

Flying Insect Classification with Inexpensive Sensors

Yanping Chen · Adena Why · Gustavo Batista ·
Agenor Mafra-Neto · Eamonn Keogh

Revised: 29 May 2014 / Accepted: 30 May 2014 /
Published online: 21 June 2014
© Springer Science+Business Media New York 2014

Abstract The ability to use inexpensive, noninvasive sensors to accurately classify flying insects would have significant implications for entomological research, and allow for the development of many useful applications in vector control for both medical and agricultural entomology. Given this, the last sixty years have seen many research efforts on this task. To date, however, none of this research has had a lasting impact. In this work, we explain this lack of progress. We attribute the stagnation on this problem to several factors, including the use of *acoustic* sensing devices, the overreliance on the single feature of wingbeat frequency, and the attempts to learn complex models with relatively little data. In contrast, we show that pseudo-acoustic *optical* sensors can produce vastly superior data, that we can exploit additional features, both intrinsic and extrinsic to the insect's flight behavior, and that a Bayesian classification approach allows us to efficiently learn classification models that are very robust to overfitting. We demonstrate our findings with large scale experiments, as measured both by the number of insects and the number of species considered.

Y. Chen (✉)

Department of Computer Science & Engineering, University of California, Riverside, CA, USA
e-mail: ychen053@cs.ucr.edu

A. Why

Department of Entomology, University of California, Riverside, CA, USA
e-mail: awhy001@ucr.edu

G. Batista

University of São Paulo - USP, São Paulo, Brazil
e-mail: gbatista@icmc.usp.br

A. Mafra-Neto

ISCA Technologies, Riverside, CA, USA
e-mail: president@iscatech.com

E. Keogh

Department of Computer Science & Engineering, University of California, Riverside, CA, USA
e-mail: eamonn@cs.ucr.edu

Keywords Automate insect classification · insect flight sound · insect wingbeat · Bayesian classifier · flight activity circadian rhythm

Introduction

The idea of automatically classifying insects using the incidental sound of their flight (as opposed to *deliberate* insect sounds produced by stridulation (Hao et al. 2012)) dates back to the very dawn of computers and commercially available audio recording equipment. In 1945,¹ three researchers at the Cornell University Medical College, Kahn, Celestin and Offenhauser, used equipment donated by Oliver E. Buckley (then President of the Bell Telephone Laboratories) to record and analyze mosquito sounds (Kahn et al. 1945).

The authors later wrote, “*It is the authors’ considered opinion that the intensive application of such apparatus will make possible the precise, rapid, and simple observation of natural phenomena related to the sounds of disease-carrying mosquitoes and should lead to the more effective control of such mosquitoes and of the diseases that they transmit.*” (Kahn and Offenhauser 1949). In retrospect, given the importance of insects in human affairs, it seems astonishing that more progress on this problem has not been made in the intervening decades.

There have been sporadic efforts at flying insect classification from audio features (Sawedal and Hall 1979; Schaefer and Bent 1984; Unwin and Ellington 1979; Moore et al. 1986), especially in the last decade (Moore and Miller 2002; Repasky et al. 2006); however, little real progress seems to have been made. By “lack of progress” we do not mean to suggest that these pioneering research efforts have not been fruitful. However, we would like to have *automatic* classification to become as simple, inexpensive, and ubiquitous as current *mechanical* traps such as sticky traps or interception traps (Capinera 2008), but with all the advantages offered by a digital device: higher accuracy, very low cost, real-time monitoring ability, and the ability to collect additional information (time of capture,² etc.).

We feel that the lack of progress in this pursuit can be attributed to three related factors:

1. Most efforts to collect data have used *acoustic* microphones (Reed et al. 1942; Belton and Costello 1979; Mankin et al. 2006; Raman et al. 2007). Sound attenuates according to an inverse squared law. For example, if an insect flies just three times further away from the microphone, the sound intensity (informally, the loudness) drops to one ninth. Any attempt to mitigate this by using a more sensitive microphone invariably results in extreme sensitivity to wind noise and to ambient noise in the environment. Moreover, the difficulty of collecting data with such devices seems to have led some researchers to obtain data in unnatural conditions.

¹ An even earlier paper (Reed et al. 1941) makes a similar suggestion. However, these authors determined the wingbeat frequencies *manually*, aided by a stroboscope.

² A commercially available rotator bottle trap made by BioQuip® (2850) does allow researchers to measure the time of arrival at a granularity of *hours*. However, as we shall show in Section *Additional Feature: Circadian Rhythm of Flight Activity*, we can measure the time of arrival at a *sub-second* granularity and exploit this to improve classification accuracy.

For example, nocturnal insects have been forced to fly by tapping and prodding them under bright halogen lights; insects have been recorded in confined spaces or under extreme temperatures (Belton and Costello 1979; Moore and Miller 2002). In some cases, insects were tethered with string to confine them within the range of the microphone (Reed et al. 1942). It is hard to imagine that such insect handling could result in data which would generalize to insects in natural conditions.

2. Unsurprisingly, the difficulty of obtaining data noted above has meant that many researchers have attempted to build classification models with very limited data, as few as 300 instances (Moore 1991) or less. However, it is known that for building classification models, more data is better (Halevy et al. 2009; Banko and Brill 2001; Shotton et al. 2013).
3. Compounding the poor quality data issue and the sparse data issue above is the fact that many researchers have attempted to learn very complicated classification models,³ especially neural networks (Moore et al. 1986; Moore and Miller 2002; Li et al. 2009). However, neural networks have many parameters/settings, including the interconnection pattern between different layers of neurons, the learning process for updating the weights of the interconnections, the activation function that converts a neuron's weighted input to its output activation, etc. Learning these on say a spam/email classification problem with millions of training data is not very difficult (Zhan et al. 2005), but attempting to learn them on an insect classification problem with a mere twenty examples is a recipe for overfitting (cf. Fig. 3). It is difficult to overstate how optimistic the results of neural network experiments can be unless rigorous protocols are followed (Prechelt 1995).

In this work, we will demonstrate that we have largely solved all these problems. We show that we can use optical sensors to record the “sound” of insect flight from meters away, with complete invariance to wind noise and ambient sounds. We demonstrate that these sensors have allowed us to record on the order of *millions* of labeled training instances, far more data than all previous efforts combined, and thus allow us to avoid the overfitting that has plagued previous research efforts. We introduce a principled method to incorporate additional information into the classification model. This additional information can be as quotidian and as easy-to-obtain as the time-of-day, yet still produce significant gains in accuracy. Finally, we demonstrate that the enormous amounts of data we collected allow us to take advantage of “*The unreasonable effectiveness of data*” (Halevy et al. 2009) to produce simple, accurate and robust classifiers.

In summary, we believe that flying insect classification has moved beyond the dubious claims created in the research lab and is now ready for real-world deployment. The sensors and software we present in this work will provide researchers worldwide robust tools to accelerate their research.

³ While there is a formal framework to define the complexity of a classification model (i.e. the VC dimension (Vapnik and Chervonenkis 1971)), informally we can think of a *complicated* or *complex* model as one that requires many parameters to be set or learned.

Background and Related Work

The vast majority of attempts to classify insects by their flight sounds have explicitly or implicitly used *just* the wingbeat frequency (Reed et al. 1942; Sotavalta 1947; Sawedal and Hall 1979; Schaefer and Bent 1984; Unwin and Ellington 1979; Moore et al. 1986; Moore 1991). However, such an approach is limited to applications in which the insects to be discriminated have very different frequencies. Consider Fig. 1. I which shows a histogram created from measuring the wingbeat frequencies of three (sexed) species of insects, *Culex stigmatosoma* (female), *Aedes aegypti* (female), and *Culex tarsalis* (male) (We defer details of how the data was collected until later in the paper).

