stereoscope may need to be adjusted by the user so that the veins contrast nicely with the wing cells and the zoom may need some fine adjustments so that edges in the wing venation are not blurry, but in focus.

The image should contain the marginal, 1st/2nd/3rd submarginal, 1st/2nd medial, and 2nd cubital cells and none of these cells should intersect the boundary of the image. The venation around these cells should be made as clear as possible.

Images with relatively low resolution may be used. This project worked mainly with images that were approximately 400 by 500 pixels or 0.2-megapixel resolution JPEG images, although the original source images were five megapixel TIFF images. The emphasis in gathering the pictures is not the resolution, but rather that the wing venation is clear and all of the required cells are in the image with none of them directly touching on the border of the image.

In collecting datasets approximately twenty bees were used for training and ten bees for testing. Both right and left wings were included in the dataset, so as to give more data to work with for rare bees. Differences in wing venation are introduced into the wings by all differences in the phylogenetic tree as well as bee gender and the location that the bee was acquired [1, 15, 16]. For this purpose, male bees were excluded in cases where there were not enough males to have a complete set of both females and males of a given species.

Filenames followed a format similar to that used by ABIS so as to maintain compatibility between the two projects, that is the name is composed of an unique identification number, the genus of the bee, the species of the bee, the subspecies of the bee, whether the right or left wing was used, the gender of the bee, and finally the magnification of the stereoscope with each value separated by a space.

Chapter 3

Preprocessing

3.1 Image Processing Overview

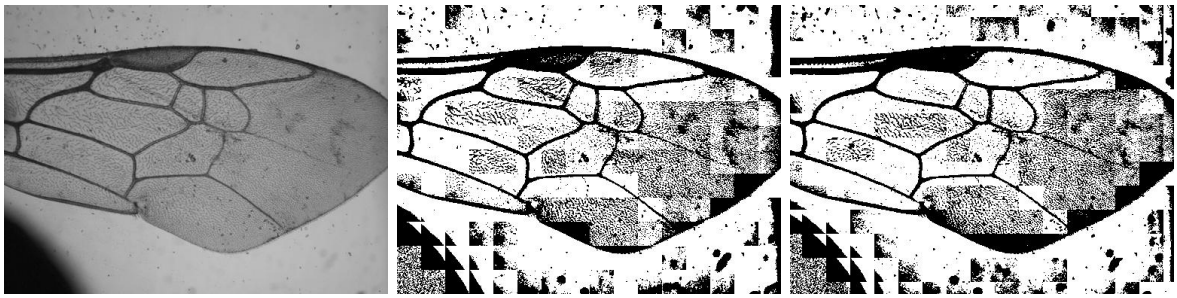
There is an old adage, "a picture is worth a thousand words." In the area of image processing, well-chosen figures can convey a lot of valuable information clearly and powerfully.

Image processing is key to accurately measuring features in the image. The goal of this image preprocessing is to make the wing's features easily measured in an automated process. Many of the features depend on having cleaned up pictures with nice veins, cells, and junctions. Ideally, all of these parts will be located and uniquely identifiable. Within the project's image preprocessing are a variety of common image processing techniques, such as thresholds, dilations, erosions, labeling of connected components, shrinking, skeletonizing, etc. [31].

The right combination of operations with good parameters, for the task, leads to an image from which it is relatively easy to identify and label vein junctions and cells. These tasks are important as any errors made in this process will propagate to the later steps, and a poor clean-up job will make correctly identifying cells and vein junctions much more difficult.

3.2 Wing venation Extraction

In some images, the wing venation can be seen so clearly and distinctly that with one global image threshold the veins can be separated reliably from the background and wing cells. However, some species of bees are known for having very faint wing venation which can be difficult for a global threshold to catch, some bees have thick hair which should not be included in the wing venation, also lighting gradients in the image cause further difficulties. To alleviate these problems, block thresholding was employed. The image was divided into several small blocks; a value to threshold with was derived for each block using Otsu's method [32]. The threshold was applied to each block, which produces a slightly blocky looking thresholded image, see Figures 12(b) and 12(c).

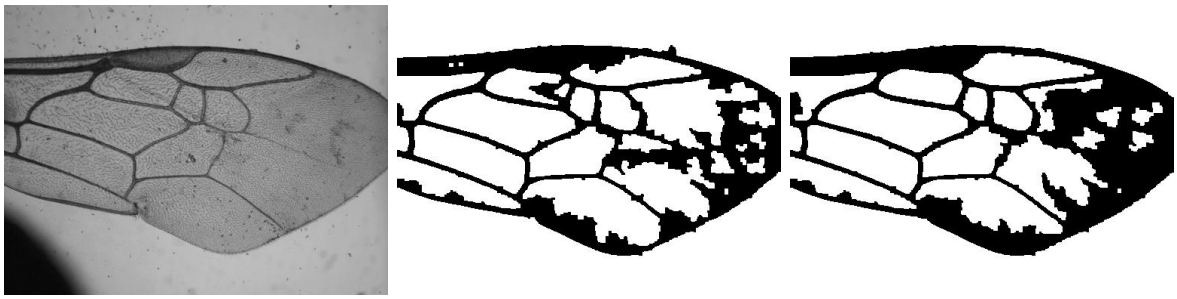


(a) Original (b) Block thresholding (c) Block thresholding different block placement

Figure 12: Block thresholding produces a black and white image of the veins

This procedure was repeated several times with the image divided into new blocks, producing several new images. For hairs close to wing veins the hairs would be thresholded to black if the image block didn't include much of the vein and white if the image block contained much of the vein, see Figures 12(b) and 13(b), but using overlapping blocks this problem was alleviated, see Figure 13(c). The actual veins were regularly

thresholded to black among the block thresholded images, although, occasionally small gaps entered in due to faint venation, breaks in the veins, or strange lighting problems. When the gaps are small enough, a simple erosion bridges the gaps. The resulting images can be summed together; a threshold is set, such that values below the threshold now correspond to wing veins and to hairs not too close to the wing veins. The wing veins are the largest connected object remaining, so can easily be extracted from the noise (dirt, grime, hairs, and pollens). Now the wing cells are remaining along with some noise, see Figure 14(b).



(a) Original

(b) Veins found using one block position

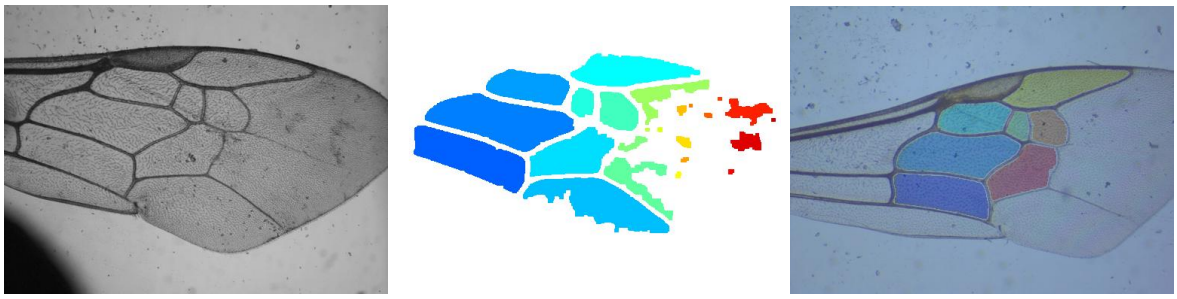
(c) Veins found using multiple block positions

Figure 13: Veins are identified in the image using a combination of different positions of block thresholds

3.3 Cell Extraction and Labeling

Using the segmented wing image, see Figure 14(b), the task is now to determine which labeled components correspond to which cells, see Figure 6. This can be a relatively easy problem on a wing for which the preprocessing has gone well and it can be a difficult problem when the preprocessing has not gone as well.

The correspondence problem is posed as a classification problem, features are gathered from the cells such as area, major axis, minor axis, Fourier Descriptors, etc. The cells are then labeled using a training set that has been manually labeled, similar to the ABIS system [3]. From this the bad cells are removed leaving only the good cells, see Figure 14(c), which have been appropriately labeled. Different colors of the cells now correspond to the label of the cell. The classification process is much more involved, but is discussed in greater detail in Chapter 5, which discusses the larger problem of bee classification in some detail. Classification of the cells is quite reliable, but options to go in and fix mislabeled cells manually are available, as the human eye and mind are able to find and label the cells more robustly than the current labeling algorithm. Depending how clean the labeled cells are they may need to be processed further to clean up some of the cells boundary conditions.



(a) Original

(b) Each connected component is labeled as a potential cell

(c) Bad cells are removed

Figure 14: Cells are found and labeled

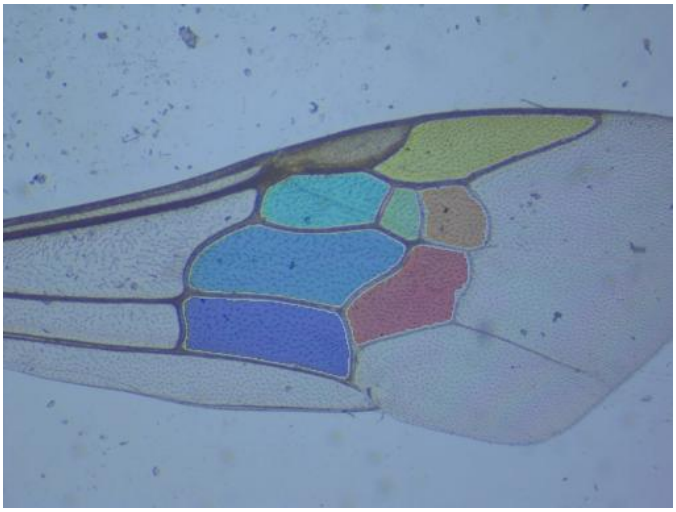
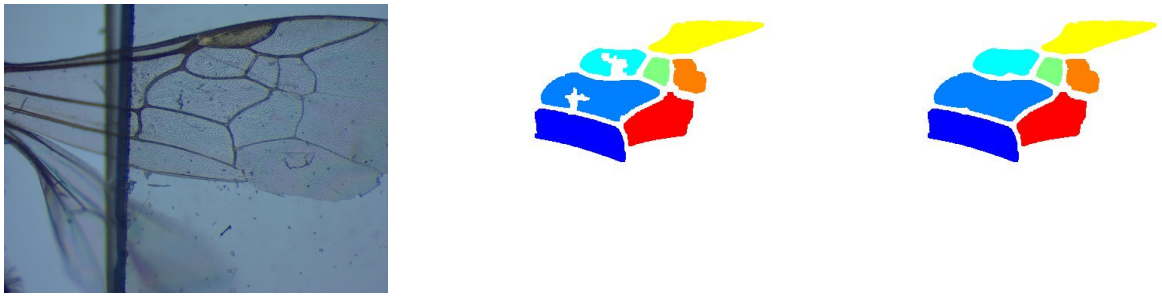


Figure 15: Cell boundaries superimposed on the original image

3.4 Cleaning Cell Boundaries

The wing cells are not quite convex shapes, however they are star-shaped, star convex, or star domains such that all lines drawn from the centroid of the cell to the cell edge are contained within the cell, as Volker Steinhage noted [3]. It is apparent that there are in fact many points within each cell aside from the centroid, which satisfy this property. The properties of star convex shapes can be used to help clean up the edges of the cells in messy images. Also, it is important to note that although the cells are not convex, the junctions of the cells occur at points which belong to the corners of the convex hull of the cell, this can be used to help find the junctions.

By treating the cell as a star convex shape, and using two points along the major axis as interior points the cell edges can be reliably smoothed eliminating impurities in the image, see Figures 16(a), 16(b), and 16(c).



(a) An image with grime inside the 1st medial cell

(b) Cells as found by algorithm

(c) Using the star convex property the cells can be reliably smoothed

Figure 16: Cell properties can be used to fill in holes

3.5 Vein Junction Extraction

The vein junction positions are described using only an x and y coordinate; yet they are often important features, albeit somewhat difficult to find in a raw image. Relative junction positions vary significantly between genera of bees, which adds to the difficulty of finding and comparing junctions, for example bees with two submarginal cells have fewer junctions than those with three submarginal cells. Among bees with the same number of submarginal cells the position of the second submarginal cell may vary enough between genera causing the vein junctions to border different cells. An example is the Nomada Obliterata vs. the Megachile genera [6]. For this reason, a large variety of bee genera were carefully examined and a numbering system that was as general and consistent as possible between all bee genera was determined, though bees with only two submarginal cells are missing some of the junctions found among bees with three submarginal cells.

Once the veins have been found and the segments labeled, as in Figure 15, a shrinking algorithm can be used on the veins to reduce the vein junctions to single pixels that can

then be found using a branchpoint finding algorithm [33]. Shrinking is superior to a similar method that takes the black and white image's skeleton and iteratively prunes endpoints to remove extraneous spurs because shrinking requires only one algorithm to take care of both the skeletonization and pruning. Also, shrinking works more quickly [22]. The branchpoints in the image skeleton are identified as junctions.

If the cells were correctly found, then the junctions between them can be reliably found, by dilating the branchpoints and observing the cells it overlaps. A dilated branchpoint overlapping three cells is the junction between those three cells. This method successfully finds all interior junctions, or junctions that border two or more cells. The exterior junctions are more difficult to find. These ideas work well for most bees, although there are species of stingless bees that have severely reduced wing venation, to the point that fewer than three interior vein junctions exist.

Once the interior junctions have been located they need to be labeled. They can be labeled by looking at the junction's cell neighbors. Some junctions with three cell neighbors are still best identified using only two cell neighbors, as the third neighbor depends on the type of bee.

The vein junctions are useful features. With as few as three touching cells accurately located, three junctions will be correctly identified. Using a template image an affine or a similarity (procrustes) transform can be performed with the three junction positions. The transformed image nearly matches the template (when they are of the same genus) and specific missing features can be more easily found, as the approximate location is known. For example, if there is a break in a vein in the image mask one can employ dilation within a region of interest specified in a genus specific template and the broken veins can be reconnected. This is especially helpful for faint venation commonly seen

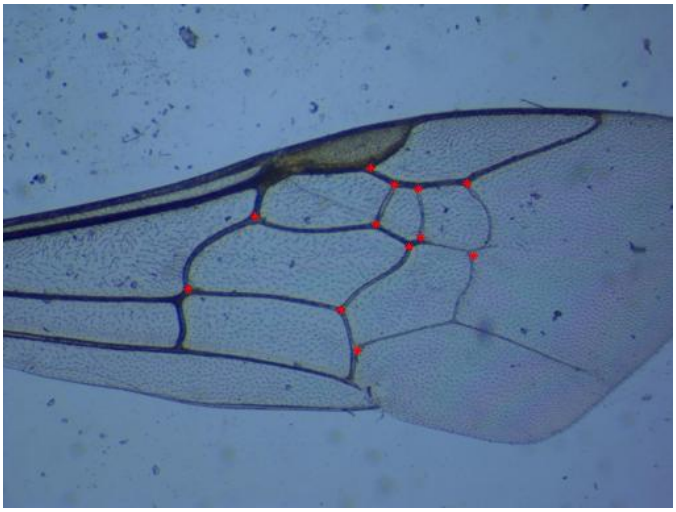


Figure 17: Interior junctions are found and labeled

along the three submarginal cells or the second medial cell.

Once all interior junctions have been found, more sophisticated techniques can be employed to find the exterior junctions. These junctions occur at local maxima in the individual cell shape signatures or at points in the convex hull. Some can be found using rules specific to the junction, for example the marginal cell's exterior junction is found along the principal axis of the marginal cell. There is a tradeoff when extra work is required to accurately identify a feature in the image, features should be easy to find relative to the problem of identifying the species. When features are found that are difficult to obtain and do not contribute significantly to improvements in identification, it is well to consider discarding them.

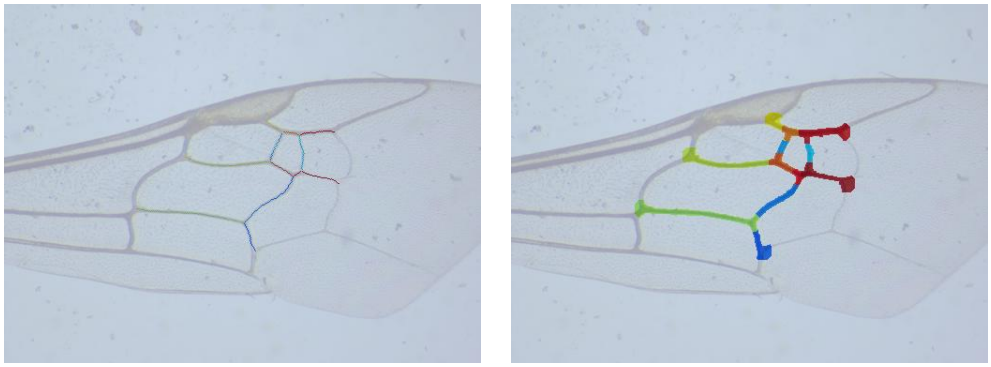
3.6 Interior Junction Labeling

Once all of the interior junctions have been found they must be labeled. The labeled cells are dilated individually one at a time to find which junctions they are adjacent to. Each junction is associated with a list of neighboring cells. Next a simple set of rules is used to determine the appropriate label of each junction. Although the rules are simple there are many, because position of the junctions and cell adjacencies can be quite different between species of bees, for an example see Figure 3. It is desirable to have the junctions across different species of bees be comparable, so the junctions were carefully labeled across species so that labels would match or at least nearly match. This also may be important in the case that there is enough genetic variability within a species that a particular junction may border different cells within individual specimens of the same species, which although I have not observed this in test data, it seems likely that this may be the case among some species.

3.7 Vein Labeling

To find and label the veins between cells, a version of the image with labeled cells and thinned veins is used. The cells are then taken two at a time and single pixels that form a bridge between the two cells are filled in and labeled as the vein between the two touching cells. This is repeated for every pair of cells and is enough to label all interior veins, or veins that border between two cells, see Figure 18.

