2014-10

Signal classification by similarity and feature extraction allows an important application in insect recognition

Joint Conference on Robotics and Intelligent Systems; Best DSc Theses / MSc Dissertations Contest in Artificial and Computational Intelligence, 9th, 2014, São Carlos.
http://www.producao.usp.br/handle/BDPI/48593

Preface

This volume contains the papers presented at CTDIAC 2014: the 9th Best DSc Theses / MSc Dissertations Contest in Artificial and Computational Intelligence, held on October 17-22, 2014 in São Carlos/SP, as a satellite event of JCRIS 2014.

There were 32 submissions. Each submission was reviewed by at least 2 program committee members. There were 5 DSc thesis papers and 6 MSc dissertation papers selected to the second phase of CTDIAC.

We greatly appreciate the members of the program committee that helped us to select a set of high quality papers. We also thank the JCRIS and BRACIS committees for helping us with the CTDIAC organization.

Anne Canuto
Graçaliz Dimuro
Aline Paes
Karina Valdivia

Table of Contents

Signal Classification by Similarity and Feature Extraction Allows an Important Application in Insect Recognition

Diego F. Silva, Gustavo E. A. P. A. Batista
ICMC/USP – São Carlos, Brazil
{diegofsilva,gbatista}@icmc.usp.br

Abstract

Insects have a strong relationship with the humanity, in both positive and negative ways. It is estimated that insects, particularly bees, pollinate at least two-thirds of all food consumed in the world. In contrast, mosquito borne diseases kill millions of people every year. Due to such a complex relationship, insect control attempts must be carefully planned. Otherwise, there is the risk of eliminating beneficial species, such as the recent threat of bee extinction. We are developing a novel sensor as a tool to control disease vectors and agricultural pests. This sensor captures insect flight information using laser light and classify the insects according to their species. Therefore, the sensor will provide real-time population estimates of species. Such information is the key to enable effective alarming systems for outbreaks, the intelligent use of insect control techniques, such as insecticides, and will be the heart of the next generation of insect traps that will capture only species of interest. In this paper, we demonstrate how we overtook the most important challenge to make this sensor practical: the creation of accurate classification systems. The sensor generates a very brief signal as result of the instant that the insect crosses the laser. Such events last for tenths of a second and have a very simple structure, consequence of the wings movements. Nevertheless, we managed to successfully identify relevant features using speech and audio analysis techniques. Even with the described challenges, we show that we can achieve an accuracy of 98% in the task of disease vector mosquitoes identification.

Student level: *MSc –* ***Date of conclusion:*** *2/27/2014 –* ***Examining board members:*** *João Luis Garcia Rosa, Alexandre Plastino de Carvalho and Estevam Rafael Hruschka Júnior –* ***Dissertation:*** *[1] –* ***Publications:*** *[2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]*

Keywords

Classification, Feature Extraction, Similarity, Time Series

1. Introduction

Insects have a strong relationship with the humanity, in both positive and negative ways. Mosquito borne diseases that kill millions of people every year. It is estimated that dengue, a disease transmitted by mosquitoes of the genus *Aedes*, affects between 50 and 100 million people every year and it is considered endemic in more than 100 countries [14]. Malaria, transmitted by mosquitoes of the genus *Anopheles*, affects around 6% of the world's population and it is estimated that there are over 200 million cases per year and about 7 million lethal cases in the last decade [15]. Another example are the insect pests that consume and destroy around US$40 billion worth of food each year [16].

In contrast, insects pollinate at least two-thirds of all the food consumed in the world, with bees alone responsible for pollinating one-third of this total [17]. Furthermore, many species have been used as bioindicators of environmental quality, since their presence/absence, distribution and density, define the quality of the ecosystem, especially in relation to contaminants in the air, soil and water [18].

Due to such a complex relationship, many researchers have developed several methods of insect control [19]. However, without the knowledge of the spatio-temporal distribution of the insects, the use of these techniques becomes costly and inefficient. One example is the recent threat of bee extinction due to insecticide exposure. Just in the summer of 2013, nearly half of the American commercial hives disappeared [20]. Although the exact reason for bee hive decline is a combination of factors, undoubtedly a central issue is the large scale pulverization of insecticides. Currently, studying the spatio-temporal distribution of insects is a costly and time consuming task. In general, insect counts are obtained with traps, usually adhesive, which are collected periodically and analyzed by experts who manually identify and count the collected species of insects.