It is visually obvious that if asked to separate *Cx. stigmatosoma* ♀ from *Cx. tarsalis* ♂, the wingbeat frequency could produce an accurate classification, as the two species have very different frequencies with minimal overlap. To see this, we can compute the optimal Bayes error rate (Fukunaga 1990), which is a strict lower bound to the actual error rate obtained by *any* classifier that considers *only* this feature. Here, the Bayes error rate is half the *overlapping* area under both curves divided by the *total* area under the two curves.

Because there is only a tiny overlap between the wingbeat frequency distributions of the two species, the Bayes error rate is correspondingly small, 0.57 % if we use the raw histograms and 1.08 % if we use the derived Gaussians.

However, if the task is to separate *Cx. stigmatosoma* ♀ from *Ae. aegypti* ♀, the wingbeat frequency will not do as well, as the frequencies of these two species overlap greatly. In this case, the Bayes error rate is *much* larger, 24.90 % if we use the raw histograms and 30.95 % if we use the derived Gaussians.

This problem can only get worse if we consider more species, as there will be increasing overlap among the wingbeat frequencies. This phenomenon can be understood as a real-value version of the Pigeonhole principle (Grimaldi 1989). Given this, it is unsurprising that some doubt the utility of wingbeat sounds to classify the insects. However, we will show that the analysis above is

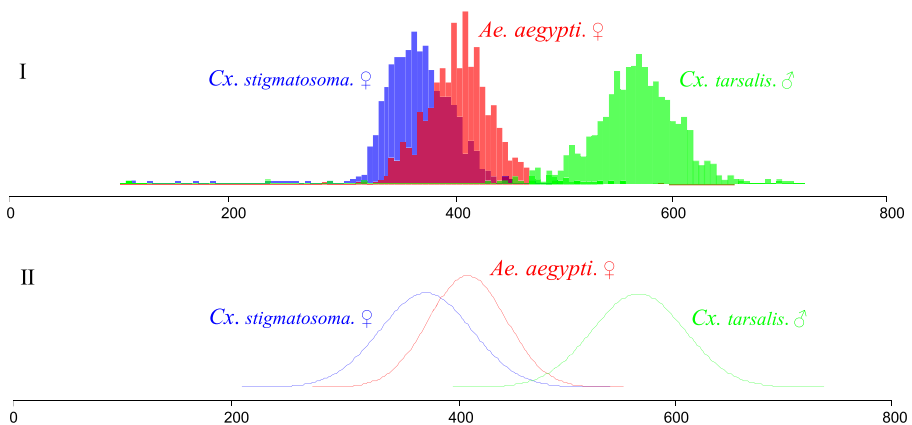


Fig. 1 I) Histograms of wingbeat frequencies of three species of insects, *Cx. stigmatosoma* ♀, *Ae. aegypti* ♀, and *Cx. tarsalis* ♂. Each histogram is derived based on 1,000 wingbeat sound snippets. II) Gaussian curves that fit the wingbeat frequency histograms

pessimistic. Insect flight sounds *can* allow much higher classification rates than the above suggests because:

- There is more information in the flight sound signal than just the wingbeat frequency. By analogy, humans have no problem distinguishing between Middle C on a piano and Middle C on a saxophone, even though both are the same 261.62 Hz fundamental frequency. The Bayes error rate to classify the three species in Fig. 2. I using *just* the wingbeat frequency is 19.13 %; however, as we shall see below in the section titled *Flying Insect Classification*, that by using the additional features from the wingbeat signal, we can obtain an error rate of 12.43 %.
- We can augment the wingbeat sounds with additional cheap-to-obtain features that can help to improve the classification performance. For example, many species may have different flight activity circadian rhythms. As we shall see below in the section titled *Additional Feature: Circadian Rhythm of Flight Activity and Geographic Distribution*, simply incorporating the *time-of-intercept* information can significantly improve the performance of the classification.

The ability to allow the incorporation of auxiliary features is one of the reasons we argue that the Bayesian classifier is ideal for this task (cf. Section *Flying Insect Classification*), as it can gracefully incorporate evidence from multiple sources and in multiple formats.

Materials and Methods

Insect Colony and Rearing

Six species of insects were studied in this work: *Cx. tarsalis*, *Cx. stigmatosoma*, *Ae. aegypti*, *Culex quinquefasciatus*, *Musca domestica* and *Drosophila simulans*.

All adult insects were reared from laboratory colonies derived from wild individuals collected at various locations. *Cx. tarsalis* colony was derived from wild individuals collected at the Eastern Municipal Water District's demonstration constructed treatment

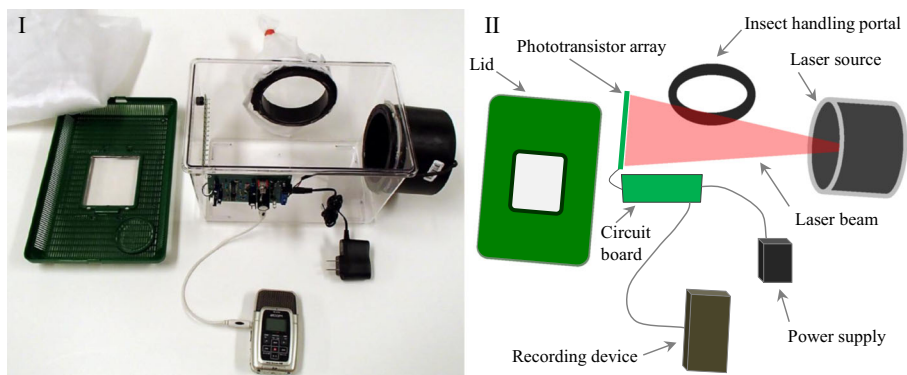


Fig. 2 One of the cages used to gather data for this project. II) A logical version of the setup with the components annotated

wetland (San Jacinto, CA) in 2001. *Cx. quinquefasciatus* colony was derived from wild individuals collected in southern California in 1990 (Georghiou and Wirth 1997). *Cx. stigmatosoma* colony was derived from wild individuals collected at the University of California, Riverside, Aquatic Research Facility in Riverside, CA in 2012. *Ae. aegypti* colony was started in 2000 with eggs from Thailand (Van Dam and Walton 2008). *Musca domestica* colony was derived from wild individuals collected in San Jacinto, CA in 2009, and *Drosophila simulans* colony were derived from wild individuals caught in Riverside, CA in 2011.

The larvae of *Cx. tarsalis*, *Cx. quinquefasciatus*, *Cx. stigmatosoma* and *Ae. aegypti* were reared in enamel pans under standard laboratory conditions (27 °C, 16:8 h light:dark [LD] cycle with 1 h dusk/dawn periods) and fed ad libitum on a mixture of ground rodent chow and Brewer's yeast (3:1, v:v). *Musca domestica* larvae were kept under standard laboratory conditions (12:12 h light:dark [LD] cycle, 26 °C, 40 % RH) and reared in a mixture of water, bran meal, alfalfa, yeast, and powdered milk. *Drosophila simulans* larvae were fed ad libitum on a mixture of rotting fruit.

Mosquito pupae were collected into 300-mL cups (Solo Cup Co., Chicago IL) and placed into experimental chambers. Alternatively, adults were aspirated into experimental chambers within 1 week of emergence. The adult mosquitoes were allowed to feed ad libitum on a 10 % sucrose and water mixture; food was replaced weekly. Cotton towels were moistened, twice a week, and placed on top of the experimental chambers and a 300-ml cup of tap water (Solo Cup Co., Chicago IL) was kept in the chamber at all times to maintain a higher level of humidity within the cage. *Musca domestica* adults were fed ad libitum on a mixture of sugar and low-fat dried milk, with free access to water. *Drosophila simulans* adults were fed ad libitum on a mixture of rotting fruit.