The exterior veins are found by dilating all cells and subtracting areas that are covered by the interior veins and the cells, which leaves an outline of all of the cells. These veins can further be associated with the nearest cells, see Figure 19.



(a) Thin veins (b) Thick veins

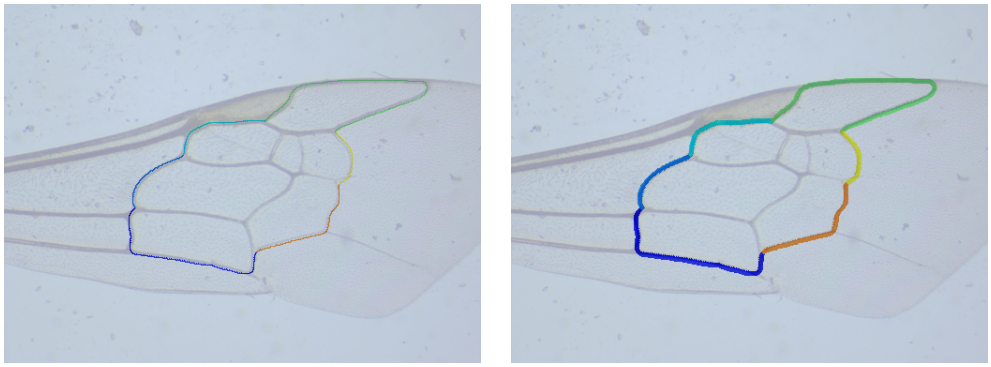
Figure 18: Interior veins

After labeling the veins, the vein lengths can be determined by summing all of the pixels that make up a thin vein. It may be useful to know how much this distance varies from the Euclidean distance between endpoints, which is easily found and can be used to create a ratio, Euclidean distance between vein endpoints to measured vein length. The veins can then be dilated to fit into the mold provided by the labeled cells without thinned veins. Average vein width is found by dividing the sum of all pixels within the thickened veins by all the pixels within the thinned veins.

3.8 Exterior Junction Extraction and Labeling

To find and label the exterior junctions, the thinned-labeled veins are used. A labeled vein containing one exterior junction is first selected. The endpoints are marked and the point furthest from the line between the endpoints is found and labeled as the endpoint, see Figure 20.

For the exterior junctions the Euclidian distance is used, for the orange and cyan



(a) Thin veins (b) Thick veins

Figure 19: Exterior veins

veins of Figure 19. The dark blue vein has two exterior junctions, so it is first split in half and then the same idea is applied. The green vein's exterior junction is the point furthest away from the midpoint of the line connecting the two endpoints. Since these junctions have been found one at a time, their labels are already known. Figure 21 shows the exterior junctions as detected and labeled by the algorithm.

3.9 Affine Transform

After the genera or species has been identified the image can be matched to a template image, to help further identify points and useful features that may help in successive classification. For this to occur it is helpful if some of the important points match between the template image and the test image. When a few junctions have been accurately found in the test image, an affine or a similarity transform can be derived from the points to match the image with the template. The affine transformation creates a transform that may involve changes in rotation, scale, translation, and/or shear to

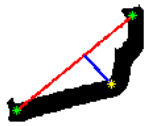


Figure 20: Endpoints of a vein are found and the point farthest from their connecting line is found

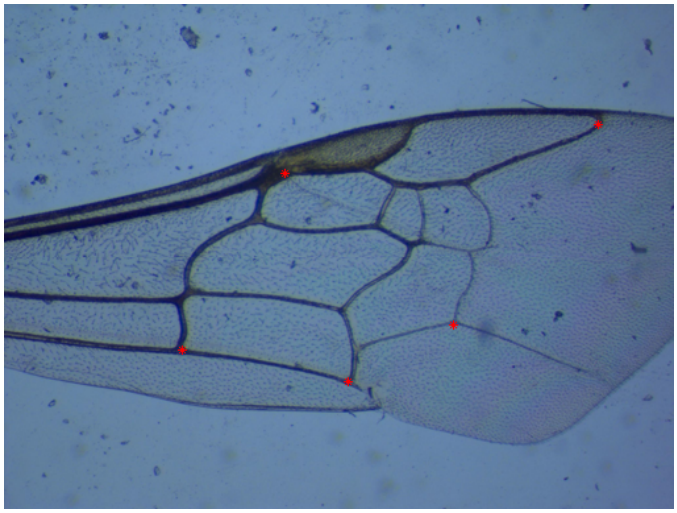


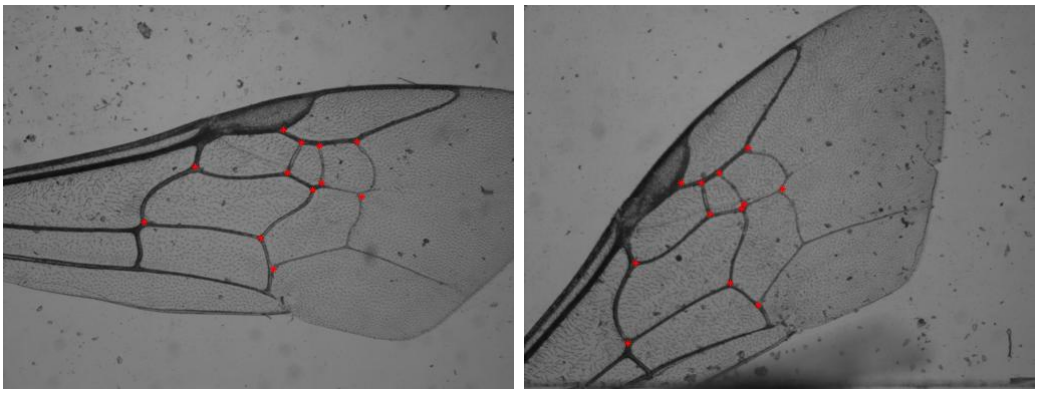
Figure 21: Exterior junctions are found and labeled

minimize the error between selected points in the transformed image and the template image. In this case, corresponding vein junctions were used to create the affine transformation. Wings of a species match closely, see Figure 22, wings within a genera match a little less closely, see Figure 23, and bees from different genera do not match so well, see Figure 24.

The similarity transform or Procrustes analysis can be used alternatively, which is a transform belonging to the affine family without the shear component. Other alternatives include a polynomial fit, a piecewise fitting, and the local weighted mean transform. Qualitatively, the affine transform and similarity transform seem to work best. The error between the test image's transformed vein junction positions and the template's vein junctions positions can also be used as a rudimentary classification technique.

3.10 Smoothing borders

To clean up the appearance of the borders further, which may still appear rough, one can simply smooth the borders by taking each border pixel to be the average of its k neighbors [31]. This produces smoother cells whose boundaries are less affected by inconsistencies in the edge boundary of the veins, and looks better to the eye. However, they may cause the cell to shrink, and the resulting features based on cells may be less reliable.



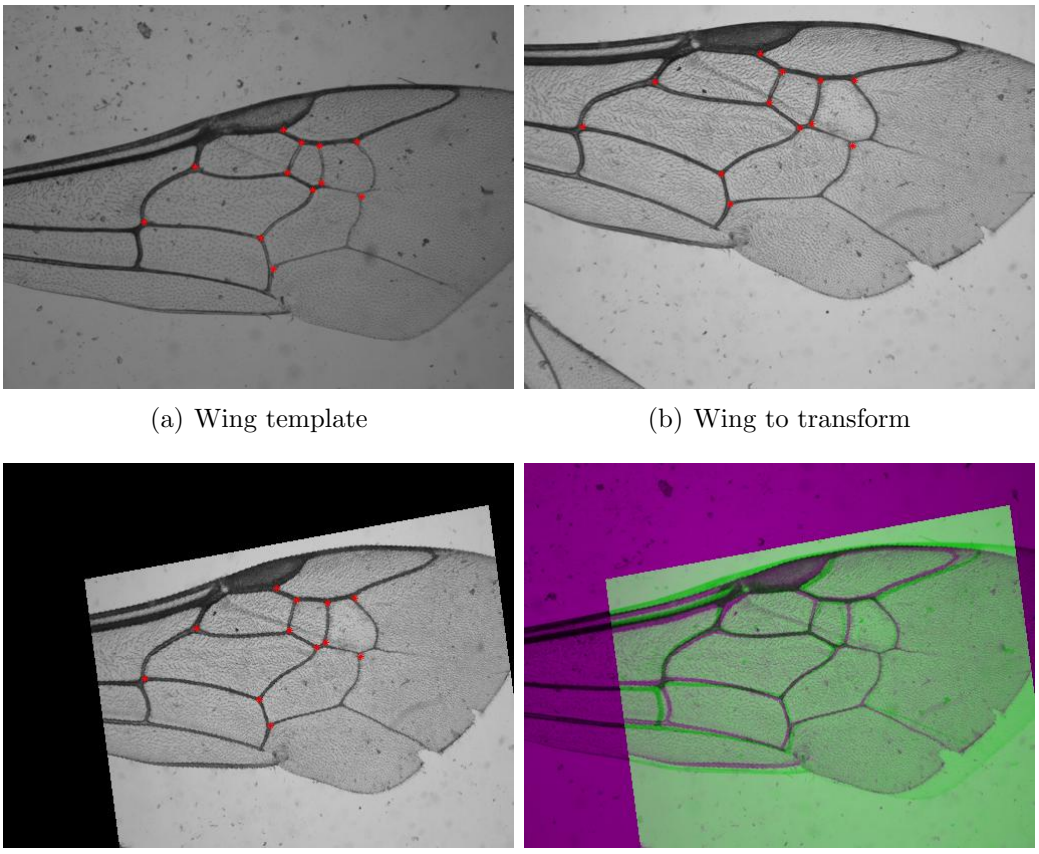
(a) Wing template

(b) Wing to transform

(c) Transformed wing

(d) Transformed wing (purple) overlaid on template wing (green)

Figure 22: Key points from a template wing and a sample wing used to align two wings of a species



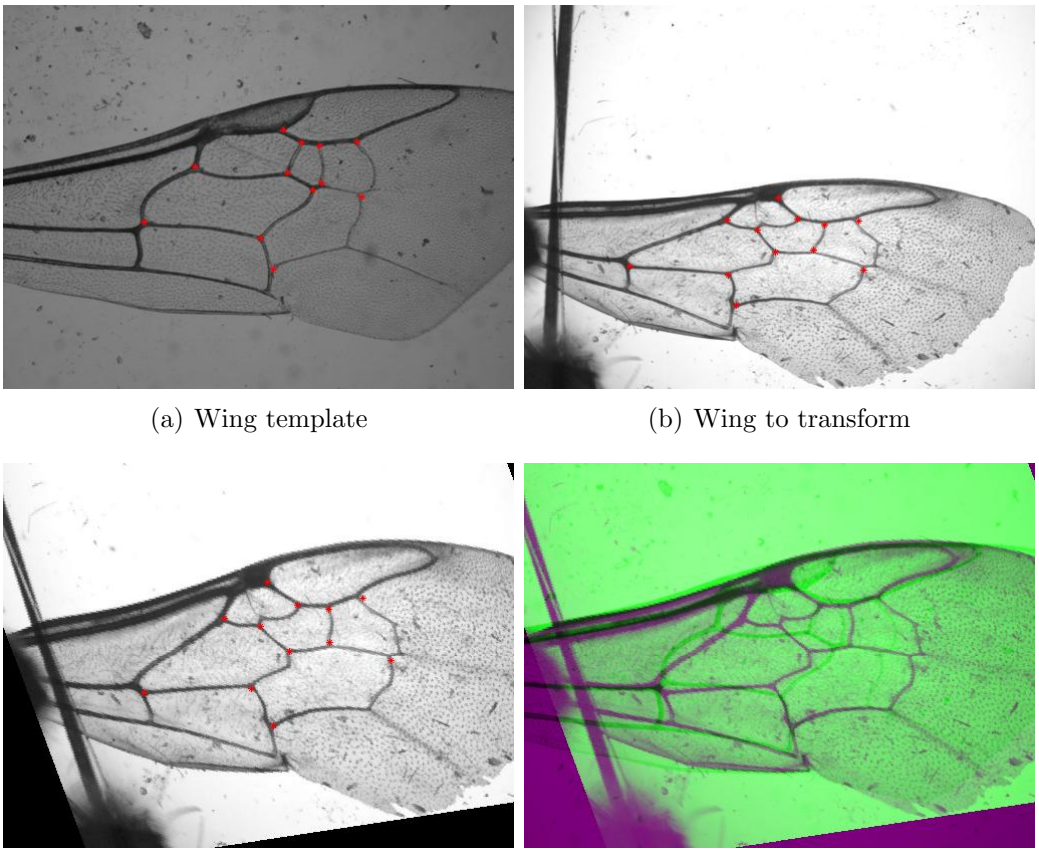
(a) Wing template

(b) Wing to transform

(c) Transformed wing

(d) Transformed wing (purple) overlaid on template wing (green)

Figure 23: Key points from a template wing and a sample wing used to align two wings of a genus



(a) Wing template

(b) Wing to transform

(c) Transformed wing

(d) Transformed wing (purple) overlaid on template wing (green)

Figure 24: Key points from a template wing and a sample wing used to align two bee wings

Chapter 4

Feature Extraction

4.1 Feature Extraction Background

Choosing good features is an essential task in the development of any successful classification system. It is desirable to have features that are:

- Easy to measure in the data
- Differentiate well between the classes (genera, species, subspecies, gender)
- Reliably found
- Robustly measured

This chapter deals with the process of extracting features used in this project, and Chapter 5 gives more details on how the features are selected for the problem at hand. Once the image preprocessing has been accomplished, the features are relatively easy to extract from the image. Some feature extraction and classification may provide feedback so that more preprocessing can be done. For example, a known genus may suggest areas to look for missing veins or missing junctions, or it may suggest a template to use to transform the image to better distinguish between subtle details.

4.2 Cell Properties

Once the cells are labeled, a number of features can be calculated, such as the area, perimeter, Fourier descriptors, orientation, solidity, major axis, minor axis, and compactness. Some of these form ratios that compare well between pictures, such as the ratio between two cell's areas, two cell's perimeters, and the ratio of a cells major to minor axis.

4.3 Key Points as Features

Key points such as the vein junctions and the cell centroids can be used to derive many features. Notably the junction locations serve as starting points for a number of easily calculated features, including the distance between junctions, ratios of distances between junctions, angles formed by three junctions, etc. These features are used by both melittologists as well as other bee identification systems [5, 3, 14]. Coordinates of cell centroids are another easily measured (in software) feature, similar to junction coordinates. Once the junctions are located, the pairwise distance between all junctions and the angle formed by any set of junctions can be calculated. Altogether, cell centroids and cell junctions define approximately twenty-five potentially useful points in the image. Successfully measuring the pairwise distance between all of these points gives $\begin{pmatrix} 25 \\ 2 \end{pmatrix} =$ 300 values, which means there are $\begin{pmatrix} 300 \\ 2 \end{pmatrix} = 44,850$ ratios of pairwise distances. Of course many of these features are redundant, but any one may be particularly useful.

To further expand the possible set of useful features, consider that the distance

doesn't need to be restricted to Euclidean distance. For wing images that have been affined transformed for a best fit to a template, the x and y components of distance also individually have meaning. Other metrics, such as L1 distance, can be considered.

Aside from using the distances between the key points, the angles between any two key points using a third key point as the vertex also forms several potentially useful features. A simple formula is then used to find this angle $cos^{-1}\left(\frac{a\cdot b}{|a||b|}\right)$ or more explicitly $cos^{-1}\left(\frac{(p_1-p_v)\cdot(p_2-p_v)}{|p_1-p_v|_2|p_2-p_v|_2}\right)$ where p_v is the position of the junction acting as the vertex and p_1 and p_2 are the positions of the remaining two junctions, which form the angle. There are $3 \times \begin{pmatrix} 25 \\ 3 \end{pmatrix} = 6,900$ angles composed of all subsets of three of these significant points.

Since the usefulness of these measurements vary in different classification problems, it would be tedious and time-consuming (if not impossible) to make all of these measurements by hand and determine which measurements are actually needed. However, by using a computer it is relatively easy to measure all of these features in the training phase so that only the most important measurements need to be taken when classifying. This requires a way of automatically determining a small set of useful features from the larger set of all measured features. Once established, this method also allows evaluation of other potential features whose usefulness can then be evaluated. The most important features are a relatively small subset of the complete feature set and once identified by the system they could be measured (by hand if necessary) for a few samples. These ideas clearly can have a significant impact not only on creating a reliable automated system, but also on the way that morphometric features are selected and used in new identification guides.

Chapter 5

Classification

5.1 Classification Overview

It should be noted that as Charles Darwin said: "No one definition has satisfied all naturalists; yet every naturalist knows vaguely what he means when he speaks of a species. Generally the term includes the unknown element of a distinct act of creation [34]." Today, the meaning of species is still somewhat ambiguous; there are a number of practical definitions in use, this makes classification of species a special challenge as there are no firm morphological methods of differentiating species and species boundaries may change with time, as they have in the past. For this reason the lines between species are often fuzzy, which means that some may classify two bees as being the same species while others may classify them as being different. Interestingly, in some cases there is enough variability in the wing venation of subspecies that they can readily be differentiated, whereas two species of another genus may be much more difficult to distinguish. In some species there is even enough variability in wing venation between the genders that males can be differentiated from females, whereas in other species the wing venation is more closely related between the genders.

After gathering feature vectors for the training samples there are still many questions to be asked before successful classification. Some of these questions include how many

features should be used, which features should be used, how the features should be scaled, how missing features should be handled, and which classifier should be used. There also needs to be a method to test the selected criteria and to interpret the results. The answers to these questions are not independent, for when a choice is made in one of these areas it may affect others.