We are developing a novel sensor as a tool to control disease vectors and agricultural pests. This sensor captures insect flight information using laser light and classify the insects according to their species. Therefore, the sensor will provide real-time population estimates of species. Such a sensor will enable effective alarming systems for outbreaks, the intelligent use of insect control techniques, such as insecticides, and will be the heart of the next generation of insect traps that will capture only species of interest.

In this paper, we demonstrate how we overtook the most important challenge to make this sensor practical: the creation of an accurate classification systems. The sensor generates a very brief signal as result of the instant that the insect crosses the laser. Such events last for tenths of a second and have a very simple structure, consequence of the wings movements. Nevertheless, we managed to successfully identify relevant features using speech and audio analysis techniques.

We show that, with the correct combination of feature extraction and machine learning techniques, we can achieve an accuracy of almost 90% in the task of identifying the correct insect species among other nine species with data collected by the sensor. More importantly, we show that we can achieve an accuracy of 98% in the task of correctly recognizing if a given event was generated by a disease vector mosquitoes.

The remaining of this paper is organized as follows. Section 2 describes the sensor used in this work, as well the data collecting procedure. Section 3 presents the main classification approaches explored in this research. Section 4 describes our experiments to evaluate the classification methods for automatic species identification. The results achieved in these experiments are briefly analyzed in Section 5. Finally, we conclude this paper in Section 6.

Since the research contributions of the dissertation related to this article transcend the limits of insect classification, we organized an appendix to briefly describe them.

2. Laser Insect Sensor

The main elements of the sensor are a laser beam and an array of phototransistors. When an insect crosses the laser beam, a variation of light is caused by partial occlusion of light due to the wings movements. Such a variation is stored as a short time series. Our main goal is to build a classification system that takes such a time series as input and provides counts of insects discriminated by species.

2.1.Sensor Description

The general design of the sensor used in this work is shown in Figure 1. It consists of a low-powered planar laser source pointed to an array of phototransistors. When a flying insect crosses the laser, its wings partially occlude the light, causing small light variations that are captured by the phototransistors. An electronic circuit board filters and amplifies the signal and the output is recorded by a digital sound recorder.

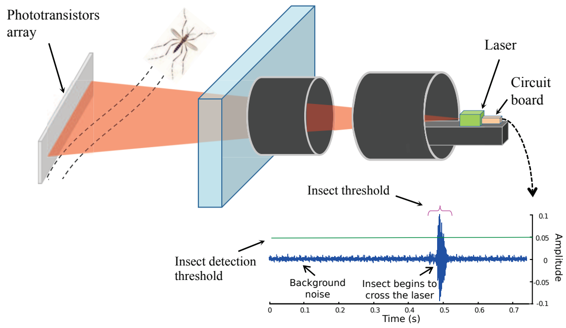

Figure 1. The logical design of the sensor. A planar laser light is directed at an array of phototransistors. When an insect flies across the laser, a light variation is registered by the phototransistors as a time series

The sensor signal is very similar to an audio signal captured by a microphone, even though the data are obtained optically. However, the sensor is totally deaf to any agent that does not cross the light; therefore, the sensor does not suffer any external interference such as bird sounds, cars, or airplane noise.

The data captured by the sensor are constituted, in general, of background noise with occasional "events", result of the brief moment that an insect flew across the laser. In the next section, we provide details about the procedure used to collect and preprocess the data used in this work.

2.2.Collecting and Preprocessing Data

After collecting the data, we preprocessed the recordings and detected the insect passages in raw data. We designed a detector responsible for identifying the events of interest and separating them from background noise. The general idea of the detector is to move a sliding window across the raw data and calculate the spectrum of the signal inside the window. As most insects have wing beat frequencies which range from 100Hz to 1000Hz, we used the maximum magnitude of the signal spectrum in this range as the detector confidence.