Experimental chambers consisted of Kritter Keepers (Lee's Aquarium and Pet Products, San Marcos, CA) that were modified to include the sensor apparatus as well as a sleeve (Bug Dorm sleeve, Bioquip, Rancho Dominguez, CA) attached to a piece of PVC piping to allow access to the insects. Two different sizes of experimental chambers were used, the larger 67 cm L x 22 cm W x 24.75 cm H, and the smaller 30 cm L x 20 cm W x 20 cm H. The lids of the experimental chambers were modified with a piece of mesh cloth affixed to the inside in order to prevent escape of the insects, as shown in Fig. 2. I. Experimental chambers were maintained on a 16:8 h light:dark [LD] cycle, 20.5–22 °C and 30–50 % RH for the duration of the experiment. Each experimental chamber contained 20 to 40 individuals of a same species, in order to capture as many flying sounds as possible while limiting the possibility of capturing more than one insect-generated sound at a same time.

Some tests were conducted with newly emerged adults, which would be virgins, but other trials were not. Anecdotally this appears to make no difference to the task-at-hand, however a formal study is currently underway by an independent group of researchers using our sensors and software.

Instruments to Record Flying Sounds

We used the sensor described in (Batista et al. 2011) to capture the insect flying sounds. The logic design of the sensor consists of a phototransistor array which is connected to an electronic board, and a laser line pointing at the phototransistor array. When an insect flies across the laser beam, its wings partially occlude the light, causing small light fluctuations. The light fluctuations are captured by the phototransistor array as

changes in current, and the signal is filtered and amplified by the custom designed electronic board. The physical version of the sensor is shown in Fig. 2. I.

The output of the electronic board feeds into a digital sound recorder (Zoom H2 Handy Recorder) and is recorded as audio data in the MP3 format. Each MP3 file is 6 h long, and a new file starts recording immediately after a file has recorded for 6 h, so the data is continuous. The length of the MP3 file is limited by the device firmware rather than the disk space. The MP3 standard is a lossy format and optimized for human perception of speech and music. However, most flying insects produce sounds that are well within the range of human hearing and careful comparisons to lossless recordings suggest that we lose no exploitable (or indeed, *detectable*) information.

Sensor Data Processing

We downloaded the MP3 sound files to a PC twice a week and used a detection algorithm to automatically extract the brief insect flight sounds from the raw recording data. The detection algorithm used a sliding window to “slide” through the raw data. At each data point, a classifier/detector is used to decide whether the audio segment contains an insect flying sound. It is important to note that the classifier used at this stage is solving the relatively simple two-class task, differentiating between *insect/non-insect*. We will discuss the more sophisticated classifier, which attempts to differentiate species and sex, in the next section.

The classifier/detector used for the *insect/non-insect* problem is a nearest neighbor classifier based on the frequency spectrum. For ground truth data, we used ten flying sounds extracted from early experiments as the training data for the *insect* sounds, and ten segments of raw recording background noise as the training data for the *non-insect* sounds. The number of training data was limited to ten, because more training data would slow down the algorithm while fewer data would not represent variability observed. Note that the training data for background sounds can be different from minute to minute. This is because while the frequency spectrum of the background sound has little variance within a short time interval, it can change greatly and unpredictably in the long run. This variability (called *concept drift* in the machine learning community (Tsybmal 2004; Widmer and Kubat 1996)) may be due to the effects of temperature change on the electronics and the slow decline of battery output power etc. Fortunately, given the high signal-to-noise ratio in the audio, the high variation of the *non-insect* sounds does not cause a significant problem. Figure 4. I shows an example of a one-second audio clip containing a flying insect generated by our sensor. As we can see, the signal of insects flying across the laser is well distinguished from the background signal, as the amplitude is much higher and the range of frequency is quite different from that of background sound.

The length of the sliding window in the detection algorithm was set to be 100 ms, which is about the average length of a flying sound. Each detected insect sound is saved into a one-second long WAV format audio file by centering the insect flying signal and padding with zeros elsewhere. This makes all flying sounds the same length and simplifies the future archiving and processing of the data. Note that we converted the audio format from MP3 to WAV at this stage. This is simply because we publicly release all our data so that the community can confirm and extend our results. Because the vast majority of the signal processing community uses Matlab, and Matlab provides native functions for working with WAV files, this is the obvious choice for an archiving format. Figure 4. II shows the saved audio of the insect sound shown in Fig. 4. I.

Flying sounds detected in the raw recordings may be contaminated by the background noise, such as the 60 Hz noise from the American domestic electricity, which “bleeds” into the recording due to the inadequate filtering in power transformers. To obtain a cleaner signal, we applied the spectral subtraction technique (Boll 1979; Ephraim and Malah 1984) to each detected flying sound to reduce noise.

Flying Insect Classification

In the section above, we showed how a simple nearest neighbor classifier can *detect* the sound of insects, and pass the sound snippet on for further inspection. Here, we discuss algorithms to actually classify the snippets down to species (and in some cases, sex) level.

While there are a host of classification algorithms in the literature (decision trees, neural networks, nearest neighbor, etc.), the Bayes classifier is optimal in minimizing the probability of misclassification (Devroye 1996), under the assumption of independence of features. The Bayes classifier is a simple probabilistic classifier that predicts class membership probabilities based on Bayes’ theorem. In addition to its excellent classification performance, the Bayesian classifier has several properties that make it extremely useful in practice and particularly suitable to the task at hand.

1. The Bayes classifier is undemanding in both CPU and memory requirements. Any devices to be deployed in the field in large quantities will typically be small devices with limited resources, such as limited memory, CPU power and battery life. The Bayesian classifier (once constructed offline in the lab) requires time and space resources that are just linear in the number of features.
2. The Bayes classifier is very easy to implement. Unlike neural networks (Moore and Miller 2002; Li et al. 2009), the Bayes classifier does not have many parameters that must be carefully tuned. In addition, the model is fast to build, and it requires only a small amount of training data to estimate the distribution parameters necessary for accurate classification, such as the means and variances of Gaussian distributions.
3. Unlike other classification methods that are essentially “black box”, the Bayesian classifier allows for the graceful introduction of user knowledge. For example, if we have external (to the training data set) knowledge that given the particular location of a deployed insect sensor we should expect to be twice as likely to encounter a *Cx. tarsalis* as an *Ae. aegypti*, we can “tell” the algorithm this, and the algorithm can use this information to improve its accuracy. This means that in some cases, we can augment our classifier with information gleaned from the text of journal papers or simply the experiences of field technicians. In section *A Tentative Additional Feature: Geographic Distribution* in (Chen et al. 2014), we give a concrete example of this. Another example of how the Bayesian classifier allows us to gracefully add domain knowledge is a consideration of the effect of temperature/humidity on flight. While the experiments reported here reflect a single temperate for simplicity, in ongoing work by the current authors, it appears it is possible to predict the changes in wingbeat frequency due to the temperatures effect on air density. This means we can make the Bayesian classifier invariant to changes in temperature, without having to explicitly collect data recorded at different temperatures.

4. The Bayesian classifier simplifies the task flagging anomalies. Most classifiers must make a classification decision, even if the object being classified is vastly different to anything observed in the training phase. In contrast, we can slightly modify the Bayesian classifier to produce an “*Unknown*” classification. One or two such classifications per day could be ignored, but a spate of them could be investigated in case it is indicative of an infestation of a completely unexpected invasive species.
5. When there are multiple features used for classification, we need to consider the possibility of missing values, which happens when some features are not observed. For example, as we discuss below, we use *time-of-intercept* as a feature. However, a dead clock battery could deny us this feature even when the rest of the system is working perfectly. Missing values are a problem for any learner and may cause serious difficulties. However, the Bayesian classifier can trivially handle this problem, simply by dynamically ignoring the feature in question at classification time.