There is no best answer to these questions for all problems. For example, the number of features required to differentiate two species of bees depends on which two species are being differentiated, for some species only a few features may be needed, whereas, for other species which bear greater resemblance more features may be required, as supported by Figures 2 through 5.

5.2 Number of Features

The number of features is an important choice; generally it is assumed that the number of samples for each class is greater than the number of features. This is especially important for the case where the features are distributed according to a multivariate Gaussian, in which to have a well-defined inverse for the covariance matrix requires more samples than there are dimensions or in this case features [35]. In theory, adding features to the classification problem should decrease the probability of error. In a truly Gaussian case the probability of error can be decreased without limit as the number of features is increased. However, in practice the addition of extra features can lead to worse rather than better performance [36]. Generally, it is a good idea to use fewer features than there are instances of each class in the training set in order to avoid overfitting and underdetermined solutions [35]. Figure 25 illustrates how generally more features used

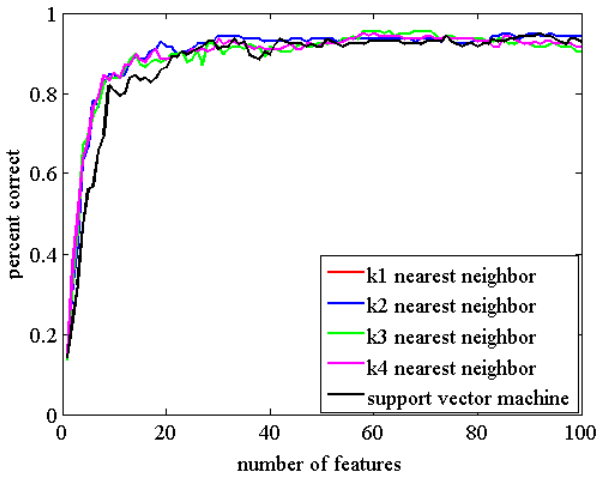


Figure 25: For a given set of features, classification accuracy is compared to the number of features used

for classification increases the classification's accuracy in the bee test set. On the other hand, it also illustrates the law of diminishing returns, since using more than thirty features does not increase the accuracy.

Just as it is generally more difficult to differentiate identical twins than it is to differentiate fraternal twins, it is more difficult (and therefore requires more or better features) to differentiate closely related species than it is to differentiate distantly related species. If the species being discriminated are closely related in a morphological sense, more features are needed and thus more samples are needed.

It is desirable to select enough features to reliably discriminate between taxa in the training set. Also, it is desirable to use fewer features because this reduces the computational complexity of the model, it means the features can be calculated more quickly, and used to classify a sample more quickly. Using fewer features also reduces the risk of overfitting the data, which leads to worse performance. However, more features

can add redundancy, which can increase reliability. Generally, more features allow for a better fit of the training data with the model. For many tasks requiring an automatic classification system, it is hard, in practice, to find an ideal feature set that gives good results with a minimum number of features.

In order to come up with the best features, a large list of easy-to-calculate candidate features was generated. The list includes items used by other notable systems as well as other features that seemed reasonable to add [1, 3, 21, 14, 17, 18, 11, 9, 37]. Distances can be measured under many different metrics, for this reason a variety of different metrics were explored when first creating the possible feature list. Some of the features in this list include: Fourier descriptors of the wing cells, cell area ratios, cell perimeter ratios, ratios of distances between cell centroids, cell major to minor axis ratio, cell eccentricity, cell orientation, solidity, extent, compactness, junction angles, junction distances, major to minor axis ratio, vein length, and average vein width. For this problem the city block or L1 metric was used, because it seemed to give the best performance.

5.3 Feature Selection

For each species, a number of features are gathered. Just as it may be easy to tell apart two people by hair color alone and another two people by eye color alone, the best features to distinguish different species of bees depends on which two species need to be told apart. For this reason, one set of features may be used to tell apart bees in the genus *Lasioglossum* and quite another set is used to tell apart bees in the genus *Bombus*. Yet another set of features is first used to tell apart the genus *Bombus* from the genus *Agapostemon*.

An initial set of features for each image is gathered for classification of the genus. Once the genus has been classified, uncertainty about which species the bee belongs to is reduced. Determining the genus is much easier than determining the species and for many research studies the genera of bee is sufficient. Because it is an easier task, this system classifies the genus before the species rather than combining the classification of the genus and species into one problem. Once the genus has been successfully identified, the test image can be matched to a template image. A new set of features is gathered which is used to classify the species. This process can be continued to attempt to differentiate subspecies and gender. It may seem surprising, but in some species it is relatively easy and reliable to tell apart the gender using only wing venation. This also means that ideally a training set should contain an ample number of both male and female specimens if both need to be identified in the deployed system stage, so the algorithm will not select features that are specific to one gender's wings.

Perhaps the most natural way to select features is to find which features have the greatest correlation as a differentiator between the classes and to continue adding features until the accuracy of the classifier is sufficient. Although this sort of Bayesian feature selector is easily implementable, it is computationally intense and it gives features that are likely to be correlated to one another.

Another approach is to seek to minimize the redundancy between the features while maximizing the relevancy to the classification task. An example of this idea is to maximize the relevancy while minimizing the redundancy between features [38]. This method is fast and gives quite good results.

A common method for combined feature de-noising and feature reduction is to use

Principal Components Analysis (PCA). Using principal components analysis is not directly a feature selection algorithm, but it does have many similarities to feature selection algorithms. First, the training data must be scaled so that no feature dominates in size, (otherwise structures with large variance dominate the calculations). Next, the principal components of a set of training data is found, which is a weight matrix. New feature vectors can be matrix multiplied by the principal components to produce a new feature vector, which orders the features from those with greatest variance to those with smallest variance. These new features are also orthogonal to one another, which means that there is no (linear) redundancy between any two features. When features are linearly redundant the principal components analysis will result in lower dimensionality, as it will not need to add weights for each feature vector. The number of components to use can further be reduced by selectively removing the new feature vectors with lowest variance, however this feature selection does not result in a decrease in the number of features that need to be gathered. Principal Components Analysis (PCA) makes the assumption that the data can be modeled with largely linear components, large variances have important structure, and the principal components are orthogonal [35, 39]. Since large variances are assumed to have important structure PCA is sensitive to outliers, as in Figure 26. These are good assumptions when the data is distributed according to a Gaussian distribution, in which case PCA will choose a basis such that the axes are independent of one another, which is desirable.

Since the problem of feature selection and classifying are intertwined, some feature selection algorithms require the classifier as an input, which it then uses to iteratively classify the dataset using different subsets of features to determine which set of features is optimal [40].

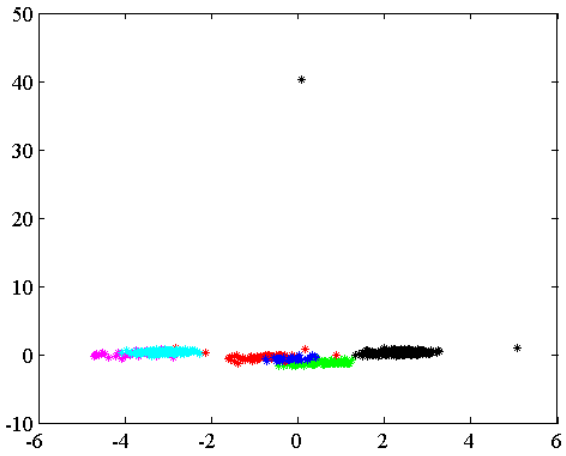


Figure 26: Principal components analysis failed to give the most useful features in this case for the two most important features, the first and third feature do quite well though, see Figure 42

Other state-of-the-art bee identification systems do not use feature selection algorithms; rather they rely on basic features that are plausibly important. This work instead uses the idea of automating the process of selecting a subset of features well-suited to the classification problem, which means its accuracy is determined by the quality of the feature selection algorithm and training set. A variety of methodologies for selecting a subset of features were tested in order to select an algorithm, which would quickly select high quality features. Specific feature selectors that were used include: Maximum Relevancy Minimum Redundancy feature selection [38], backward and forward sequential feature selection [40], Sparse multinomial logistic regression via Bayesian L1 regularisation [41], Relief-F feature selection [42], Kruskal-Wallis feature selection [43], Information Gain [44], Gini Index [45], a fast correlation based filter [46, 47], χ^2 feature selection [48], and the Bayesian logistic regression method [49]. These algorithms need

not be used independently but rather can be used in sequence to further refine a feature set.

5.4 Feature Scaling

Once the feature vector is ready for use, a natural question is that of how best to scale the features. One doesn't wish to give extra weight to features simply because they have been described with larger numbers. Rather, one would like the features to be approximately on the same scale, with extra weight given to those features deemed to be of greatest value. This way all of the features can influence the decision.

A common approach to scaling is to consider each of the features as random variables distributed according to some normal distribution, where each variable has its own mean and variance. Sometimes this is a pretty good assumption, see Figure 27, other times it is not as in Figure 28. With the assumption that the features are distributed according to the normal distribution, it is natural to subtract the mean and divide by the standard deviation so that each of the features are now distributed as the standard normal distribution (with zero mean and unit variance) and features are directly comparable. Clearly, many of the features used are not actually distributed normally, for example distance cannot be distributed according to a normal distribution, as it does not take on negative values. Even when features are not normally distributed, it may still be helpful to model them according to a normal distribution.

Yet another way to scale the features is to divide by the maximum magnitude of each feature vector, which normalizes the training data between one and negative one. In this way it is certain that there are no outliers in the training data that could significantly

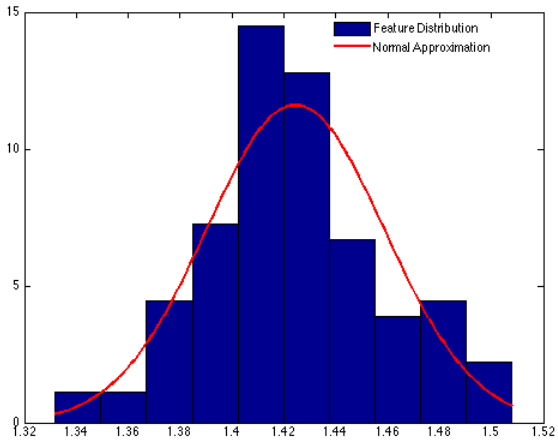


Figure 27: Normalized histogram of one feature for one hundred samples of a species with corresponding Gaussian estimate overlaid, a good fit

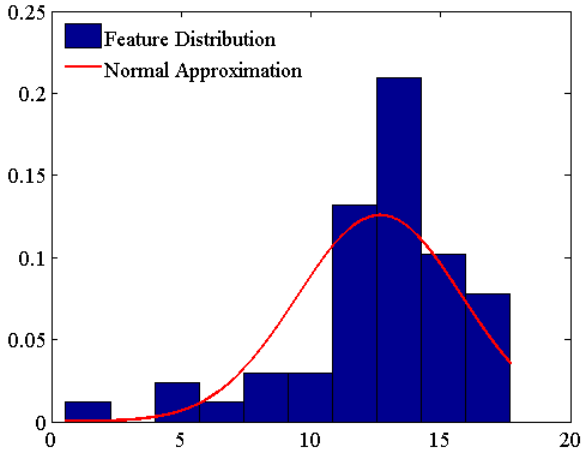
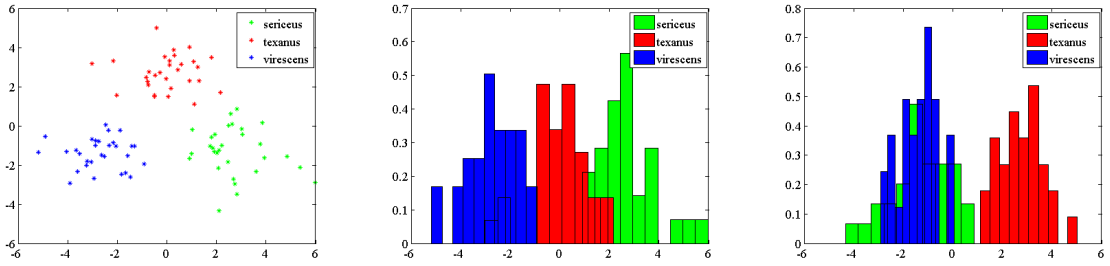


Figure 28: Normalized histogram of a different feature for one hundred samples of a species with corresponding Gaussian estimate overlaid, not a good fit

offset calculations, but there is still no guarantee that features in the test data will remain between negative one and one. A similar method is to scale the features in the training set between zero and one. Even after this initial scaling, questions remain of how to weight the features so that features of greater value are given greater weight.

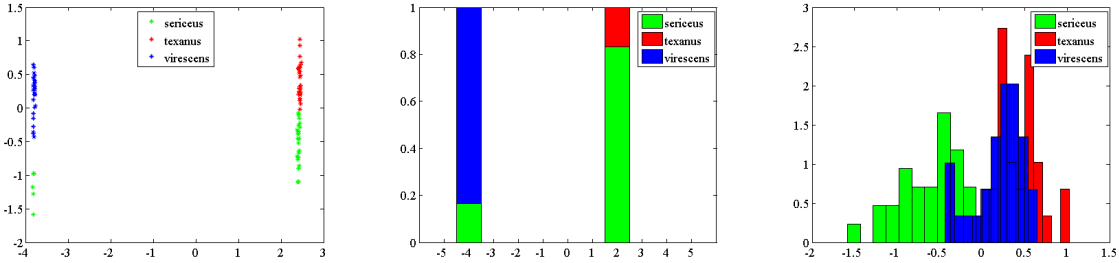
Some feature selectors also determine the proper weighting of the features it selects, since the problem of which features to use is related to how important (how heavy a weight) each feature should be given [41]. This weight vector is returned with the best features to use. The scaling of the features affects principal components analysis, which can affect the distribution of features, which can in turn affect the optimal classification algorithm. For example, Figure 29 shows a set of features selected by the Sparse multinomial logistic regression via Bayesian L1 regularisation with weights derived under a Gaussian assumption, then processed via a principal components analysis. Figure 30 uses weights that have been selected by the Sparse multinomial logistic regression via Bayesian L1 regularisation algorithm also run through principal components [41]. Using estimated Gaussian probability density functions as a classifier seems like a natural choice for the data scaled under a Gaussian assumption (Figure 29), but although the features and samples used are the same the data scaled according the Sparse multinomial logistic regression via Bayesian L1 regularisation heuristic (Figure 29) would have relatively poor performance using the same classifier. In this case, k-nearest neighbor (for small k) and the SVM classifier will perform well under either scenario.

Performance of these methods seemed to vary based on the classification problem. For these classification problems, empirical results did not make the choice clear. Since many of the collected features can be modeled as a normal distribution from empirical results (see Figure 27) it was determined to subtract the mean and divide by the standard



(a) Two most important features (b) First most important feature (c) Second most important feature

Figure 29: PCA run on features scaled to be distributed according to a standard normal distribution



(a) Two most important features (b) First most important feature (c) Second most important feature

Figure 30: PCA run on features scaled by the Sparse multinomial logistic regression via Bayesian L1 regularisation

deviation when using a feature selector that does not incorporate weights. Otherwise, the feature selector algorithm's weights were used.

5.5 Missing Features

Ideally, all of the features will be found for every image, but in practice features may be missing from a sample. For example, a cell may not be found in the wing of one picture, which leads to several missing features. In the training set, it is possible to discard images with missing features. If one has a sufficiently large training set a few missing data samples will not have a large impact. The approach to handle missing values may depend on whether the feature vector with missing values belongs to the training set, or whether it belongs to the test set or is an unlabeled feature vector (ie a sample for which the bee genus or species is unknown).

The simplest method is to discard feature vectors with missing values, since imputing missing values takes some care, modeling, and experimentation. The simplest way to impute is to replace all of the values with a constant, such as zero or one. Unfortunately, this can cause features with a common missing value to cluster together. Therefore, if it is anticipated that a species of bee will often be missing a particular cell, it may be advantageous to impute missing values with a constant since the result may be a cluster of the species without the missing cell and another cluster of the species with the missing cell. Techniques such as the k-nearest neighbor classifier (discussed in section 5.6) may be able to take advantage of such clusters.

If the features have been scaled so that they are zero mean, imputing with a zero is equivalent to imputing with the expected value of the feature, which is often a sensible

thing to do. In training data where the label is known, it may be advantageous to impute missing values with the mean of the available values within the class the feature vector is represents. If the label is unknown, this extra information does not exist and therefore the mean of the training set may be taking into account any known priors. Since the training and test data are coupled, it may make the most sense to impute the training data in the same manner that the test data is imputed, although this depends partially on the classifier and how regularly features are expected to be missing.

Another scheme to impute missing features is using the Singular Value Decomposition (SVD) [36]. Features can be imputed two ways using the SVD. The first way is to use the SVD of only the clean training set (where all feature vectors are full). The second way is to use the whole training set (which may contain missing values). This requires an iterative solution that can converge to reasonable solutions, although it does have the disadvantage of penalizing labels that are underrepresented.

A different heuristic is to impute using the k-nearest neighbor feature vector [36]. The idea is that the values of the missing features will most likely be close to the values found in complete feature vectors whose other features most closely match the available values in the incomplete feature vector [50, 51, 52]. Imputing based on the SVD or the k-nearest neighbor is of course computationally more intense, which means that in the early stages of a project one may benefit from using a simpler method for more rapid evaluation of new ideas and features.

Of course measurements of the actual features as they exist are best. It is generally best to re-photograph a sample if the original image has hard-to-find features. Without the original specimen, known features in the image can be used to extrapolate regions of interest, so that specialized processing on those regions can be used to find the missing

features. Another alternative to imputing in the test and implementation phase is to use the subset of the feature vector that is complete with corresponding features in the training set. For many classifiers this means it is necessary to train the classifier on all different possible combinations of features. This can be computationally prohibitive.