Because of the considerations listed above, we argue that the Bayesian classifier is the best for our problem at hand. Note that our decision to use Bayesian classifier, while informed by the above advantages, was also informed by an extensive empirical comparison of the accuracy achievable by other methods, given that in some situations *accuracy* trumps all other considerations. While we omit exhaustive results for brevity, in Fig. 3 we show a comparison with the neural network classifier, as it is the most frequently used technique in the literature (Moore and Miller 2002). We considered only the frequency spectrum of wingbeat snippets for the three species discussed in Fig. 2. The training data was randomly sampled from a pool of 1,500 objects, and the test data was a completely disjoint set of 1,500 objects, and we tested over 1,000 random resamplings. For the neural network, we used a single hidden layer of size ten, which seemed to be approximately the default parameters in the literature.

The results show that while the neural network classifier eventually converges on the performance of the Bayesian classifier, it is significantly worse for smaller datasets. Moreover, for any dataset size in the range examined, it can occasionally produce pathologically poor results, doing worse than the default rate of 33.3 %.

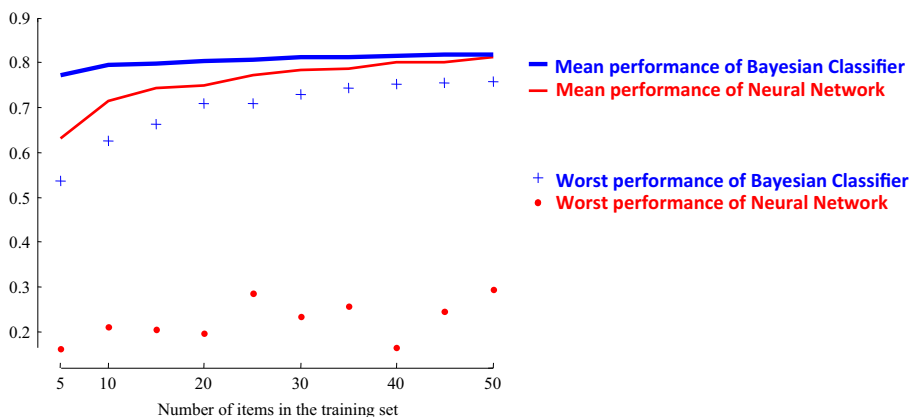


Fig. 3 A comparison of the mean and worst performance of the Bayesian versus Neural Networks Classifiers for datasets ranging in size from five to fifty

Note that our concern about performance on small datasets is only apparently in conflict with our claim that our sensors can produce massive datasets. In some cases, when dealing with new insect species, it may be necessary to bootstrap the modeling of the species by using just a handful of annotated examples to find more (unannotated) examples in the archives, a process known as semi-supervised learning (Chen et al. 2013).

The intuition behind Bayesian classification is to find the mostly likely class given the data observed. When the classifier is based on a single feature F_1 , the probability that an observed data f_1 belongs to a class C_i is calculated as:

$$p(C_i|F_1 = f_1) \propto P(C_i)P(F_1 = f_1|C_i) \quad (1)$$

Where $P(C_i)$ is the prior probability of class C_i and $P(F_1 = f_1|C_i)$ is the class-conditioned probability of observing feature f_1 in class C_i .

For insect classification, the primary data we observed are the flight sounds, as illustrated in Fig. 4. I. The flying sound signal is the non-zero amplitude section (red/bold) in the center of the audio, and can be represented by a sequence $S = \langle s_1, s_2, \dots, s_N \rangle$, where s_i is the signal sampled in the instance i and N is the total number of samples of the signal. This sequence contains a lot of acoustic information, and features can be extracted from it.

The most obvious feature to extract from the sound snippet is the wingbeat frequency. For more details on how to compute wingbeat frequency, please refer to (Chen et al. 2014). Fig. 2. I shows a wingbeat frequency histogram plot for three species of insects (each for a single sex only). We can observe that the histogram for each species is well modeled by a Gaussian distribution. Hence, we fit a Gaussian for each distribution as shown in Fig. 2. II. Note that as hinted at in the introduction, the Bayesian classifier does not have to use the idealized Gaussian distribution; it could use the raw histograms to estimate the probabilities instead. However, using the Gaussian distributions is computationally cheaper at classification time and helps guard against overfitting.

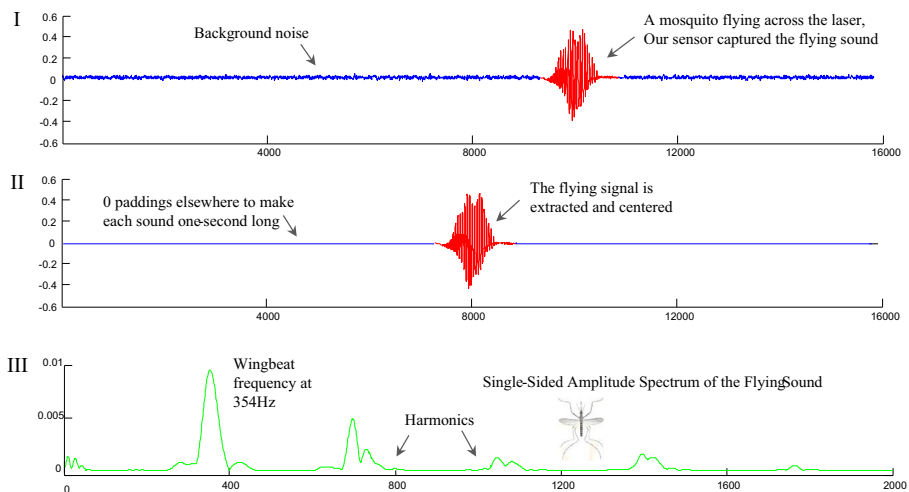


Fig. 4 I) An example of a one-second audio clip containing a flying sound generated by the sensor. The sound was produced by a female *Cx. stigmatosoma*. The insect sound is highlighted in red/bold. II) The insect sound that is cleaned and saved into a one-second long audio clip by centering the insect signal and padding with 0 s elsewhere. III) The frequency spectrum of the insect sound obtained using DFT

For high-dimensional features, such as the frequency spectrum of a sound clip, we can use the k-Nearest-Neighbors (kNN) density estimation approach (Mack and Rosenblatt 1979) to learn the class-conditioned density functions. A more detailed description of the kNN approach, as well as how to estimate the probability of observing an unknown object in class C_i using the density function can be found in (Chen et al. 2014). As such, we are able to estimate the class-conditioned probability for features in any format, including the feature of distance returned from an opaque similarity function, and thus generalize the Bayesian classifier to subsume some of the advantages of the nearest neighbor classifier.

Table 1 outlines the Bayesian classification algorithm. The algorithm begins in Lines 1–3 by estimating the *prior* probability for each class. This is done by counting the number of occurrences of each class in the training data set. It then estimates the conditional probability for each unknown data using the kNN approach. Specifically, given an unknown insect sound, the algorithm first searches the entire training data to find the top k nearest neighbors using some distance measure (Lines 5–9); it then counts for each class the number of neighbors which belong to that class and calculates the class-conditioned probability. With the prior probability and the class-conditioned probability known for each class, the algorithm calculates the *posterior* probability for each class (Lines 13, 15–18) and predicts the unknown data to belong to the class that has the highest posterior probability (Line 19).

Table 1 The Bayesian Classification Algorithm Using a High-dimensional Feature

Notation

k: the number of nearest neighbors in kNN approach

disfunc: a distance function to calculate the distance between two data

C: a set of classes

TRAIN: the training dataset

T_{C_i} : number of training data that belong to class C_i

```

1  for i = 1 : |C|
2       $P(C_i) = T_{C_i} / |\text{TRAIN}|$ ; //estimate prior probability
3  end
4  for each unknown data  $F_1$ 
5      for j = 1 : |TRAIN|
6           $d(j) = \text{disfunc}(F_1, \text{TRAIN}[j])$ ; //the distance of  $F_1$  to each training data
7      end
8       $[d, \text{sort\_index}] = \text{sort}(d, \text{'ascend'})$ ; //sort the distance in ascending order
9       $\text{top\_k} = \text{sort\_index}(1 \text{ to } k)$ ; // find the top-k nearest neighbors
10     for i = 1 : |C|
11          $k_{C_i}$  = number of data in top_k that are labeled as class  $C_i$ ;
12          $P(F_1|C_i) = k_{C_i} / k$ ; // calculate the conditional probability with kNN approach
13          $P(C_i|F_1) = P(C_i) P(F_1|C_i)$ ; // calculate posterior probability
14     end
15      $\text{normalize\_factor} = \sum_{i=1}^{|C|} P(C_i|F_1)$  // normalize the posterior probability
16     for i = 1 : |C|
17          $P(C_i|F_1) = P(C_i|F_1) / \text{normalize\_factor}$ ;
18     end
19      $\hat{C} = \text{argmax}_{C_i \in C} P(C_i|F_1)$  // assign the unknown data  $F_1$  to the class  $\hat{C}$ 
20 end

```

Table 2 Our Distance Measure for two Insect Flight Sounds

Notation:	
S_1, S_2 : two sound sequences	
dis: the distance between the two sounds	
	function dis = disfunc(S_1, S_2)
1	spectrum1 = DFT(S_1);
2	spectrum2 = DFT(S_2);
3	truncateSpectrum1 = spectrum1(frequency range= [100, 2000]);
4	truncateSpectrum2 = spectrum2(frequency range= [100, 2000]);
5	dis = $\sqrt{\sum (\text{truncateSpectrum}_1 - \text{truncateSpectrum}_2)^2}$

The algorithm outlined in Table 1 requires two inputs, including the parameter k . The goal is to choose a value of k that minimizes the probability estimation error. One way to do this is to use *validation* (Kohavi 1995). The idea is to keep part of the training data apart as validation data, and evaluate different values of k based on the estimation accuracy on the validation data. The value of k which achieves the best estimation accuracy is chosen and used in classification. This leaves only the question of which *distance measure* to use, that is, how to decide the distance between any two insect sounds. Our empirical results showed that a simple algorithm which computes the Euclidean distance between the truncated frequency spectrums of the insect sounds works quite well. Our distance measure is further explained in Table 2. Given two flying sounds, we first transform each sound into frequency spectrums using DFT (Lines 1–2). The spectrums are then truncated to include only those corresponding to the frequency range from 100 to 2,000 (Lines 3–4); the frequency range is thus chosen, because according to entomological advice,⁴ all other frequencies are unlikely to be the result of insect activity, and probably reflect noise in the sensor. We then compute the Euclidean distance between the two truncated spectrums (Line 5) and return it as the distance between the two flying sounds.

Our flying-sounds-based insect classification algorithm is obtained by “plugging” the distance measure explained in Table 2 into the Bayesian classification framework outlined in Table 1. To demonstrate the effectiveness of the algorithm, we considered the data that was used to generate the plot in Fig. 2. These data were randomly sampled from a dataset with over 100,000 sounds generated by our sensor. We sampled in total 3,000 flying sounds, 1,000 sounds for each species, so the *prior* probability for each class is one-third. Using our insect classification algorithm with k set to eight, which was selected based on the validation result, we achieved an error rate of 12.43 % using leave-one-out. We then compared our algorithm to the *optimal* result possible using only the wingbeat frequency, which is the most commonly used approach in previous research efforts. The optimal Bayes error-rate to classify the insects using wingbeat frequency is 18.13 %, which is the lower bound for any algorithm that uses just that feature. This means that using the truncated frequency spectrum is able to reduce the error rate by almost a third. To the best of our knowledge, this is the first explicit demonstration that there is exploitable information in the flight sounds *beyond* the wingbeat frequency.

⁴ Many large insects, i.e. most members of Odonata and/or Lepidoptera, have wingbeat frequencies that are significantly slower than 100 Hz; our choice of truncation level reflects our special interest in Culicidae.

It is important to note that we do not claim that the distance measure we used in this work is optimal. There may be better distance measures, especially if we are confining our attention to *just* Culicidae or *just* Tipulidae, etc. However, if and when a better distance measure is found, we can simply “plug” the distance measure in the Bayesian classification framework to get a better classification performance.

Additional Features: Circadian Rhythm of Flight Activity and Geographic Distribution

In addition to the insect flight sounds, there are other features that can be used to reduce the error rate. The features can be very cheap to obtain, as simple as noting the *time-of-intercept*, yet the improvement can be significant.

It has long been noted that different insects often have different circadian flight activity patterns (Taylor 1969), and thus the time when a flying sound is intercepted can be used to help classify insects. Figure 5 shows the flight activity circadian rhythms of *Cx. stigmatosoma* (female), *Cx. tarsalis* (male), and *Ae. aegypti* (female). Those circadian rhythms were learned based on hundreds of thousands of individual observations collected over one month. Note that although all three species are most active at dawn and dusk, *Ae. aegypti* females are significantly more active during daylight hours. Thus, if an unknown insect sound is captured at noon, it is more probable to be produced by an *Ae. aegypti* female than by a *Cx. tarsalis* male based on this *time-of-intercept* information.

A detailed description on how to incorporate new features into a Bayesian classifier can be found in (Chen et al. 2014). To demonstrate the benefit of incorporating the additional feature, we again revisit the toy example in Fig. 2. With the *time-of-intercept* feature incorporated and the accurate flight activity circadian rhythms learned using our sensor data, we achieve a classification accuracy of 95.23 %. Recall that the classification accuracy using *just* the *insect-sound* is 87.57 % (cf. the paragraph right below Table 2). Simply by incorporating this cheap-to-obtain feature, we reduce the classification error rate by about two-thirds, from 12.43 % to only 4.77 %.

In addition to the *time-of-intercept*, we can also use the *location-of-intercept* as an additional feature to reduce classification error rate. The *location-of-intercept* is also very cheap-to-obtain., which is simply the location where the sensor is deployed, yet it carries useful information for classification because insects are rarely evenly distributed at any spatial granularity we consider.

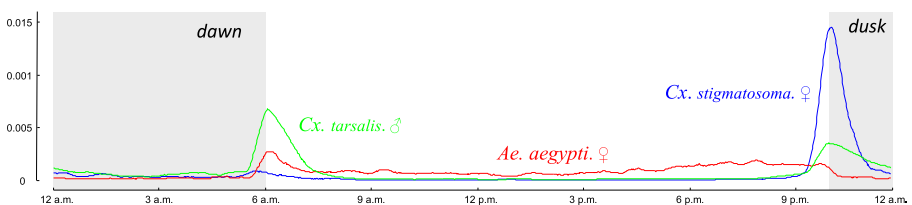


Fig. 5 The flight activity circadian rhythms of *Cx. stigmatosoma* (female), *Cx. tarsalis* (male), and *Ae. Aegypti* (female), learned based on observations generated by our sensor that were collected over one month

A General Framework for Adding Features

There may be dozens of additional features that could help improve the classification performance. In this section, we generalize our classifier to a framework that is easily extendable to incorporate arbitrarily many features.

With n independent features, the posterior probability that an observation belongs to a class C_i is calculated as:

$$P(C_1 | F_1 = f_1, F_2 = f_2, \dots, F_n = f_n) \propto P(C_1) \prod_{j=1}^n P(F_j = f_j | C_1) \quad (2)$$

Where $P(F_j = f_j | C_i)$ is the probability of observing f_i in class C_i

Note that the posterior probability can be calculated incrementally as the number of features increases. That is, if we have used some features to classify the objects, and later on, we have discovered more useful features and would like to add those new features to the classifier to re-classify the objects, we do not have to re-compute the entire classification from scratch. Instead, we can keep the posterior probability obtained from the previous classification (based on the old features), update each posterior probability by multiplying it with the corresponding class-conditioned probability of the new features, and re-classify the objects using the new posterior probabilities.