Initially this project had numerous missing features, since all of the cells were found irregularly. With the change in image quality, the number of missing features was reduced significantly. Imputing with a constant value of zero was used since it is easy and provided fast and good results. In this case, since the features were scaled to be distributed according to a standard normal distribution, zero was also the mean value. No significant difference between the different heuristics was observed in the test stage, so replacing missing values with zero was used since it requires little computation.

5.6 Classification Schemes

While developing the system with only a small feature set that changed rapidly, the k-nearest neighbor classifier was of great use for evaluation of changes made in the other parts of the system. This is because the k-nearest neighbor requires no learning phase and is relatively accurate and quick. However, in a finished system where no new learning phases are to occur, the system can be significantly sped up without much change in performance by using a model designed from a prior learning phase.

A number of classifiers were explored when developing, including decision trees with bootstrap aggregation (tree bagger) [53, 54, 55], Maximum-Likelihood using estimated

multivariate normal distributions probability density functions with and without class-wise principal components analysis [36, 56, 57, 58, 59], a Bayes Classifier [60, 61], k-nearest neighbor [62, 36], and support vector machines [63, 64, 65, 66, 67, 36]. Many of these classifiers require significant computation in the learning phase, but require much less computation in the classifying phase. Which means they are not feasible in the development stage, but used in the testing and deployment stage are quite feasible. Optionally bagging or boosting techniques can be used in conjunction with a classifier to improve performance [68, 69, 70, 71, 72, 73].

Each classifier works on a unique set of heuristics and will work well when certain assumptions about the data hold. Decision tree classifiers ask a binary question about one or more of the features and based on the answer branch out to a new question. When all necessary questions are answered a leaf node is reached which represents the class of the classification. Maximum-Likelihood classifiers assume some known distribution (such as a multivariate normal distribution or a Laplacian distribution) and classify based on the class with the highest probability density function at a given point in feature space. Maximum A-Posteriori classifiers work off of the same principle, but consider prior probabilities or the likelihood of an event happening. The k-nearest neighbor classifier finds the k-closest neighbors in feature space and takes a majority vote for the classification decision - which is quite robust when training points cluster, but requires many computations per classification. The support vector machine seeks to find separating hyperplanes between classes, which is easy to compute and works well when boundaries between classes are linear.

5.7 Validation

There is an unfortunate tendency in classification to overfit the data in an effort to "get really good results." In order to avoid this, a strict separation between training data and test data was kept. After selecting features to use from the training data, leave-one-out cross-validation was applied, followed by true validation using feature vectors from the test data against the model. A wide discrepancy was found between these two methods in the initial dataset, initial results showed a gap of 20% error between the cross-validation and the true validation. Having test data separate was invaluable in predicting how well the model worked.

In this way, keeping the test data separate from the training data gives an accurate representation of how much over-fitting may have occurred and provides a more accurate portrayal of how the system will do at identifying the bee species. Typically cross-validation and validation are used hand-in-hand, because using cross-validation alone can lead to overfitting, and if the validation results are used without the cross-validation results the algorithm can be tweaked until it works best on the test set. Ultimately it is desirable for the training data's cross-validation results to match well with the validation results. A poor match means that likely the model overfits the training data.

After cross-validation and true validation, attempts can be made to further simplify the model and thereby reduce any over-fitting that may have occurred. Care must be taken since adjustments can lead to simply over-fitting a new model containing both the training and the test data.

5.8 Interpreting Classification Results

There are a number of ways to interpret and visualize classification results such as the probability of correct identification, visualizing where mis-classifications occur using a confusion matrix [74], using a dendrogram to measure how related each taxon is [75], and projecting the samples into a subspace for easy visualization (often using principal components analysis).

The most straightforward way to visualize the results is to look at how well the classification system does in terms of correctly identifying labeled samples of different taxa. A slightly more elaborate method is to interpret results by means of the confusion matrix, see Figures 31 and 32. The confusion matrix is formed during cross-validation or during validation using a model. In this work, the data was split into two sets: training data to be used to create a model for classifying the samples and test data to be used to test the model created using the training data. Figures 31 and 32 are sample results in the form of graphical displays of the confusion matrix from cross-validation of the model using the training data (Figure 31) and validation of the model using the test data (Figure 32). Ideally, all of the data should lie on the diagonal of the matrix or image. Values not along the diagonal represent instances of "confusion" where the model predicts the wrong class for a given sample.

A dendrogram can be created by comparing the distances between the centroids of each class, giving a measure of how closely related the species are. The shorter the lines connecting two classes in the dendrogram, the more closely related the classes are. In this case, it is how closely related the species are in terms of wing morphology with a given set of features, see Figure 33. This interpretation neglects the variance, using only the mean

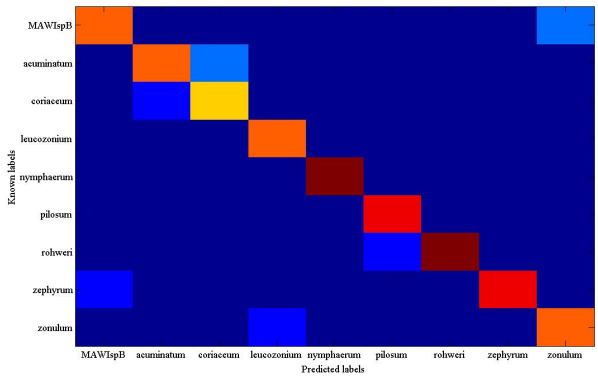


Figure 31: Confusion matrix from Leave-One-Out Cross-Validation of the training data

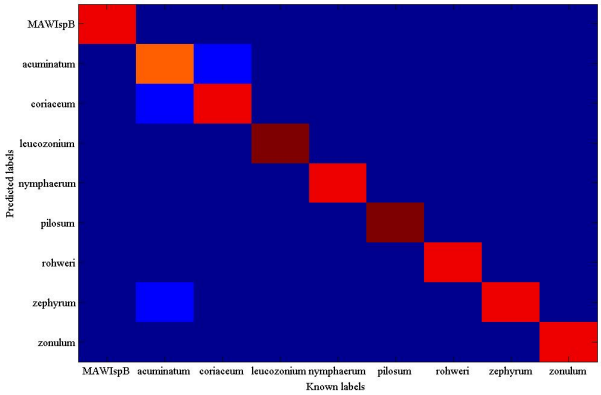


Figure 32: Confusion matrix from validation of the model using test data

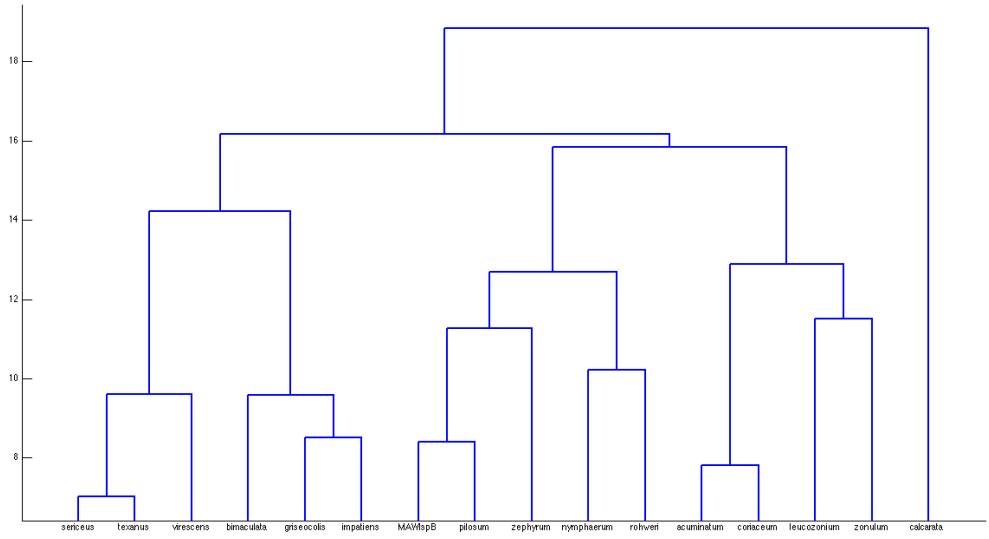


Figure 33: Dendrogram of sixteen species

within the species and so it does not paint a completely clear picture. Although, the dendrogram of Figure 33 was created based on wing morphological features, the species within a genus clumped together nicely. Specifically the species *sericeus*, *texanus*, and *virescens* are green bees belonging to the genus *Agapostemon*, the species *bimaculata*, *griseocolis*, and *impatiens* are larger bees belonging to the genus *Bombus*, the species *calcarata* belongs to the genus *Ceratina*, and the other nine bees are small black bees belonging to the genus *Lasioglossum*. Looking at the family level reveals that *Ceratina* is more closely related evolutionarily to *Bombus* than the others, which is not shown by the dendrogram based on wing morphology.

A dendrogram can also be called a cladogram, which has been used in the past as a phylogenetic tree. Cladograms are commonly used to show ancestral relations between

species. Originally, such interspecies relationships were determined using morphological features (which wing venation features is an example of), however, today genetic sequencing of data and computational phylogenetics are most commonly used. Although, the use of morphological features can be used to identify leaves in the tree of the dendrogram as shown by any successful species identification system, it is too much to say that it also will correctly identify and place the branches. Thus morphological features should only be used as a provisional hypothesis to where a species may fit on a cladogram, as morphological and molecular features can disagree [76].

For this reason it is desirable that a measure of differentiability be used. This may be inferred from the results of cross-validation during the learning stage or by setting aside some test samples for validation to use after the learning phase is complete, or by using some kind of measure to see how closely related two subjects are. Other systems do this with their features after projecting the feature space into a lower dimensional space, using methods such as running a principal components analysis or using a kernel discriminant analysis (KDA) to project the features into a two-dimensional subspace so that an operator can view clusters of species and see how closely they are related or how well they are differentiated in the two most important dimensions [3, 16]. This measure is purely visual and mostly qualitative. Sometimes in these plots, taxon group samples are intermingled, other times there is a great distance between the taxon groups. This is similar to Figures 34 and 36, which are the results of using the two largest principal components to plot the data. Scree plots indicate how much of the variability is contained in each feature. Ideally, to plot the species in two dimensions the variability should be contained in the first two features, as in Figure 35. Often for large data sets more of the variability will be contained in features beyond the first two, even after PCA, see Figure

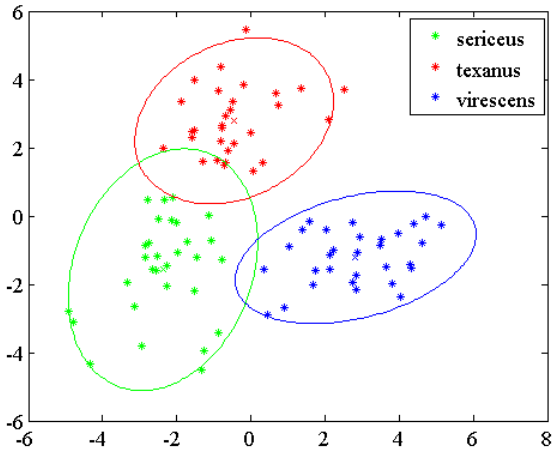


Figure 34: Three species plotted with the top two features resulting from principal components analysis

37.

5.9 Limitations

Accurate classification requires that the classes can be distinctly separated in feature space. Unfortunately, this is not always the case and is certainly not the case if a poor set of features is chosen. When classes overlap in feature space there will be some contribution of error as a result. While adding features tends to increase the accuracy in cross-validation it can result in over-fitting. A good rule of thumb is to use fewer features than there are specimens of each class in the training set [35]. Features that work well for discriminating one set of bees may not help in discriminating another set of bees, this is why it is helpful to have a flexible system that can adjust and adapt as its training data is expanded to encompass more bee species.

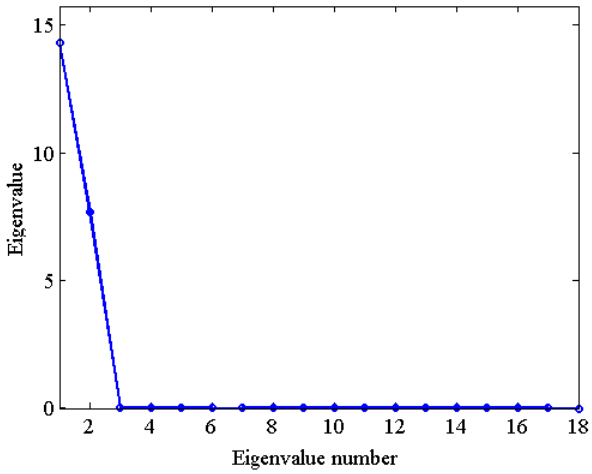


Figure 35: Scree plot of three species

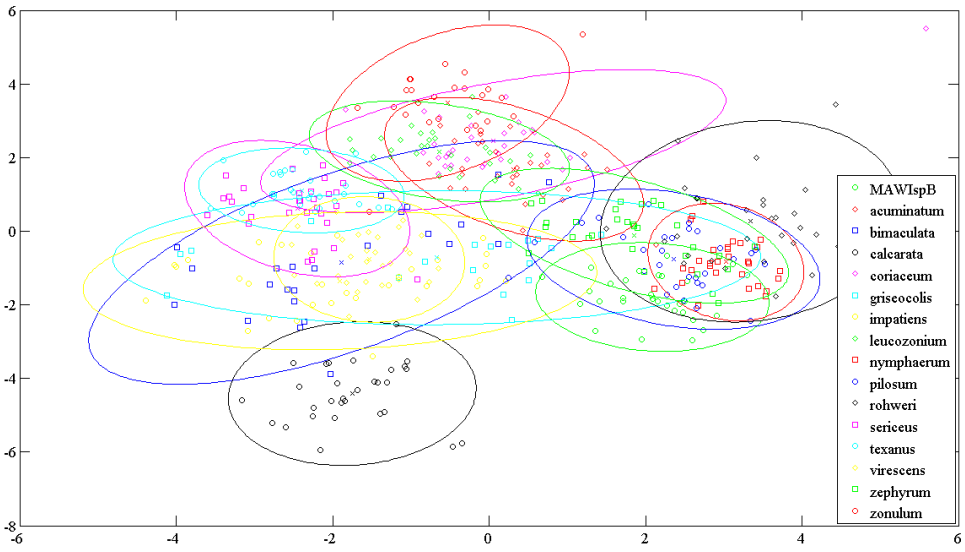


Figure 36: Sixteen species plotted with the top two features resulting from principal components analysis

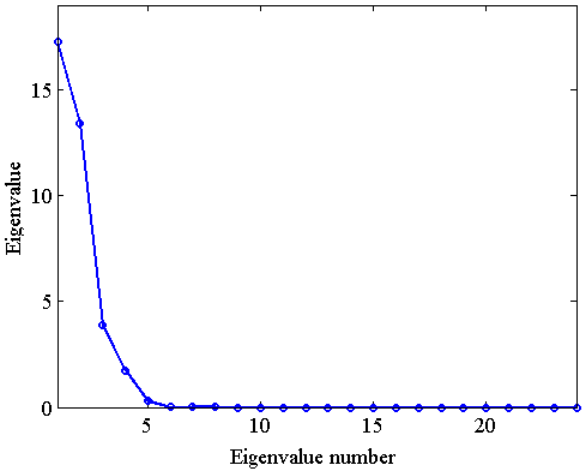


Figure 37: Scree plot of sixteen species

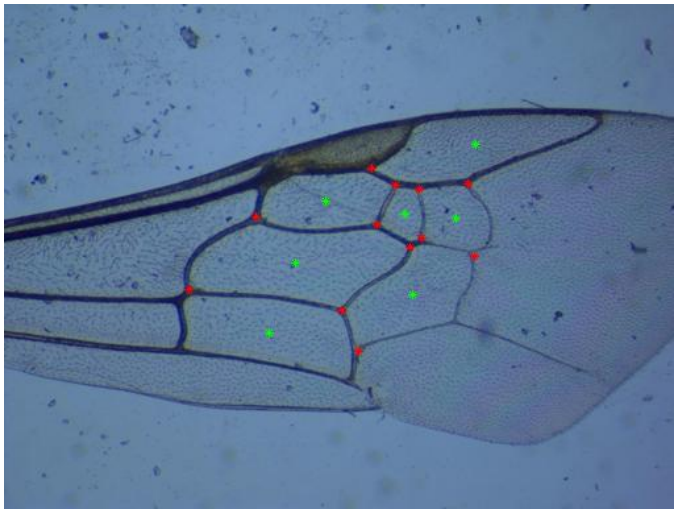


Figure 38: Junctions and cell centroids form excellent key points

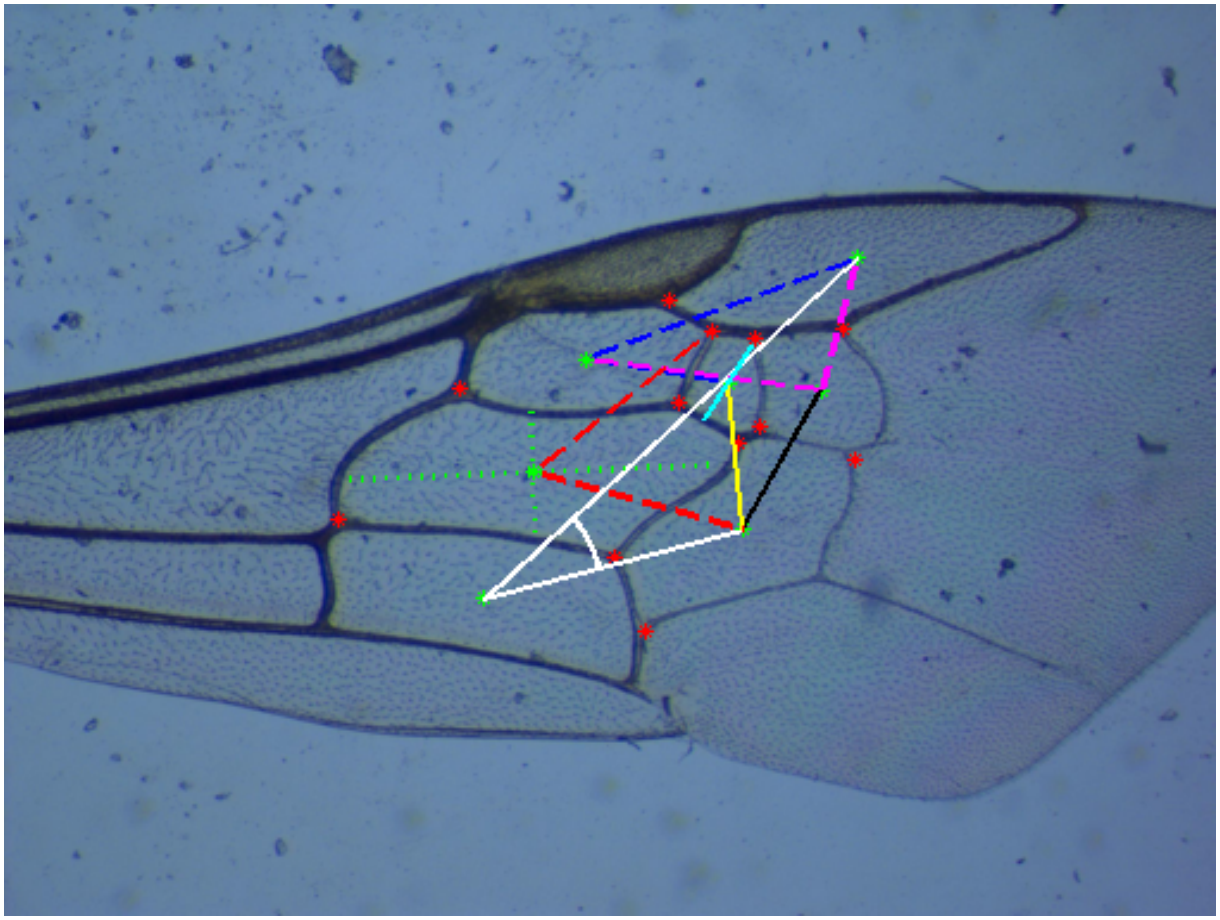


Figure 39: Representative features, dashed lines represent ratios between two lines, single solid lines are absolute distance, a pair of solid lines with a semicircle between them represents an angle