In our discussions thus far, we have assumed that all the features are independent given the class. In (Chen et al. 2014), it was shown that this independence assumption is reasonable for the Bayesian classifier to work well. However, it is also possible that users may wish to use features that clearly violate the independence assumption in our general framework. For example, if the sensor was augmented to obtain *insect mass* (a generally useful feature), it is clear from basic principles of allometric scaling that the frequency spectrum feature would *not* be independent (Deakin 2010). The good news is that as shown in Fig. 6, the Bayesian network can be generalized to encode the dependencies among the features. In the cases where there is clear dependence between some features, we can consider adding an arrow between the dependent features to represent this dependence. For example, suppose there is dependence between features F_2 and F_3 , we can add an arrow between them, as shown by the

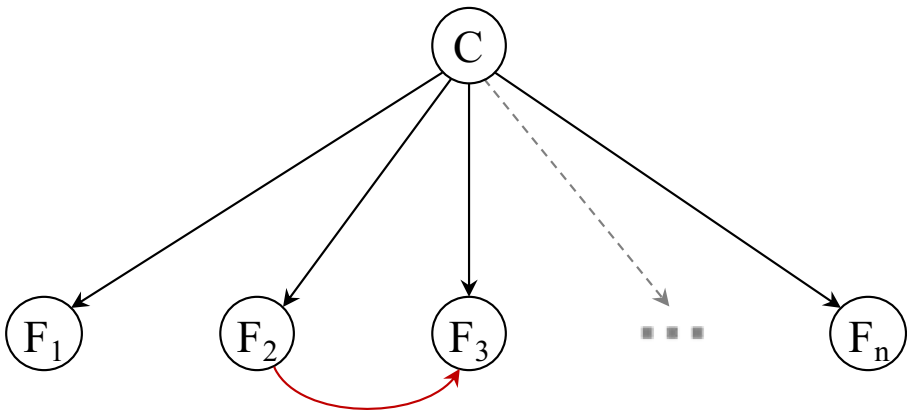


Fig. 6 The Bayesian network that uses n features for classification, with feature F_2 and F_3 being conditionally dependent

red arrow in Fig. 6. The direction of the arrow represents *causality*. The only drawback to this augmented Bayesian classifier (Keogh and Pazzani 1999) is that more training data is required to learn the classification model if there are feature dependences, as more distribution parameters need to be estimated (e.g., the covariance matrix is required instead of just the standard deviation) .

A Case Study: Sexing Mosquitoes

Sexing mosquitoes is required in some entomological applications. For example, the Sterile Insect Technique, a method which eliminates large populations of breeding insects by releasing only sterile males into the wild, has to separate the male mosquitoes from the females before being released (Papathanos et al. 2009). Here, we conducted an experiment to see how well it is possible to distinguish female and male mosquitoes from a single species using our proposed classifier.

In this experiment, we would like to distinguish male *Ae. aegypti* mosquitoes from females. The only feature used in this experiment is the *frequency spectrum*. We did not use the *time-of-intercept*, as there is no obvious difference between the flight activity circadian rhythms of the males and the females that belong to a same species (A recent paper offers evidence of minor, but measurable differences for the related species *Anopheles gambiae* (Rund et al. 2012); however, we ignore this possibility here for simplicity). The data used were randomly sampled from a pool of over 20,000 exemplars. We varied the number of exemplars from each sex from 100 to 1,000 and averaged over 100 runs, each time using random sampling with replacement. The average classification performance using leave-one-out cross validation is shown in Fig. 7.

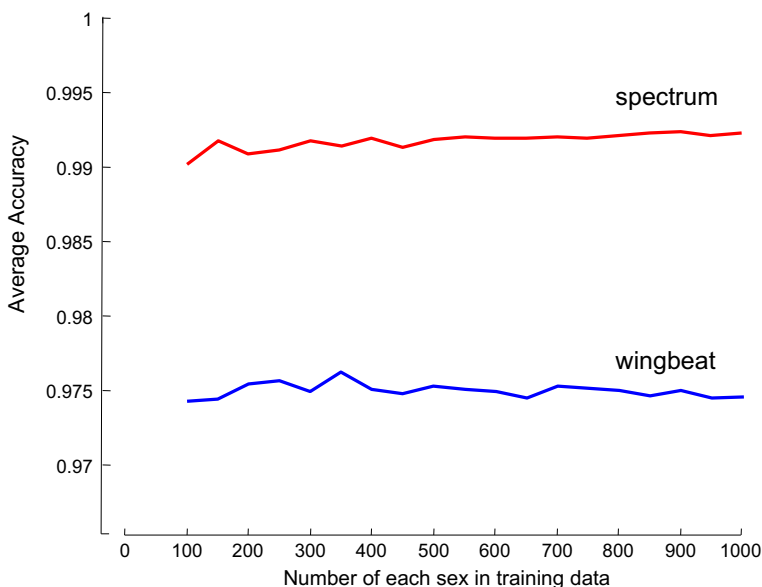


Fig. 7 The classification accuracy of sex discrimination of *Ae. aegypti* mosquitoes with different numbers of training data using our proposed classifier and the wingbeat-frequency-only classifier

We can see that our classifier is quite accurate in sex separation. With 1,000 training data for each sex, we achieved a classification accuracy of 99.22 % using *just* the truncated frequency spectrum. That is, if our classifier is used to separate 1,000 mosquitoes, we will make about eight misclassifications. Note that, as the amount of training data increases, the classification accuracy increases. This is an additional confirmation of the claim that more data improves classification (Halevy et al. 2009).

We compared our classifier to the classifier using just the wingbeat frequency. As shown in Fig. 7, our classifier consistently outperforms the wingbeat frequency classifier across the entire range of the number of training data. The classification accuracy using the wingbeat classifier was 97.47 % if there are 1,000 training data for each sex. Recall that the accuracy using our proposed classifier was 99.22 %. By using the frequency spectrum instead of the wingbeat frequency, we reduced the error rate by more than two-thirds, from 2.53 % to 0.78 %. It is important to recall that in this comparison, the data and the basic classifier were *identical*; thus, all the improvement can be attributed to the additional information available in the frequency spectrum beyond just the wingbeat frequency. This offers additional evidence for our claim that wingbeat frequency by itself is insufficient for accurate classification.

In this experiment, we assume the cost of female misclassification (misclassifying a female as a male) is the same as the cost of male misclassification (misclassifying a male as a female). The confusion matrix of classifying 2,000 mosquitoes (equal size for each sex) with the same cost assumption from one experiment is shown in Table 3. I.

However, there are cases in which the misclassification costs are asymmetric. For example, when the Sterile Insect Technique is applied to mosquito control, failing to release an occasional male mosquito because we mistakenly thought it was a female does not matter too much. In contrast, releasing a female into the wild is a more serious mistake, as it is only the females that pose a threat to human health. In the cases where we have to deal with asymmetric misclassification costs, we can change the decision boundary of our classifier to lower the number of high-cost misclassifications in a principled manner. Of course, there is *no free lunch*, and a reduction in the number of high-cost misclassifications will be accompanied by an increase in the number of low-cost misclassifications.

In the previous experiment, with equal misclassification costs, an unknown insect is predicted to belong to the class that has the higher posterior probability. This is the equivalent of saying the threshold to predict an unknown insect as female is 0.5. That is, only when the posterior probability of belonging to the class of females is larger than

Table 3 (I) The confusion matrix for sex discrimination of *Ae. aegypti* mosquitoes with the decision threshold for female being 0.5 (i.e., same cost assumption). (II) The confusion matrix of sexing the same mosquitoes with the decision threshold for female being 0.1

		Predicted class	
		female	male
Actual class	female	993	7
	male	5	995

		Predicted class	
		female	male
Actual class	female	1,000	0
	male	22	978

0.5 will an unknown insect be predicted as a female. Equivalently, we can replace Line 19 in Table 1 with the code in Table 4 by setting the threshold to 0.5.

We can change the threshold to minimize the *total* cost when the costs of different misclassifications are different. In the Sterile Insect Technique, the goal is to reduce the number of female misclassifications. This can be achieved by lowering the threshold required to predict an exemplar to be female. For example, we can set the threshold to be 0.1, so that if the probability of an unknown exemplar belonging to a female is no less than this value, it is predicted as a female. While changing the threshold may result in a lower overall accuracy, as more males will be misclassified as females, it reduces the number of females that are misclassified as male. By examining the experiment summarized in Table 3. I, we can predict that by setting the threshold to be 0.1, we reduce the female misclassification rate to 0.075 %, with the male misclassification rate rising to 0.69 %. We chose this threshold value because it gives us an approximately one in a thousand chance of releasing a female. However, any domain specific threshold value can be used; the practitioner simply needs to state her preference in one of two intuitive and equivalent ways: “*What is the threshold that gives me a one in (some value) chance of misclassifying a female as a male*” or “*For my problem, misclassifying a male as a female is (some value) times worse than the other type of mistake, what should the threshold be?*” (Elkan 2001).

We applied our 0.1 threshold to the data which was used to produce the confusion matrix shown in Table 3. I and obtained the confusion matrix shown in Table 3. II. As we can see, of 2,000 insects in this experiment, twenty-two males, and *zero* females where misclassified, numbers in close agreement to theory.

Experiment: Insect Classification with Increasing Number of Species

When discussing our sensor/algorithm, we are invariably asked, “*How accurate is it?*” The answer to this depends on the insects to be classified. For example, if the classifier is used to distinguish *Cx. stigmatosoma* (female) from *Cx. tarsalis* (male), it can achieve near perfect accuracy as the two classes are radically different in their wingbeat sounds; whereas when it is used to separate *Cx. stigmatosoma* (female) from *Ae. aegypti* (female), the classification accuracy will be much lower, given that the two species have quite similar sounds, as hinted at in Fig. 2. Therefore, a single absolute value for classification accuracy will not give the reader a good intuition about the performance of our system. Instead, in this section, rather than reporting our classifier’s accuracy on a fixed set of insects, we applied our classifier to datasets with an incrementally increasing number of species and therefore increasing classification difficulty.

Table 4 The decision making policy for the sex separation experiment

```

if (  $P(\text{female}|X) \geq \text{threshold}$  )
  X is a female
else
  X is a male
end

```

We began by classifying just two species of insects; then at each step, we added one more species (or a single *sex* of a sexually dimorphic species) and used our classifier to classify the increased number of species. We considered a total of ten classes of insects (different sexes from the same species counting as different classes), 5,000 exemplars in each class. Our classifier used both *insect-sound* (frequency spectrum) and *time-of-intercept* for classification. The classification accuracy measured at each step and the relevant class added is shown in Table 5. Note that the classification accuracy at each step is the accuracy of classifying all the species that come *at* and *before* that step. For example, the classification accuracy at the last step is the accuracy of classifying *all* ten classes of insects.

As we can see, our classifier achieves more than 96 % accuracy when classifying no more than five species of insects, significantly higher than the default rate of 20 % accuracy. Even when the number of classes considered increases to ten, the classification accuracy is never lower than 79 %, again significantly higher than the default rate of 10 %. Note that the ten classes are not easy to separate, even by human inspection. Among the ten species, eight of them are mosquitoes; six of them are from the same genus.

The Utility of Automatic Insect Classification

The reader may already appreciate the utility of automatic insect classification. However, for completeness, we give some examples of how the technology may be used.

- Electrical Discharge Insect Control Systems EDICS (“bug zappers”) are insect traps that attract and then electrocute insects. They are very popular with consumers who are presumably gratified by the characteristic “buzz” produced when an insect is electrocuted. While most commercial devices are sold as mosquito deterrents, studies have shown that as little as 0.22 % of the insects killed are mosquitoes (Frick and Tallamy 1996). This is not surprising, since the attractant is typically just an ultraviolet light. Augmenting the traps with CO₂ or other chemical attractants helps, but still allows the needless electrocution of beneficial insects. ISCA technologies (owned by author A. M-N) is experimenting with building a “smart trap” that classifies insects as they approach the trap, selectively killing the target insects but blowing the non-target insects away with compressed air.

Table 5 Classification accuracy with increasing number of classes

Step	Species added	Classification accuracy	Step	Species added	Classification accuracy
1	<i>Ae. aegypti</i> ♂	N/A	6	<i>Cx. quinquefasciatus</i> ♂	92.69%
2	<i>Musca domestica</i>	98.99%	7	<i>Cx. stigmatosoma</i> ♀	89.66%
3	<i>Ae. aegypti</i> ♀	98.27%	8	<i>Cx. tarsalis</i> ♂	83.54%
4	<i>Cx. stigmatosoma</i> ♂	97.31%	9	<i>Cx. quinquefasciatus</i> ♀	81.04%
5	<i>Cx. tarsalis</i> ♀	96.10%	10	<i>Drosophila simulans</i>	79.44%

- As noted above, the Sterile Insect Technique has been used to reduce the populations of certain target insects, most notably with Screwworm flies (*Cochliomyia hominivorax*) and the Mediterranean fruit fly (*Ceratitis capitata*). The basic idea is to release sterile males into the wild to mate with wild females. Because the males are sterile, the females will lay eggs that are either unfertilized, or produce a smaller proportion of fertilized eggs, leading to population declines and eventual eradication in certain areas. (Benedict and Robinson 2003). Note that it is important not to release females, and sexing mosquitoes is notoriously difficult. Researchers at the University of Kentucky are experimenting with our sensors to create insectaries from which only male hatchlings can escape. The idea is to use a modified EDICS or a high powered laser that selectively turns on and off to allow males to pass through, but kills the females.
- Much of the research on insect behavior with regard to color, odor, etc., is done by having human observers count insects as they move in dual choice olfactometer or on landing strips etc. For example, (Cooperband et al. 2013) notes, “*Virgin female wasps were individually released downwind and the color on which they landed was recorded* (by a human observer).” There are several problems with this: human time becomes a bottleneck in research; human error is a possibility; and for some host seeking insects, the presence of a human nearby may affect the outcome of the experiment (unless costly isolation techniques/equipment is used). We envision our sensor can be used to accelerate such research by making it significantly cheaper to conduct these types of experiments. Moreover, the unique abilities of our system will allow researchers to conduct experiments that are currently impossible. For example, a recent paper (Rund et al. 2012) attempted to see if there are sex-specific differences in the daily flight activity patterns of *Anopheles gambiae* mosquitoes. To do this, the authors placed individual sexed mosquitoes in small glass tubes to record their behavior. However, it is possible that both the small size of the glass tubes and the fact that the insects were in isolation affected the result. Moreover, even the act of physically sexing the mosquitoes may affect them due to metabolic stress etc. In contrast, by using our sensors, we can allow unsexed pupae to hatch out and the adults fly in cages with order of magnitude larger volumes. In this way, we can automatically and noninvasively sex them to produce sex-specific daily flight activity plots.

Conclusion and Future Work

In this work we have introduced a sensor/classification framework that allows the inexpensive and scalable classification of flying insects. We have shown experimentally that the accuracies achievable by our system are good enough to allow the development of commercial products and to be a useful tool for entomological research. To encourage the adoption and extension of our ideas, we are making all code, data, and sensor schematics freely available at the UCR Computational Entomology Page (Chen 2013). Moreover, within the limits of our budget, we will continue our practice of giving a complete system (as shown in Fig. 2) to any research entomologist who requests one.