Chapter 6

Reliability Measures

6.1 Reliability Measures

Having an estimate of how reliably each specimen has been identified is important and useful to entomologists. There are only two reasons for a specimen to be misclassified: it is either confused with another class in the training data or it belongs to a class not found in the training data. Reliability measures should therefore convey these ideas. Misclassifications are likely to occur when features have not been robustly extracted, the class of the specimen wasn't incorporated into the training set, or even simply that the feature vector used does not provide a clean enough separation between the classes under the classifier in use.

Entomologists could take samples that were classified with a low reliability and have them examined by an expert in the area of bee identification. Alternatively, the system could report several top class choices using associated reliability measures and provide the user with specific characteristics to look for that may easily differentiate the species perhaps using features not found in the wing or given to the system such as the length of the tongue or color of the abdomen.

New species, which do not match a known class in the dataset, could be discerned as

not fitting well in the expected classification scheme, which would accelerate their identification by taxonomists. A reliability measure could also be used to provide excellent quality results for studies where a higher level of accuracy is required. The reliability measures should provide insight into how likely it is that the test specimen has been classified correctly given the training set ("confusion" measure) and how much of an outlier the specimen is to the other specimens in the dataset ("outlier" measure).

The problem is how best to create reliability measures. Things like feature distributions, scaling, and processing all affect the best way to create a reliability measure. When a certain type of classifier works well for the problem of classification, related reliability measures derived under the assumptions of the classifier should also perform well.

6.2 Multivariate Normal Distribution Measures

Perhaps the simplest case is when it is known that the features are distributed according to a multivariate Gaussian distribution. In this case, the optimal classifier, in a Bayesian sense, is to compare the probability density functions of each class at the test sample's point in feature space and choose the largest [36, 77]. Since the distributions are rarely known, their parameters (ie the mean and standard deviation) must first be estimated using samples from the training set [78]. For a test sample, it is optimal to evaluate its point in feature space across all of the probability densities and choose the largest. Further use can be made from the densities though, by comparing the probability densities across classes at the given point. In the case where the classification occurred at a point where the distributions overlap significantly the decision would be less reliable,

even if it were the best decision given the features. By creating a ratio of the probability density function at the point of the query of the chosen class over the sum of all of the probability density functions evaluated at the same point, a measure similar to that used by the likelihood ratio test can be created [78]. The range of this ratio is between zero and one where one is a reliable decision and zero is an unreliable decision.

If the probability densities are too small, several standard deviations from the class means, then the species may not yet be incorporated into the system. Such an idea can be used to help identify outliers in the data, that may not belong or do not fit the model. Since the spread of the data is known in terms of the covariance matrix, the distance from the chosen class' centroid to the specimen in feature space is measured using Mahalanobis distance [36, 79]. Small distances give confidence that the point is not an outlier while large distances indicate it is more likely to be an outlier. In the one-dimensional case this boils down to the number of standard deviations the test point is from the mean, which intuitively is a good measure.

More explicitly for n classes and d features per specimen fitting the multivariate normal distribution model means there are n multivariate normal distributions in d-dimensional space. The classifier's optimal decision of the class, $\hat{c}$, given a sample point, x, can be written as

$$\hat{c} = \arg\max_k \frac{\mathcal{N}(\mu_k, \Sigma_k)(x)}{\sum_{\ell=1}^{n} \mathcal{N}(\mu_\ell, \Sigma_\ell)(x)} \ \forall k \in \{1, 2, ..., n\}.$$

The "confusion" reliability measure is then:

$$R_c = \frac{\mathcal{N}(\mu_{\hat{c}}, \Sigma_{\hat{c}})(x)}{\sum_{\ell=1}^{n} \mathcal{N}(\mu_\ell, \Sigma_\ell)(x)}$$

and the "outlier" reliability measure is

$$R_o = \sqrt{(x - \mu_{\hat{c}})^T \Sigma_{\hat{c}}^{-1} (x - \mu_{\hat{c}})},$$

where μ_k and Σ_k are estimated from the training set using standard techniques [78].

Extra information provided by these reliability measures can be compared to user determined thresholds so that specimens that are potential outliers or potentially confused can be examined more closely. Although this set of reliability measures was determined for a multivariate normal distribution, it can be generalized such that a measure can be found to match the assumed underlying continuous distribution. If the model is good, it should provide an insightful measure of reliability.

These reliability measures were applied to the species and the two plotted features from Figure 34 to provide the reader with an intuitive understanding of these measures. First all three species went through leave-one-out cross-validation (loocv) to produce the reliability measures for the ones correctly and incorrectly classified, then one species was removed from the training set and classified using the other two species as training data to simulate an unknown species. The loocv results were 97.78% correct species classification, naturally the simulated unknown species resulted in 0% correctly classified; the reliability measures are recorded in Table 2.

This process was then repeated for the species in Figure 36, a large classification problem with few features. The two plotted features were used as well as an additional eight features which yielded 86.68% correct classification, the reliability measures are recorded in Table 3. Observe the results are in part a function of the number of classes and how densely packed the classes are. The species classified correctly have a mean value near one for the "confusion" measure and a relatively small mean value for the

	type	min	max	mean	median
Correct	R_c	0.7273	1.0000	0.9790	0.9993
Incorrect in training	R_c	0.5458	0.5700	0.5579	0.5579
Incorrect not in training	R_c	0.5323	1.0000	0.8898	0.9738
Correct	R_o	0.1983	2.7437	1.3068	1.2934
Incorrect in training	R_o	2.5474	2.7945	2.6710	2.6710
Incorrect not in training	R_o	2.6896	6.7443	5.0291	5.1171

Table 2: Reliability measures for three species under the multivariate normal distribution assumption

	type	min	max	mean	median
Correct	R_c	0.4844	1.0000	0.9654	0.9999
Incorrect in training	R_c	0.4047	1.0000	0.8609	0.9677
Incorrect not in training	R_c	0.5103	1.0000	0.8733	0.9986
Correct	R_o	1.6130	10.1605	3.7962	3.6456
Incorrect in training	R_o	2.2338	174.6600	7.8018	4.4974
Incorrect not in training	R_o	3.2955	7.9708	5.2308	5.0384

Table 3: Reliability measures for sixteen species under the multivariate normal distribution assumption

"outlier" measure. The simulated "unknown" species has a higher mean than those classified correctly, which means when classifying using the system based solely on wing venation many of these specimens would be set aside to be examined more closely, given a proper threshold set in the system. An expert or classification guide could then be used to make a final determination, which is just how the reliability measures should perform.

6.3 Measures for k-nearest neighbor

The k-nearest neighbor algorithm uses the nearest k neighbors in feature space to the test point and takes a majority vote to determine the class of the query [36]. A useful measure of reliability would then be how close the vote is. For example, if one hundred votes are taken between two parties and the first gets 51 votes while the second gets 49 votes, the decision is much less decisive than if one party received 99 votes and the second received one vote. Many k-nearest neighbor algorithms weight neighbors in the vote according to the inverse of their distance, which can also be taken into account in this reliability measure. The k-nearest neighbor algorithm approaches the Bayes Error Rate as the number of training points approaches infinity, because it in effect creates a histogram at any given point. As the number of training points becomes large, the histogram approximates the underlying probability density function [36]. Since the actual probability density function is approximated, with many training points and a large k using the closeness of the vote as a reliability measure will give good results in terms of the likelihood the bee belongs to another group. The distance to these neighbors can be used to determine whether the sample point is an outlier.

The ideas used by the k-nearest neighbor measure needs a large k for the numbers to be meaningful. Unfortunately, empirically, large k's do not seem to classify this dataset, probably due to the small number of sample points. However, the reliability measure will give have a small range of values for small k. The k-nearest neighbor (knn) algorithm can be modified so that it still classifies well and yet gives meaningful results for the reliability measure, this modification is the nearest k-class-neighbors (nkcn) classifier. Traditionally, the k-nearest neighbor takes a majority vote of the neighbors within

a minimum sized hypersphere that contains k neighbors. Instead, the nkcn takes a majority vote of the neighbors within a minimum sized hypersphere, which contains k neighbors of any one class. It can be written as:

1. Compute distance (Euclidean in example) from test point to training points
2. Order distances by class from smallest to largest.
3. Calculate the average distance for the first k-neighbors of each class.
4. Classify as the class with the smallest average distance for the first k neighbors.

This will provide a variable number of neighbors in the hypersphere. Following the proof for the k-nearest neighbor algorithm in [36], it can be verified that the nkcn classifier will reach the Bayes Error Rate under the same conditions as knn, which inspires some confidence in this classifier. Define the ratio of the number in the winning majority divided by the total number of neighbors in the hypersphere, to be a measure of reliability of the decision. The average distance between the query point and the k neighbors of the given class can be used to give a measure of how much of an outlier the query is.

More explicitly, for n classes and f features per specimen to derive the classifiers decision, $\hat{c}$, the distances from the test point, x, to the points in the training data, t, for each class are first calculated

$$d_\ell = |x - t_\ell|_2 \ \forall \ell \in \{1, 2, ..., n\}.$$

the distances to neighbors are sorted smallest to largest for each class

$$d_\ell = \text{sort}(d_\ell) \ \forall \ell \in \{1, 2, ..., n\}.$$

The average distance of the first k-neighbors of each class is used to choose the class estimate $\hat{c}$

$$\hat{c} = \arg\min_{\ell} \frac{d_\ell(1:k)}{k} \ \forall \ell \in \{1, 2, ..., n\}.$$

The "confusion" metric is then defined to be the number of neighbors in the minimum-sized hypersphere containing k-class-neighbors belonging to the class estimate ($\hat{c}$) divided by the number of neighbors in the same hypersphere, that is

$$R_c = \frac{\sum_{i=1}^{\#\text{ of points in } \hat{c}} \mathbf{1}_{\{d_{\hat{c}}(i) \leq d_{\hat{c}}(k)\}}}{\sum_{j=1}^{\#\text{ of points in training}} \mathbf{1}_{\{|x - t(j)|_2 \leq d_{\hat{c}}(k)\}}}.$$

The "outlier" reliability measure is the average distance to the first k neighbors in the class estimate $\hat{c}$

$$R_o = \frac{d_{\hat{c}}(1:k)}{k}.$$

Assuming no two neighbors are at the same distance (which is true with high probability when the distribution is continuous), it is possible to bound this ratio between zero and one where the numerator will be a constant k and the denominator will be between k and $k + (k-1) \times (\text{total number of classes} - 1)$. One is a clear decision and zero is an ambiguous decision for R_c. The larger R_o is, the more likely it is an outlier. To set a threshold based on this value requires some assumptions based on the spread of the data.

To make a connection between the reliability measures for the multivariate normal distribution and the reliability measures for the nkcn classifier. Note that as the total number of points in the training set approaches infinity, the nkcn classifier perfectly approximates the probability density function. For large enough k and an underlying normal distribution R_c is the same under either assumption. If the Mahalanobis distance were used in place of average Euclidean distance to the k-th neighbor, R_o would be the

	type	min	max	mean	median
Correct	R_c	0.6000	1.0000	0.9839	1.0000
Incorrect in training	R_c	0.4286	0.7500	0.5929	0.6000
Incorrect not in training	R_c	0.5000	1.0000	0.9000	1.0000
Correct	R_o	0.1442	1.3387	0.4399	0.4021
Incorrect in training	R_o	0.7608	1.0615	0.8647	0.7720
Incorrect not in training	R_o	1.9726	6.4566	4.3656	4.4821

Table 4: Reliability measures for three species using the nearest k-class-neighbors algorithm with k=3

same for an underlying multivariate normal distribution. However, for an unknown mean and covariance matrix, assuming the features are independent and identically distributed the distance to the k-th neighbor for large k is a function of the distance from the mean and the covariance matrix or spread of the data.

Just as the multivariate normal distribution assumption's reliability measures were determined, the same test was applied to the nearest k-class-neighbors. The species and the two plotted features from Figure 34 were once again used to produce the reliability measures recorded in Table 4, the result of the loocv were 96.67% correct species classification. Next, the sixteen species in Figure 36 using the two plotted features and an additional twenty features were used to create the reliability measures recorded in Table 5. In this case the loocv classified 90.83% of the species correctly. Once again, the "unknown" species and misclassified specimens have a higher average outlier measure and the "confusion" measure is lower on the misclassified specimens.

	type	min	max	mean	median
Correct	R_c	0.2000	1.0000	0.9160	1.0000
Incorrect in training	R_c	0.2727	1.0000	0.5992	0.6000
Incorrect not in training	R_c	0.5000	1.0000	0.8789	1.0000
Correct	R_o	4.1223	24.4030	8.3361	7.9703
Incorrect in training	R_o	6.2524	64.3817	11.1200	9.0212
Incorrect not in training	R_o	7.8332	17.0182	11.3599	11.1924

Table 5: Reliability measures for sixteen species using the nearest k-class-neighbors algorithm with k=3

6.4 Summary

These reliability measures are helpful, but they certainly are not perfect. The "confusion" measurement of reliability doesn't work well when the decision of which class to choose seems clear in the Bayesian sense, but the choice is wrong due to instances of a class within areas of low probability. Furthermore, although the "outlier" measurement of reliability should work well for discovering species that are distantly related in feature space to the species in the training set, it will not work as well for discovering species that are closely related to other species in the feature space. Considering the nature of the information available at the classifier level this is not unreasonable.

When classifying if two or more classes have high probability relative to the class with highest probability, it may be helpful to list the top choices with their associated reliability measures as is often done in reporting search engine results. Additional information could then be used from a classification guide on how to discriminate such specimens using more than just the wing features. These could be given to the user to make the best possible decision.

Chapter 7

Results

7.1 Final Setup

The results using MOBS on gathered test sets representing a large number of genera and species were promising. Researchers in entomology at the University of Wisconsin-Madison gathered many bees, from which images of wing venation were collected. The ABIS test set provided with their software was also used, which is the source of the genera *Megachile* and *Osmia* [3]. The combined number of unique bee wing images was well over one thousand images. However, many of the different classes lacked a sufficient number of samples to be reliably classified.

The dataset contained over one thousand images of bee wings in which the pre-processing algorithm was able to find all "important" cells well over 90% of the time. Correct labeling of cells when all "important" cells were found occurred approximately 90% of the time.

The classification was accomplished in a hierarchical fashion [37]. More general to more specific classes were distinguished as follows:

1. genus

2. species

3. subspecies or gender

The classification process was carried out to the furthest extent from the ordered list above for all classes containing a diversity of subclasses with a minimum of ten members per class. The dataset contained six unique genera. Each of the genera in the set had between one and nine unique species. Two species had two unique subpecies labeled. A few of the species had a large representation of labeled genders. Appendix B describes the results of training and testing along with the the specifics of which bees were used.

The dataset for each classification problem was divided with two thirds in the training set and one third of the data in the test set. An equally likely assumption for each class of bees was desired so as to not bias the results during classification. For this purpose, at the species and subspecies level an effort was made to keep the number of bees in each class approximately the same. Sparse multinomial logistic regression via Bayesian L1 regularisation was applied to the training set to find a small subset of features for classification, fewer than the number of members in the smallest class (a maximum of twenty). The weights from the algorithm were neglected in favor of principal components analysis, which was then used to reduce noise and find dimensions of highest variance. Leave-one-out cross-validation was then applied to the training set with the 3rd nearest neighbor classifier applied. The results were validated using the training information (selected features, PCA loadings, training point locations). Refer to Appendix B for detailed results.