Acknowledgments We would like to thank the Vodafone Americas Foundation, the Bill and Melinda Gates Foundation and São Paulo Research Foundation (FAPESP) for funding this research, and the many faculties from the Department of Entomology at UCR that offered advice and expertise.

References

- Banko M, Brill E (2001) Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. Proceedings of the first international conference on Human language technology research (pp. 1–5). Association for Computational Linguistics
- Batista GE, Keogh EJ, Mafra-Neto A, Rowton E (2011) SIGKDD demo: sensors and software to allow computational entomology, an emerging application of data mining. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 761–764
- Belton P, Costello RA (1979) Flight sounds of the females of some mosquitoes of Western Canada. *Entomologia experimentalis et applicata* 26(1):105–114
- Benedict M, Robinson A (2003) The first releases of transgenic mosquitoes: an argument for the sterile insect technique. *TRENDS in Parasitology* 19(8):349–355, Accessed March 8, 2012
- Boll S (1979) Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27 (2), 113–120
- Capinera, JL (2008). Encyclopedia of entomology. Springer. Epsky ND, Morrill WL, Mankin R (2005) Traps for capturing insects. In Encyclopedia of Entomology, pp. 2319–2329. Springer Netherlands
- Chen Y (2013) Supporting Materials <https://sites.google.com/site/insectclassification/>
- Chen Y, Hu B, Keogh E, Batista GE (2013) DTW-D: time series semi-supervised learning from a single example. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 383–391
- Chen Y, Why A, Batista G, Mafra-Neto A, Keogh E (2014) supporting technique report <http://arxiv.org/abs/1403.2654>
- Cooperband MF, Hartness A, Lelito JP, Cosse AA (2013) Landing surface color preferences of *Spathius agrili* (Hymenoptera: Braconidae), a parasitoid of emerald ash borer, *Agrilus planipennis* (Coleoptera: Buprestidae). *J Insect Behav* 26(5):721–729
- Deakin MA (2010) Formulae for insect wingbeat frequency. *Journal of Insect Science*, 10 (96):1
- Devroye L (1996) A probabilistic theory of pattern recognition. Springer Vol 31
- Elkan, C (2001) The foundations of cost-sensitive learning. In international joint conference on artificial intelligence, vol. 17, No. 1, pp. 973–978. LAWRENCE ERLBAUM ASSOCIATES LTD.
- Ephraim Y, Malah D (1984) Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 32(6):1109–1121
- Frick TB, Tallamy DW (1996) Density and diversity of non-target insects killed by suburban electric insect traps. *Entomological News* 107:77–82
- Fukunaga K (1990) Introduction to statistical pattern recognition. Online via Elsevier, Access
- Georghiou GP, Wirth MC (1997) Influence of exposure to single versus multiple toxins of *Bacillus thuringiensis* subsp. *israelensis* on Development of Resistance in Mosquito *Culex quinquefasciatus* (Diptera: Culicidae). *Appl Environ Microbiol* 63(3):1095–1101
- Grimaldi RP (1989) Discrete and Combinatorial Mathematics: An Applied Introduction 2nd Ed. Addison-Wesley Longman Publishing Co., Inc
- Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. *IEEE Intell Syst* 24(2):8–12
- Hao Y, Campana B, Keogh EJ (2012) Monitoring and mining animal sounds in visual space. *J Insect Behav* 1–28
- Kahn MC, Celestin W, Offenhauser W (1945) Recording of sounds produced by certain disease-carrying mosquitoes. *Science* 101:335–336
- Kahn MC, Offenhauser W (1949) The identification of certain West African mosquitos by sound. *Amer J trop Med* 29:827–836
- Keogh E, Pazzani M (1999) Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. In Proceedings of the seventh international workshop on artificial intelligence and statistics. pp. 225–230.
- Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (Vol. 14, No. 2, pp. 1137–1145

- Li Z, Zhou Z, Shen Z, Yao Q (2009) Automated identification of mosquito (diptera: Culicidae) wingbeat waveform by artificial neural network. *Artificial Intell Applications and Innovations* 187(2009):483–489
- Mack YP, Rosenblatt M (1979) Multivariate k-nearest neighbor density estimates. *J Multivar Anal* 9(1):1–15
- Mankin RW, Machan R, Jones R (2006) Field testing of a prototype acoustic device for detection of Mediterranean fruit flies flying into a trap. *Proc. 7th Int. Symp. Fruit Flies of Economic Importance*, pp. 10–15
- Moore A (1991) Artificial neural network trained to identify mosquitoes in flight. *J Insect Behav* 4(3):391–396
- Moore A, Miller RH (2002) Automated identification of optically sensed aphid (Homoptera: Aphidae) wingbeat waveforms. *Ann Entomol Soc Am* 95:1–8
- Moore A, Miller JR, Tabashnik BE, Gage SH (1986) Automated identification of flying insects by analysis of wingbeat frequencies. *J Econ Entomol* 79:1703–1706
- Papathanos PA, Bossin HC, Benedict MQ, Catteruccia F, Malcolm CA, Alphey L, Crisanti A (2009) Sex separation strategies: past experience and new approaches. *Malar J* 8(Suppl 2)
- Prechelt L (1995) A quantitative study of neural network learning algorithm evaluation practices. In *proceedings of the 4th Int'l Conference on Artificial Neural Networks*. pp. 223–227
- Raman DR, Gerhardt RR, Wilkerson JB (2007) Detecting insect flight sounds in the field: Implications for acoustical counting of mosquitoes. *Trans ASABE* 50(4):1481
- Reed SC, Williams CM, Chadwick LE (1942) Frequency of wing-beat as a character for separating species races and geographic varieties of *Drosophila*. *Genetics* 27:349–361
- Repasky KS, Shaw JA, Scheppele R, Melton C, Carsten JL, Spangler LH (2006) Optical detection of honeybees by use of wing-beat modulation of scattered laser light for locating explosives and land mines. *Appl Opt* 45:1839–1843
- Rund SSC, Lee SJ, Bush BR, Duffield GE (2012) Strain- and sex-specific differences in daily flight activity and the circadian clock of *Anopheles gambiae* mosquitoes. *J Insect Physiol* 58:1609–19
- Sawedal L, Hall R (1979) Flight tone as a taxonomic character in Chironomidae (Diptera). *Entomol Scand Suppl* 10:139–143
- Schaefer GW, Bent GA (1984) An infra-red remote sensing system for the active detection and automatic determination of insect flight trajectories (IRADIT). *Bull Entomol Res* 74:261–278
- Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R (2013) Real-time human pose recognition in parts from single depth images. *Commun ACM* 56(1):116–124
- Sotavalta O (1947) The flight-tone (wing-stroke frequency) of insects (Contributions to the problem of insect flight I.). *Acta Entomol Fenn* 4:1–114
- Taylor B (1969) Geographical range and circadian rhythm. *Nature* 222:296–297
- Tsymbol A (2004) The problem of concept drift: definitions and related work. Computer Science Department, Trinity College Dublin
- Unwin DM, Ellington CP (1979) An optical tachometer for measurement of the wing-beat frequency of free-flying insects. *J Exp Biol* 82(1):377–378
- Van Dam AR, Walton WE (2008) The effect of predatory fish exudates on the ovipositional behavior of three mosquito species: *Culex quinquefasciatus*, *Aedes aegypti* and *Culex tarsalis*. *J Med Vet Entomol* 22(4):399–404
- Vapnik VN, Chervonenkis AY (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications* 16(2):264–280
- Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. *Mach Learn* 23(1):69–101
- Zhan C, Lu X, Hou M, Zhou X (2005) A lqv-based neural network anti-spam email approach. *ACM SIGOPS Oper Syst Rev* 39 (1):34–39 ISSN 0163–5980