7.2 Discussion of Results

The results of each classification problem seemed promising although it is clearly not perfect. Genera were classifed correctly in the high nineties to one hundred percent of the time in both validation and cross-validation, results depended on the choice of training set. Species were classified correctly over ninety percent of the time. Subspecies and gender were classified in the eighties to nineties percent correct. Specific results including the number of principal components used, the percent correct in training and in test data as well as visualization of confusion matrices, meaningful principal components, and dendrograms are in Appendix B.

Classification in problems with smaller training sets performed more poorly, which leads me to expect that better performance for classifying some species, the subspecies, and the genders could be improved with more training data. Other likely causes of failure include overlap in features of the bees, poor segmentation, inaccurate feature identification, and poor image quality.

Chapter 8

Conclusion

8.1 Conclusion

Pollination is an important industry that has typically been managed primarily using the European honeybee. Colony Collapse Disorder has encouraged researchers to explore other pollination options. Unfortunately, it is difficult to assess which types of bees might contribute most to the pollination needs. It is particularly difficult to monitor bee species' population dynamics, in large part due to the difficulty of species identification. There are over sixteen thousand different known species of bees with potentially many undiscovered species, which further complicates this problem [5]. Typically specimens to be identified are sent to an expert who labels the bees by hand using his/her expert domain knowledge. Recently, several automated systems have been developed to aid the process of identification using morphometric techniques including ABIS, DrawWing, and tpsdig2 [3, 14, 15].

This work has extended previous automated species identification efforts by creating a system, which uses manually acquired images for automated image preprocessing, feature extraction, feature selection, and classification for identification of bee genera, species, subspecies, and gender.

The data acquisition methodology supports collection of samples from both live bees

as well as museum specimens. Image preprocessing uses a combination of standard image processing techniques to identify wing cells. Once the wing cells have been found and labeled, key points such as the vein junctions are identified. Features are derived from these points, which may first be fit to templates based on an affine or procrustes transformation.

Feature classification including considerations and decisions taken in selecting the number of features, which features to use, feature scaling, imputing, and classification algorithms were discussed. Visualizations of results were also provided with interpretation.

Reliability measures are set forth with theoretical backing for both a maximum likelihood classifier under a multivariate normal distribution assumption and a k-nearest neighbor (or nearest k-class-neighbors) classification scheme. These reliability measures are shown to be helpful in identifying specimens that have likely been confused between classes as well as specimens whose class does not fit into the classification's training data, under the assumption that the class they belong to have a different mean and/or spread in feature space from the known classes.

Unique to MOBS are the automated feature selection process, which automatically chooses a set of plausibly useful features from a larger set. This technique reduces overfitting. When new species are introduced it encourages helps choose a set of features to separate the classes well for the new classification problem, which a fixed set of features may not naturally separate new classes well. The measures of reliability are yet another unique contribution that can be used to identify possibly confused specimen as well as assist in the discovery of new classes of genera, species, and subspecies. A report of results is given with leave-one-out cross-validation and also validation.

Results from a dataset containing several genera and species of bees suggest that automated techniques for the identification of bee genera, species, subspecies, and gender, while not perfect, may be useful. Such automated techniques could be used in research studies for identification of bees in lieu of identification by an expert or use of genetic material.

8.2 Further Work

One of the greatest unfinished works is to gather data on more or all bee genera and species, so that the tool will be practically usable by researchers around the world. This is a large task with over four hundred unique genera and over sixteen thousand known species, likely with many unknown species and genera yet be discovered and catalogued. This work is likely to take years, but with the results found at this point it seems probable that such efforts would be well rewarded.

New challenges may arise as this problem is scaled to accommodate a new larger collection of bees. The system will need to continue to adapt as the training set and classification problem changes. Among such changes may be additional features, different classifiers, and support for bees with further reduced wing venation, such as stingless bees with only the first submarginal cell present [15].

The Jenoptik ProgRes digital microscope camera used in this study has a software development kit; it would be helpful to integrate the basic cell identification into the camera's software. This would allow a user identifying live bees in the field to know instantly if the picture was usable (or if it would be best to make some fine adjustment and take another picture). This would create feedback in the system, such that the

quality of the cell-finding algorithm would feed directly into the data collection stage, increasing the chance that new data collected would be usable. To increase the number of usable images, intelligent scissors could be used at this point to efficiently select the wings cells as observed by the user [80].

After integrating cell-finding, the next step would be to integrate the classification process in the field. Classification results could be given to the user as an ordered list from most to least likely with associated reliability measures. When the classification of a live bee is in question, information in the system should be available to the user so he/she can gather any necessary extra information to reliably classify the specimen before releasing it. This extra information should be in the form of readily identifiable features that best serve to discriminate the species in question. Additionally, extra information could be given to the system in advance such as geographic location of collection to bias the prior probabilities.

It would, furthermore, be helpful if researchers around the world could upload their pictures to a centralized website which would then identify the bees. This would be especially ideal in that there would be one, easy to access, up-to-date version of the program that could be updated as data for more genera, species, subspecies, and bee genders are added.

Among the long-term goals, it would be useful to have such an identification procedure incorporated into a popular search engine. With the semantic web, one could upload a picture of a wing to a search engine, and the search engine would return results for the identification of the taxonomy of the subject as well as provide links to webpages that host information about the bee. In this way, researchers could be guided to information about bees they are less familiar with, beekeepers could monitor their

bees for hybridization (such as hybridization with Africanized honeybees), and the tool could be usable by people who are interested in knowing what is pollinating their garden or what has infested their shed. Extra information beyond just the image, such as geographic location, time and date found, or subject being pollinated could be included as an "advanced" search which would help increase the reliability.

Recently, it has been noted that many small insects of the order Hymenoptera have stable color structural patterns, and it is thought that this may provide extra information for the classification of bees [81]. One area that may be explored is whether it is more effective and/or easier to gather color information from the wings, in addition to collecting wing venation information. Such information may be particularly helpful among bees, which have weak venation.

Appendix A

Image Processing Overview

An overview of the steps each image goes through prior to feature extraction and classification is shown in Figure 40.

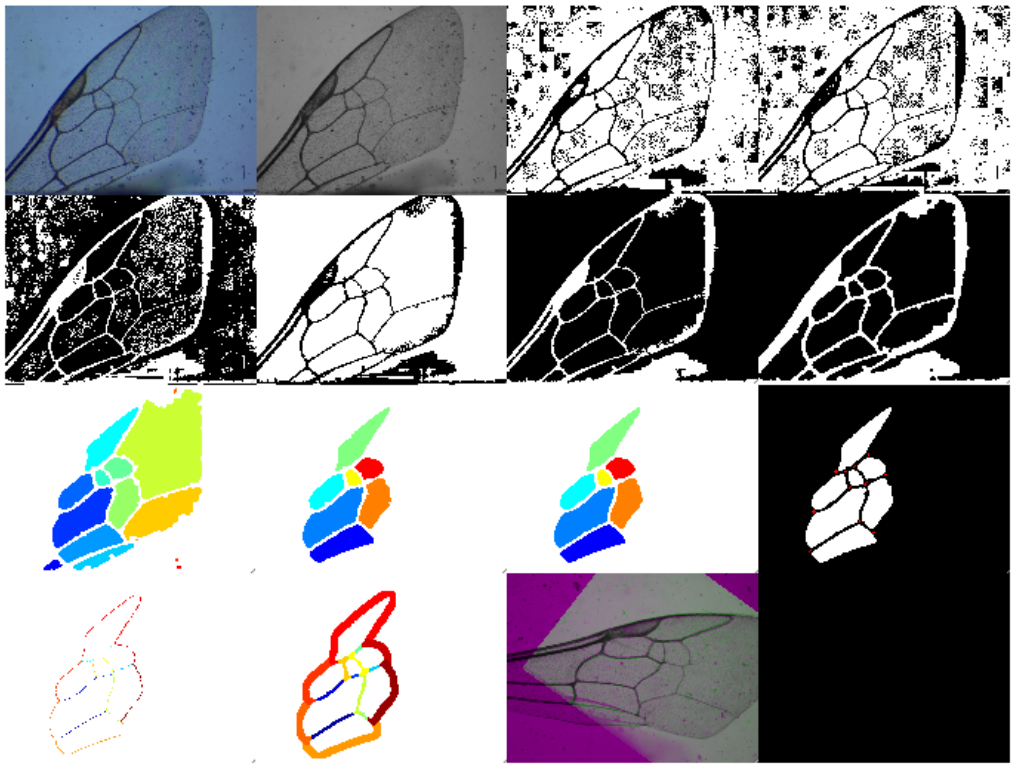


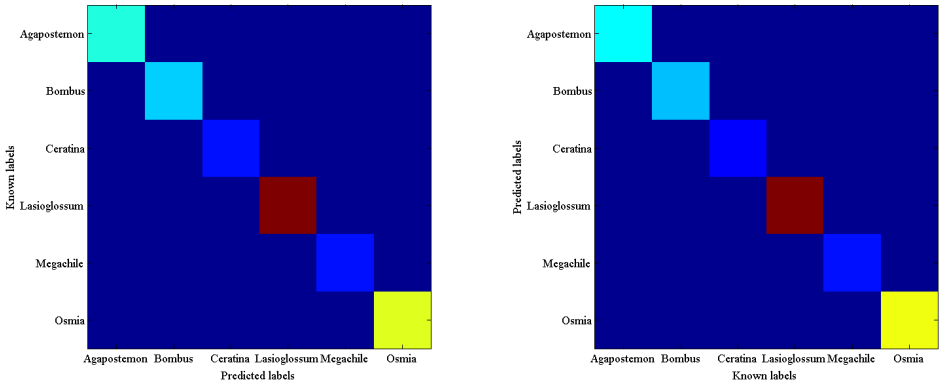
Figure 40: Image processing in brief

Appendix B

Results from Datasets

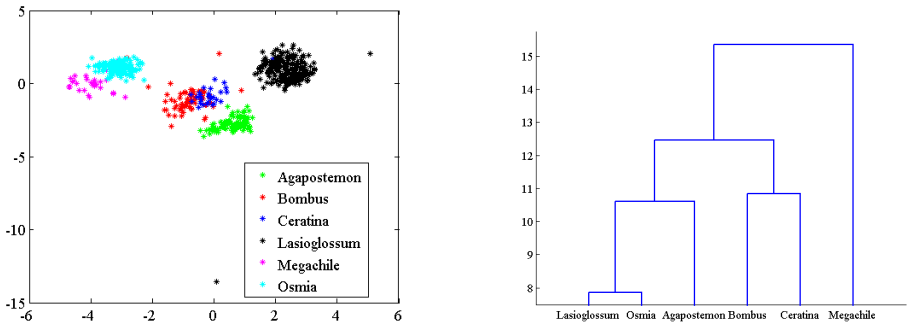
B.1 Genera Classification

First the genera were classified. For genera classification the first five principal components were used. Leave-one-out cross-validation showed 100% of the training specimens were classified correctly while 99.49% of the specimens were correctly classified in validation. Figures 41 and 42 show corresponding confusion matrices, two principal components, and a dendrogram. The different colors along the diagonal of Figure 41 are due to a different number of specimens in each genera of the set. The plotted principal components use the data in both the training set and the test set, as does the dendrogram, see Figure 42. Rerunning the feature selection with a different training set in experimentation gave 100% in both cross-validation and validation, but it is important to note that in reality it is not expected that this or any of the other classifications will truly achieve 100% accuracy in difficult test sets, although they may perform at the level of an expert.



(a) Training (b) Test

Figure 41: Confusion matrices from genus classification



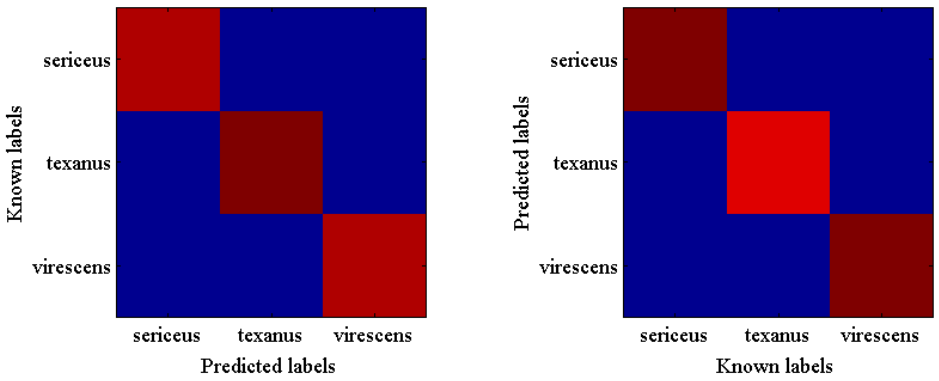
(a) PCA (b) Dendrogram

Figure 42: Genus discrimination visualizations for evaluation

B.2 Species Classification

This section explores the results of several species classifcations fitting in the genera *Agapostemon*, *Bombus*, *Lasioglossum*, and *Osmia*. No classification in the genus *Megachile* was performed because only one to three samples of each species were available. Also no classification within the genus *Ceratina* was performed, as there was only one species within *Ceratina* in the dataset.

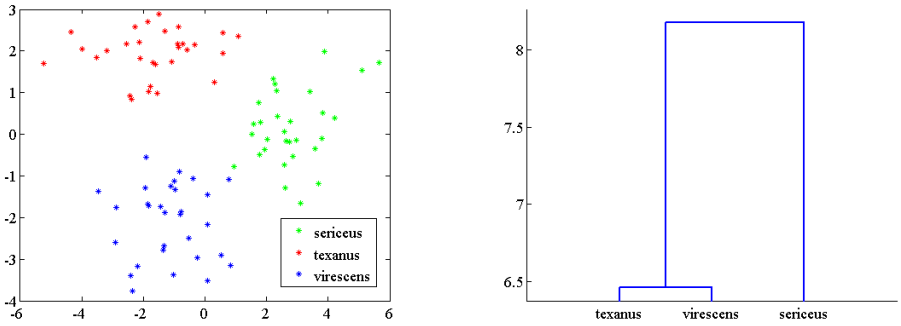
Results from classification of species within the genus *Agapostemon* are shown in Figures 43 and 44. The top eight principal components were used in this classification, which gave 100% of the species identified correctly in leave-one-out cross-validation and 100% in the validation step.



(a) Training (b) Test

Figure 43: Confusion matrices for species in the genus *Agapostemon*

Results from classification of species within the genus *Bombus* are shown in Figures 45 and 46. Results from this set were likely complicated by a lack of sufficient specimens as well as the difficulty of differentiating such specimens [18, 17]. The top three principal



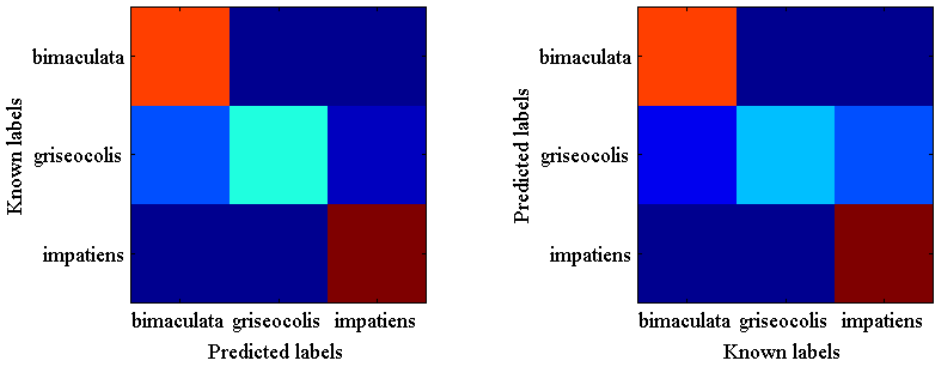
(a) PCA (b) Dendrogram

Figure 44: Species in genus *Agapostemon* discrimination visualizations for evaluation

components were used in this classification, which gave 93.88% of the species identified correctly in leave-one-out cross-validation and 91.67% in the validation step.

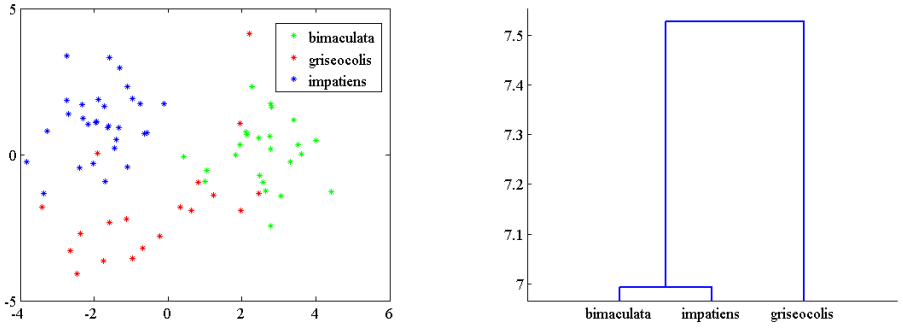
Many species in the genus Lasioglossum are noted to have faint venation [5]. Faint venation was the major contributor of classification error before the use of substage lighting. The results from classification of species within the genus *Lasioglossum* are now quite promising, see Figures 47 and 48. Even with nine closely related yet unique species in this classification problem only the top fifteen principal components were used, which gave 97.18% of the species identified correctly in leave-one-out cross-validation and 97.73% in the validation step.

The results from classification of species within the genus *Osmia* are shown in Figures 49 and 50. In this classification problem the top five principal components were used, which gave 98.51% of the species identified correctly in leave-one-out cross-validation and 96.97% in the validation step.



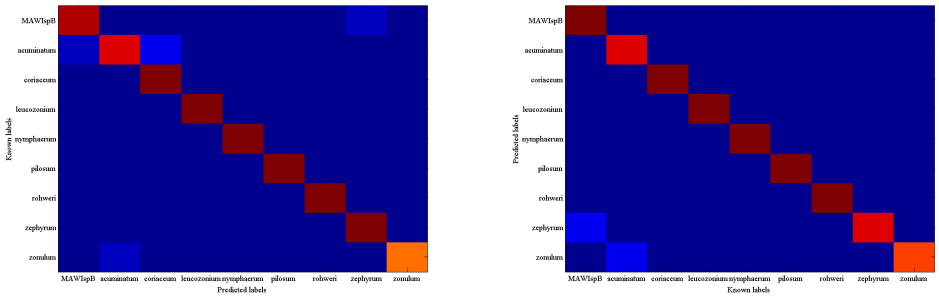
(a) Training (b) Test

Figure 45: Confusion matrices for species in the genus *Bombus*



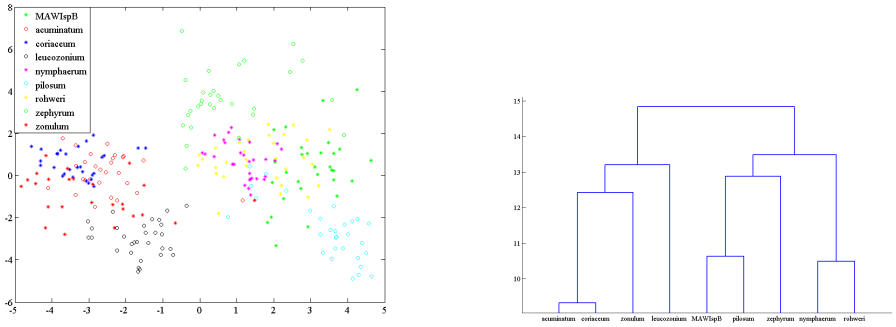
(a) PCA (b) Dendrogram

Figure 46: Species in genus *Bombus* discrimination visualizations for evaluation



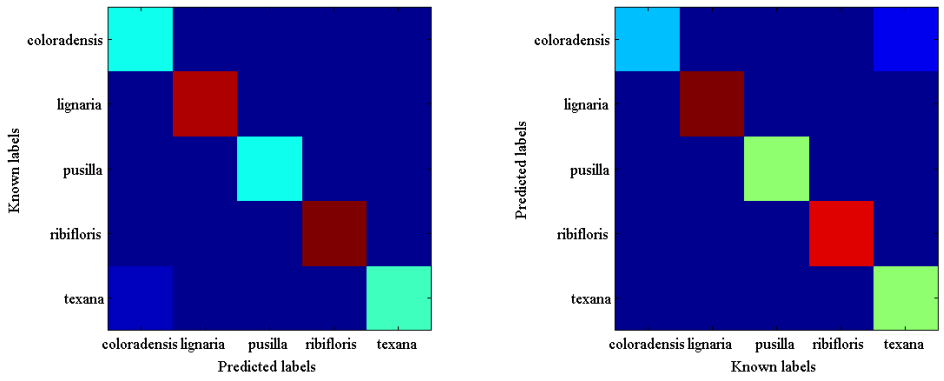
(a) Training (b) Test

Figure 47: Confusion matrices for species in the genus *Lasioglossum*



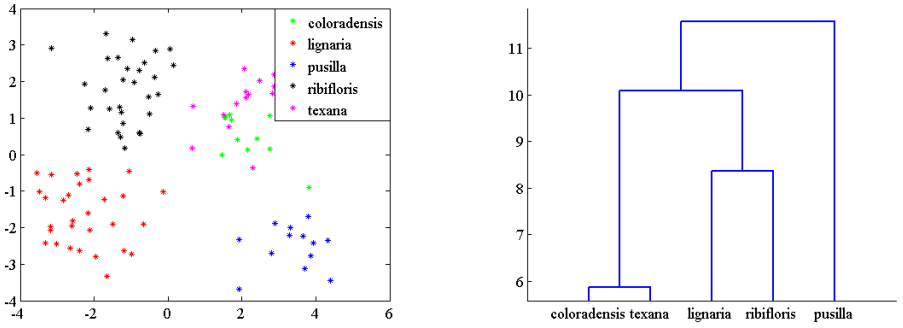
(a) PCA (b) Dendrogram

Figure 48: Species in genus *Lasioglossum* discrimination visualizations for evaluation



(a) Training (b) Test

Figure 49: Confusion matrices for species in the genus *Osmia*



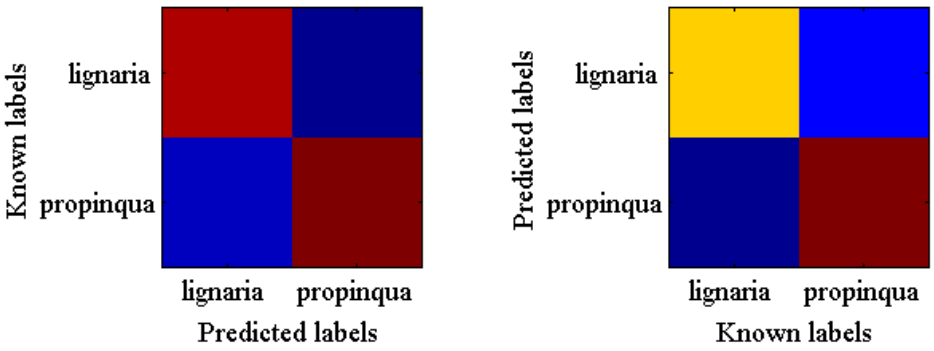
(a) PCA (b) Dendrogram

Figure 50: Species in genus *Osmia* discrimination visualizations for evaluation

B.3 Subspecies Classification

Subspecies classification results in general were worse than the species classification results. This is most likely a result of having fewer specimens of each unique subspecies, although an increase in the difficulty of classification may also contribute.

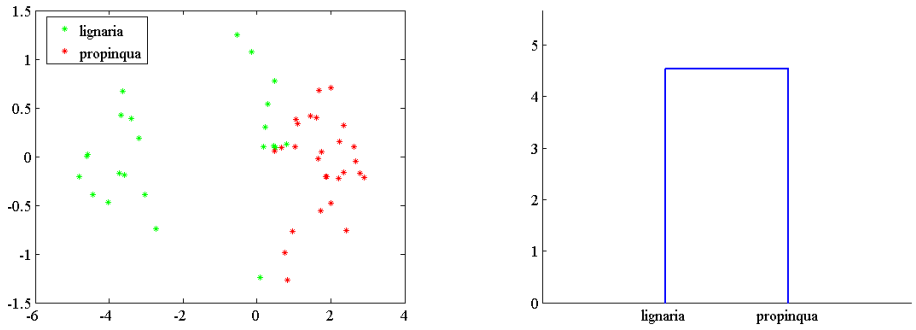
The results from classification of subspecies within *Osmia lignaria* are shown in Figures 51 and 52. In this classification problem the first principal component was used, which gave 97.06% of the species identified correctly in leave-one-out cross-validation and 93.75% in the validation step.



(a) Training (b) Test

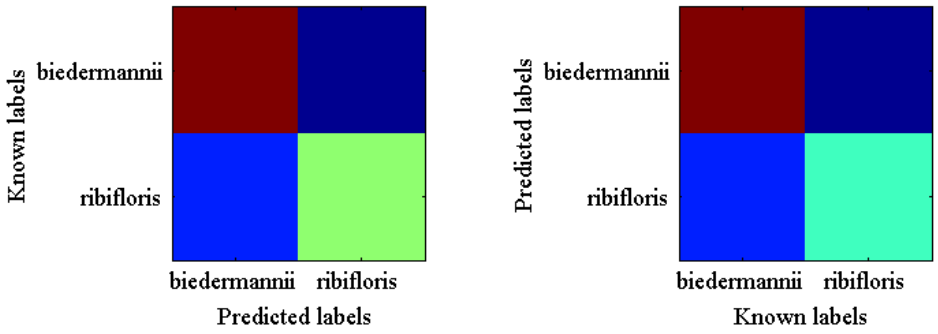
Figure 51: Confusion matrices for subspecies in *Osmia lignaria*

The results from classification of subspecies within *Osmia ribifloris* are shown in Figures 53 and 54. In this classification problem the first two principal components were used, which gave 91.30% of the species identified correctly in leave-one-out cross-validation and 90.91% in the validation step.



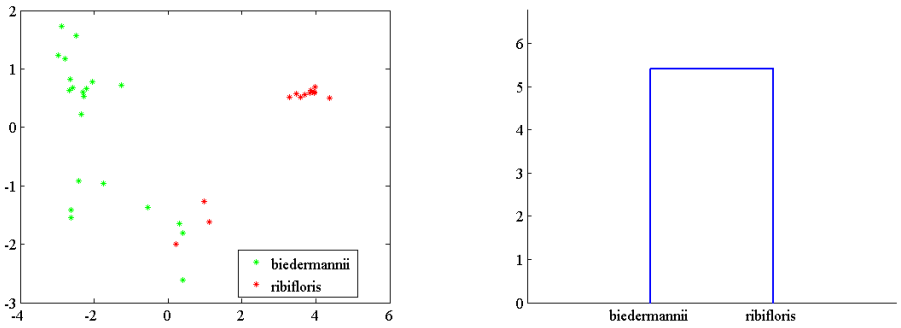
(a) PCA (b) Dendrogram

Figure 52: Subspecies in *Osmia lignaria* discrimination visualizations for evaluation



(a) Training (b) Test

Figure 53: Confusion matrices for subspecies in *Osmia ribifloris*



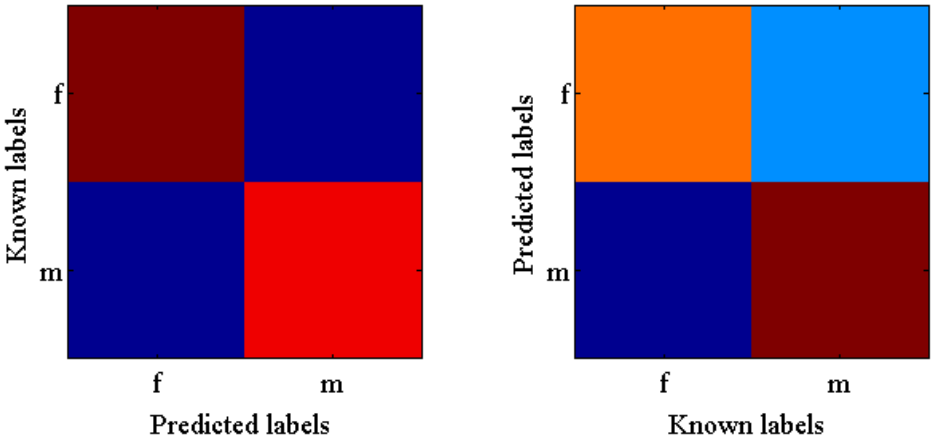
(a) PCA (b) Dendrogram

Figure 54: Subspecies in *Osmia ribifloris* discrimination visualizations for evaluation

B.4 Gender Classification

Like the subspecies classification results the gender classification results were in general worse than the species classification results, and had smaller training sets. It would be desirable to create larger training sets to see if the results could be improved.

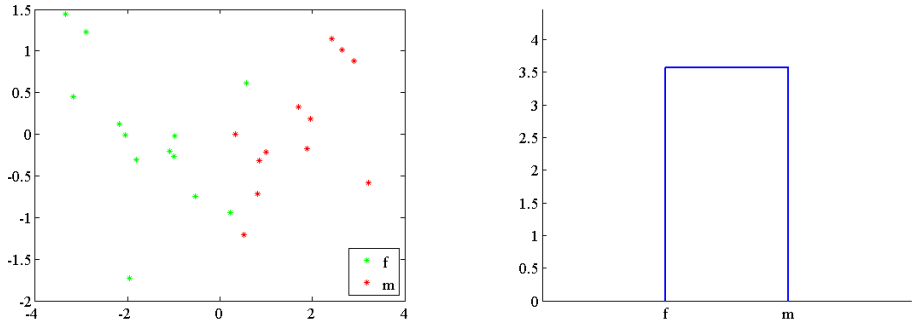
The results from classification of gender within *Osmia pusilla* are shown in Figures 55 and 56. In this classification problem the first principal component was used, which gave 100% of the species identified correctly in leave-one-out cross-validation and 87.5% in the validation step.



(a) Training (b) Test

Figure 55: Confusion matrices for gender in *Osmia pusilla*

The results from classification of gender within *Osmia coloradensis* are shown in Figures 57 and 58. In this classification problem the first principal component was used, which gave 100% of the species identified correctly in leave-one-out cross-validation and 87.5% in the validation step.

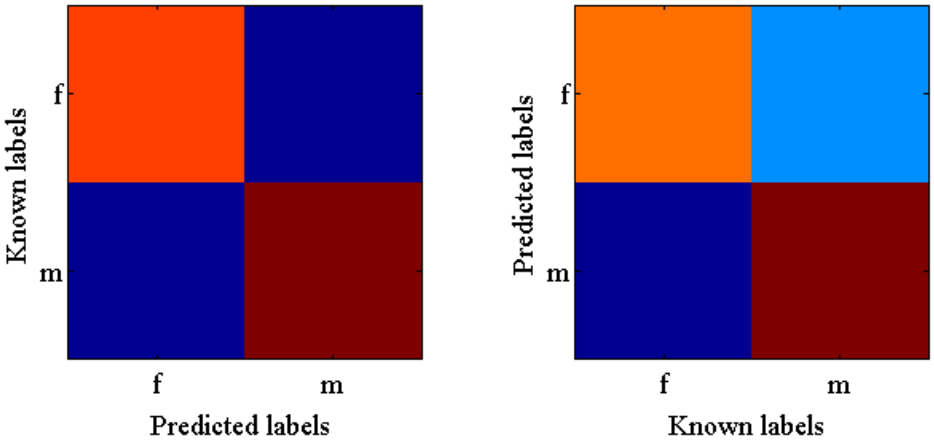


(a) PCA (b) Dendrogram

Figure 56: Gender in *Osmia pusilla* discrimination visualizations for evaluation

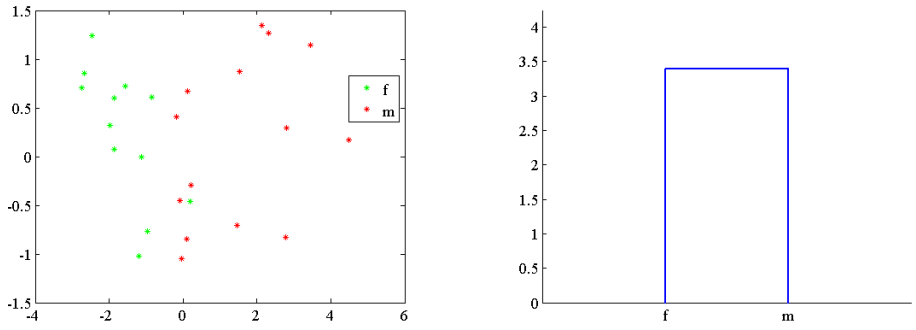
The results from classification of gender within *Osmia ribifloris* are shown in Figures 59 and 60. In this classification problem the first principal component was used, which gave 93.94% of the species identified correctly in leave-one-out cross-validation and 93.75% in the validation step.

The results from classification of gender within *Osmia texana* are shown in Figures 61 and 62. In this classification problem the first two principal components were used, which gave 89.47% of the species identified correctly in leave-one-out cross-validation and 88.89% in the validation step.



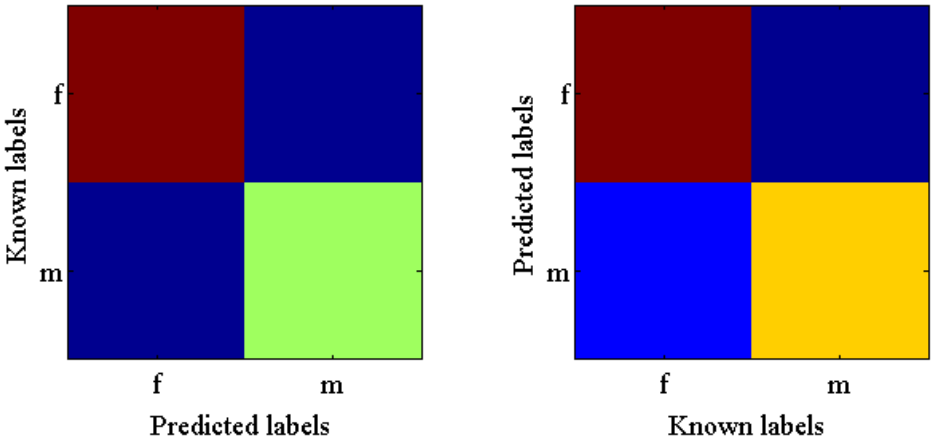
(a) Training (b) Test

Figure 57: Gender in *Osmia coloradensis* discrimination visualizations for evaluation



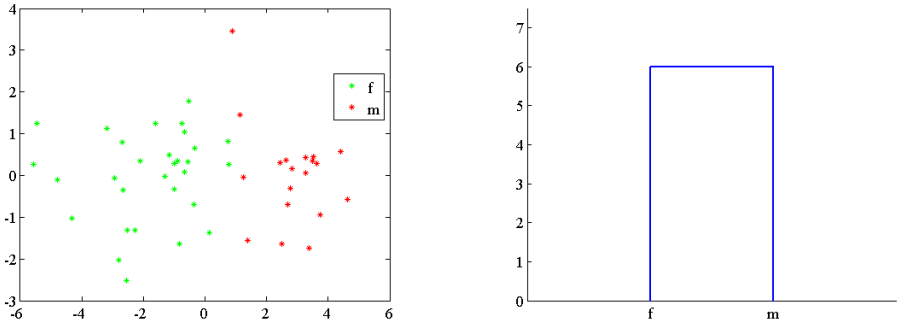
(a) PCA (b) Dendrogram

Figure 58: Gender in *Osmia coloradensis* discrimination visualizations for evaluation



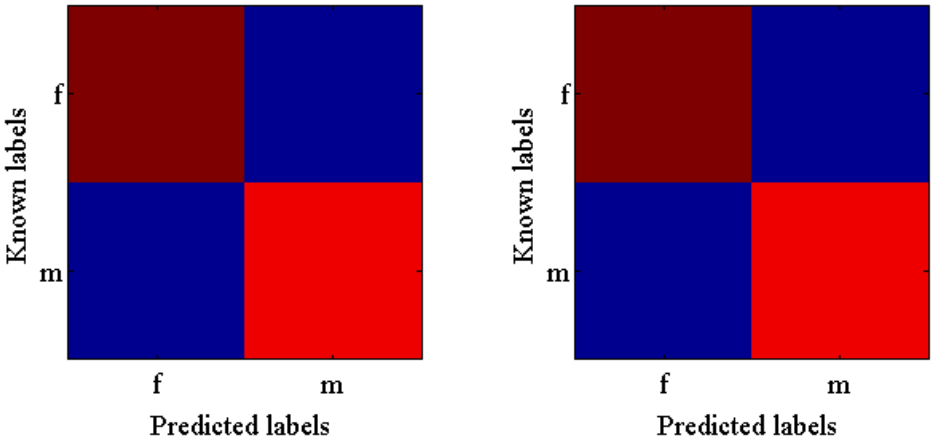
(a) Training (b) Test

Figure 59: Confusion matrices for gender in *Osmia ribifloris*



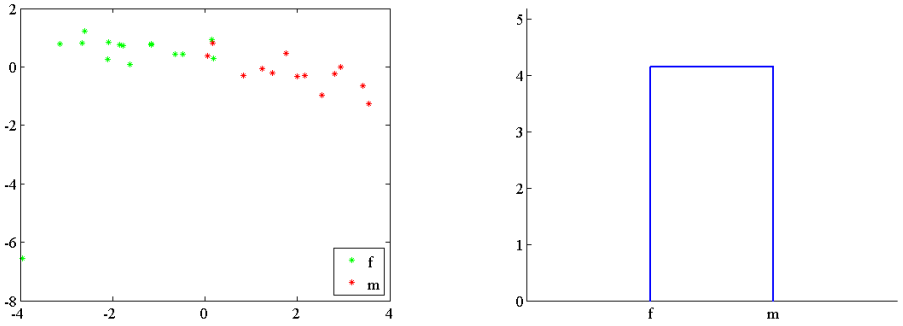
(a) PCA (b) Dendrogram

Figure 60: Gender in *Osmia ribifloris* discrimination visualizations for evaluation



(a) Training (b) Test

Figure 61: Confusion matrices for gender in *Osmia texana*



(a) PCA (b) Dendrogram

Figure 62: Gender in *Osmia texana* discrimination visualizations for evaluation

Appendix C

Classification Results Overview

Classification was done in a hierarchical fashion, that is genus was classified before classifying the species. This approach gave much better results then classifying each specimen to all of its classes at once as the following tables show [37].

Genus	**Training**	**Test**
Agapostemon	88.33	65.00
Bombus	85.00	89.63
Ceratina	93.75	88.89
Lasioglossum	81.73	83.03
Osmia	81.12	91.11
Total Average	86.09	83.27

Table 6: Species classification, percent correct by genus with genus unknown

Genus	**Training**	**Test**
Agapostemon	100	100
Bombus	93.88	91.67
Ceratina	100	100
Lasioglossum	97.18	97.73
Osmia	97.87	100
Total Average	97.79	97.88

Table 7: Species classification, percent correct by genus with genus known from a prior classification

Species	**Training**	**Test**
Agapostemon sericeus	75.00	80.00
Agapostemon texanus	90.00	90.00
Agapostemon virescens	100.00	25.00
Bombus impatiens	85.00	100.00
Bombus bimaculata	90.00	80.00
Bombus griseocolis	80.00	80.00
Ceratina calcarata	93.75	88.89
Lasioglossum MAWIspB	100.00	100.00
Lasioglossum acuminatum	52.94	44.44
Lasioglossum coriaceum	80.00	90.00
Lasioglossum leucozonium	41.67	42.86
Lasioglossum nymphaerum	90.00	70.00
Lasioglossum pilosum	100.00	100.00
Lasioglossum rohweri	90.00	100.00
Lasioglossum zephyrum	100.00	100.00
Lasioglossum zonulum	80.95	100.00
Osmia coloradensis	52.63	66.67
Osmia lignaria	90.48	88.89
Osmia pusilla	100.00	100.00
Osmia ribifloris	100.00	100.00
Osmia texana	62.50	100.00
Total Average	84.71	84.34

Table 8: Species classification, percent correct by species with genus unknown

Species	**Training**	**Test**
Agapostemon sericeus	100.00	100.00
Agapostemon texanus	100.00	100.00
Agapostemon virescens	100.00	100.00
Bombus impatiens	100.00	100.00
Bombus bimaculata	76.92	66.67
Bombus griseocolis	100.00	100.00
Ceratina calcarata	100.00	100.00
Lasioglossum MAWIspB	95.00	100.00
Lasioglossum acuminatum	85.00	70.00
Lasioglossum coriaceum	100.00	100.00
Lasioglossum leucozonium	100.00	100.00
Lasioglossum nymphaerum	100.00	100.00
Lasioglossum pilosum	100.00	100.00
Lasioglossum rohweri	100.00	100.00
Lasioglossum zephyrum	100.00	100.00
Lasioglossum zonulum	93.75	77.78
Osmia coloradensis	100.00	75.00
Osmia lignaria	100.00	100.00
Osmia pusilla	100.00	100.00
Osmia ribifloris	100.00	100.00
Osmia texana	90.00	100.00
Total Average	97.17	94.74

Table 9: Species classification, percent correct by species with genus known from a prior classification

Bibliography

[1] V. Steinhage, S. Schröder, V. Roth, A. B. Cremers, W. Drescher, and D. Wittmann, "The science of "Fingerprinting" bees," *German Research*, vol. 28, no. 1, p. 19–21, 2006.

[2] C. D. Michener, R. J. McGinley, and B. N. Danforth, *The bee genera of North and Central America (Hymenoptera: Apoidea)*. Smithsonian Institution Press, 1994.

[3] N. MacLeod and S. Association, *Automated Taxon Identification in Systematics: Theory, Approaches and Applications*. Boca Raton: CRC Press, 2007.

[4] D. Vanengelsdorp, J. D. Evans, C. Saegerman, C. Mullin, E. Haubruge, B. K. Nguyen, M. Frazier, J. Frazier, D. Cox-Foster, Y. Chen, *et al.*, "Colony collapse disorder: a descriptive study," *PLoS One*, vol. 4, no. 8, p. e6481, 2009.

[5] P. C. D. Michener, *The Bees of the World*. The Johns Hopkins University Press, 1 ed., June 2000.

[6] T. B. Mitchell, "Bees of the eastern united states.," 1960.

[7] N. MacLeod, M. Benfield, and P. Culverhouse, "Time to automate identification," *Nature*, vol. 467, no. 7312, pp. 154–155, 2010.

[8] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Human Genetics*, vol. 7, no. 2, p. 179–188, 1936.

[9] P. Belhumeur, D. Chen, S. Feiner, D. Jacobs, W. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, *et al.*, "Searching the world's herbaria: A system for visual identification of plant species," *Computer Vision–ECCV 2008*, p. 116–129, 2008.

[10] M. Mayo and A. T. Watson, "Automatic species identification of live moths," *Knowledge-Based Systems*, vol. 20, no. 2, p. 195–202, 2007.

[11] F. J. Rohlf and J. W. Archie, "A comparison of fourier methods for the description of wing shape in mosquitoes (Diptera: culicidae)," *Systematic Biology*, vol. 33, no. 3, p. 302, 1984.

[12] C. Bi, M. C. Saunders, and B. A. McPheron, "Wing Pattern-Based classification of the rhagoletis pomonella species complex using genetic neural networks," *International Journal of Computer Science & Applications*, vol. 4, p. 1–14, 2007.

[13] P. R. Prado, L. da Fontoura Costa, E. M. Moraes, M. H. Manfrin, and F. M. Sene, "Curvature measurement as a tool for the analysis of morphometric variation using drosophila wings as a model," *Braz. J. morphol. Sci*, vol. 23, no. 3-4, p. 333–342, 2006.

[14] A. Tofilski, "DrawWing, a program for numerical description of insect wings," *Journal of Insect Science*, vol. 4, 2004.

[15] T. M. Francoy, R. A. O. Silva, P. Nunes-Silva, C. Menezes, and V. L. Imperatriz-Fonseca, "Gender identification of five genera of stingless bees (Apidae, meliponini) based on wing morphology," *Genet Mol Res*, vol. 8, p. 207–214, 2009.

[16] T. M. Francoy, M. L. Grassi, V. L. Imperatriz-Fonseca, W. de Jesús May-Itzá, and J. J. Quezada-Euán, "Geometric morphometrics of the wing as a tool for assigning genetic lineages and geographic origin to melipona beecheii (Hymenoptera: meliponini)," *Apidologie*, p. 1–9, 2010.

[17] S. Schroder, D. Wittmann, W. Drescher, V. Roth, V. Steinhage, and A. B. Cremers, "The new key to bees: Automated identification by image analysis of wings," *Pollinating bees–the Conservation Link Between Agriculture and Nature, Ministry of Environment, Brasilia*, 2002.

[18] V. Roth, A. Pogoda, V. Steinhage, and S. Schröder, "Pattern recognition combining feature- and pixel-based classification within a real world application," 1999.

[19] V. Steinhage, "Automated identification of bee species in biodiversity information systems," *Computer Science for Environmental Protection*, p. 4–6, 2000.

[20] V. Steinhage, "Re: ABIS," July 2011.

[21] T. M. Francoy, D. Wittmann, M. Drauschke, S. Müller, V. Steinhage, M. A. Bezerra-Laure, D. D. Jong, and L. S. Gonçalves, "Identification of africanized honey bees through wing morphometrics: two fast and efficient procedures," *Apidologie*, vol. 39, no. 5, p. 488–494, 2008.

[22] S. Zhao, W. Ding, and D. Liu, "Technical note: Extraction of wing venation characteristics by image analysis," 2010.

[23] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, p. 404–417, 2006.

[24] D. G. Lowe, "Object recognition from local scale-invariant features," in *iccv*, p. 1150, 1999.

[25] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, p. 381–395, 1981.

[26] S. Droege, V. J. Tepedino, G. Lebuhn, W. Link, R. L. Minckley, Q. Chen, and C. Conrad, "Spatial patterns of bee captures in north american bowl trapping surveys," *Insect Conservation and Diversity*, vol. 3, no. 1, p. 15–23, 2010.

[27] R. Grundel, K. J. Frohnapple, R. P. Jean, and N. B. Pavlovic, "Effectiveness of bowl trapping and netting for inventory of a bee community," *Environmental Entomology*, vol. 40, no. 2, p. 374–380, 2011.

[28] J. K. Tuell and R. Isaacs, "Elevated pan traps to monitor bees in flowering crop canopies," *Entomologia Experimentalis et Applicata*, vol. 131, no. 1, p. 93–98, 2009.

[29] T. H. Roulston, S. A. Smith, and A. L. Brewster, "A comparison of pan trap and intensive net sampling techniques for documenting a bee (Hymenoptera: apiformes) fauna," *Journal of the Kansas Entomological Society*, vol. 80, no. 2, p. 179–181, 2007.

[30] J. S. Wilson, T. Griswold, and O. J. Messinger, "Sampling bee communities (Hymenoptera: apiformes) in a desert landscape: are pan traps sufficient?," *Journal of the Kansas Entomological Society*, vol. 81, no. 3, p. 288–300, 2008.

[31] R. C. Gonzalez and R. E. Woods, *Digital Image Processing.* Prentice Hall, 3 ed., Aug. 2007.

[32] N. Otsu *et al.*, "A threshold selection method from gray-level histograms," *IEEE Transactions on systems, Man, and Cybernetics*, vol. 9, no. 1, p. 62–66, 1979.

[33] L. Lam, S. W. Lee, and C. Y. Suen, "Thinning methodologies-a comprehensive survey," *IEEE Transactions on pattern analysis and machine intelligence*, p. 869–885, 1992.

[34] C. Darwin, *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life.* London, 1907.

[35] J. Shlens, "A tutorial on principal component analysis," *Measurement*, vol. 51, 2005.

[36] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* Wiley, 2001.

[37] V. Steinhage, B. Kastenholz, S. Schröder, and W. Drescher, "A hierarchical approach to classify solitary bees based on image analysis," *Bees Based On Image Analysis, In: Mustererkennung*, 1997.

[38] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, p. 1226–1238, 2005.

[39] L. I. Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, p. 52, 2002.

[40] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," *LECTURE NOTES IN STATISTICS-NEW YORK-SPRINGER VERLAG-*, p. 199–206, 1996.

[41] G. C. Cawley, N. L. Talbot, and M. Girolami, "Sparse multinomial logistic regression via bayesian l1 regularisation," in *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*, vol. 19, p. 209, 2007.

[42] H. Liu and H. Motoda, *Computational methods of feature selection*. Chapman & Hall, 2008.

[43] L. J. Wei, "Asymptotic conservativeness and efficiency of Kruskal-Wallis test for k dependent samples," *Journal of the American Statistical Association*, p. 1006–1009, 1981.

[44] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 2008.

[45] C. Gini, *Memorie di metodologia statistica*, vol. 1. EV Veschi, 1955.

[46] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," *Data Mining and Knowledge Discovery*, vol. 6, no. 4, p. 393–423, 2002.

[47] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Machine Learning-International Workshop then Conference*, vol. 20, p. 856, 2003.

[48] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Tools with Artificial Intelligence, 1995. Proceedings., Seventh International Conference on*, p. 388–391, 1995.

[49] G. C. Cawley and N. L. Talbot, "Gene selection in cancer classification using sparse logistic regression with bayesian regularization," *Bioinformatics*, vol. 22, no. 19, p. 2348, 2006.

[50] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," *Division of Biostatistics, Stanford University, Palo Alto, CA, Technical report*, 1999.

[51] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, p. 520, 2001.

[52] T. P. Speed, *Statistical analysis of gene expression microarray data*. CRC Press, 2003.

[53] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, p. 123–140, 1996.

[54] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, p. 5–32, 2001.

[55] L. Breiman, *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[56] Z. Nenadic, "Classwise principal component analysis," 2010.

[57] K. Das and Z. Nenadic, "An efficient discriminant-based solution for small sample size problem," *Pattern Recognition*, vol. 42, no. 5, p. 857–866, 2009.

[58] K. Das, S. Osechinskiy, and Z. Nenadic, "A classwise pca-based recognition of neural data for brain-computer interfaces," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, p. 6519–6522, 2007.

[59] K. Das, D. S. Rizzuto, and Z. Nenadic, "Mental state estimation for Brain–Computer interfaces," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 8, p. 2114–2122, 2009.

[60] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes-which naive bayes," in *Third conference on email and anti-spam (CEAS)*, vol. 17, p. 28–69, 2006.

[61] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proceedings of the eleventh conference on uncertainty in artificial intelligence*, vol. 1, p. 338–345, 1995.

[62] T. M. Mitchell, "Machine learning," *Burr Ridge, IL: McGraw Hill*, 1997.

[63] V. Kecman, *Learning and soft computing*. Massachusetts Institute of Technology, 2001.

[64] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press, 2002.

[65] J. A. Suykens, *Least squares support vector machines*. World Scientific Pub Co Inc, 2002.

[66] N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge Univ Pr, 2000.

[67] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," *Recent advances in learning and control*, p. 95–110, 2008.

[68] Y. Freund and R. E. Schapire, "Game theory, on-line prediction and boosting," in *Proceedings of the ninth annual conference on Computational learning theory*, p. 325–332, 1996.

[69] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *The Journal of Machine Learning Research*, vol. 4, p. 933–969, 2003.

[70] A. Vezhnevets and V. Vezhnevets, "Modest AdaBoost-teaching AdaBoost to generalize better," *Graphicon-2005. Novosibirsk Akademgorodok, Russia*, 2005.

[71] T. G. Dietterichl, "Ensemble learning," 2002.

[72] Y. Freund and R. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*, p. 23–37, 1995.

[73] R. E. Schapire, "A brief introduction to boosting," in *International Joint Conference on Artificial Intelligence*, vol. 16, p. 1401–1406, 1999.

[74] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, vol. 30, no. June, p. 271–274, 1998.

[75] J. B. Phipps, "Dendrogram topology," *Systematic Biology*, vol. 20, no. 3, p. 306, 1971.

[76] R. M. Bateman, J. Hilton, and P. J. Rudall, "Morphological and molecular phylogenetic context of the angiosperms: contrasting the 'top-down' and 'bottom-up' approaches used to infer the likely characteristics of the first flowers," *Journal of Experimental Botany*, vol. 57, pp. 3471 –3503, Oct. 2006.

[77] S. M. Kay, *Fundamentals of statistical signal processing -: detection theory*. PTR Prentice-Hall, Feb. 1998.

[78] S. M. Kay, *Fundamentals Of Statistical Signal Processing*. Prentice Hall, Mar. 2001.

[79] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Science, Calcutta*, vol. 12, p. 49, 1936.

[80] E. N. Mortensen, L. J. Reese, and W. A. Barrett, "Intelligent selection tools," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2, p. 776–777, 2000.

[81] E. Shevtsova, C. Hansson, D. H. Janzen, and J. Kjærandsen, "Stable structural color patterns displayed on transparent insect wings," *Proceedings of the National Academy of Sciences*, vol. 108, no. 2, p. 668, 2011.