The detector uses a sliding window and calculates the magnitude of signal components within the window. Then, the maximum magnitude is taken as a confidence value for the detector. The larger the magnitude, the higher the confidence that the signal is not background noise. All signals with magnitude above a user-specified threshold are considered an event generated by an insect. The high signal to noise ratio of the data collected by the sensor allows the user to specify low values for the threshold without the risk of false positives. Figure 2 illustrates how the detector works.

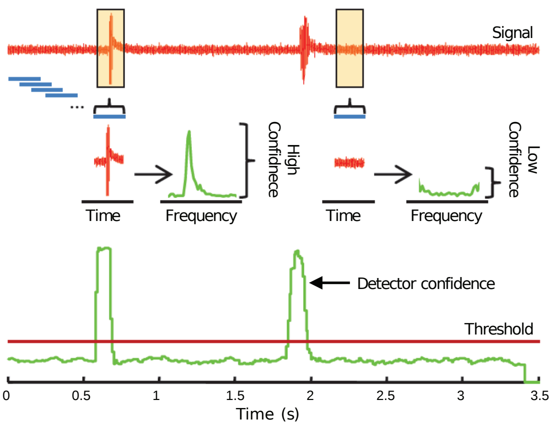

Figure 2. General design of the wing-beat detector [21]

The detector outputs audio fragments which usually last for a few tenths of a second and have at least one insect passage. Due to the simplicity of the design of the electronic circuit, there is some noise combined with the insect signals. So, we filtered most of the noise using a digital filter based on spectral subtraction, responsible for the removal of certain frequency ranges of signal [22]. An example of a filtered and segmented signal is shown in Figure 3.

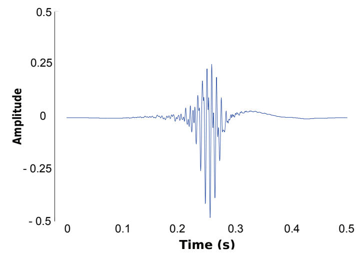

Figure 3. Example of a segmented and filtered signal. Species: *Aedes aegypti*

3. Signal Classification Approaches

In this section, we describe the main strategies explored for classifying the signals obtained by the sensor. First, we review the representations used in audio signal analysis. Then, we discuss the use of these representations in signal classification. A first use is the direct comparison of signals under different representations in a similarity-based classification approach. The second is a feature extraction approach that use machine learning systems to induce a classifier.

3.1.Signal Representations

Audio signals can be represented in several ways. The primitive signal representation describes the amplitude of its waveform at each moment of time. This representation is called temporal. A periodic signal, like a sine wave, can be simply described by its amplitude and period. However, a sine wave, as well as any other periodic wave, is not common in practical applications. Thus, the features present in the signal are not as simple as the amplitude and period as mentioned above. Therefore, other features are frequently used to summarize the signal such as mean amplitude, interval, zero-crossing rate, among others.

As an example, Figure 4 shows the temporal representation of a signal obtained by a recording of three seconds of a single note emitted by an acoustic bass.

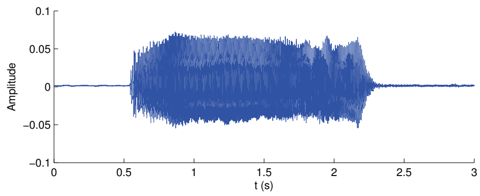

Figure 4. Temporal representation of a signal obtained by a single note emitted by an acoustic bass

Another way to revel important information about a signal is performing a change of representation. Many important features are evident in the frequency (spectral) domain. To obtain the frequency representation of a complex signal, it is necessary to break it down into a sum of simple wave forms, specifically of the family of sines and cosines. When a signal is periodic, this decomposition becomes a series of sinusoidal and cosinusoidal signals of different amplitudes and frequencies, called Fourier series. Formally, a Fourier series with period T is defined by Equation 1, where a_i and b_j are the weights of each cosine and sine component, respectively, and c is constant.

$$f(t) = \sum_{i=0}^{\infty} a_i \cos(\frac{2\pi it}{T}) + \sum_{j=0}^{\infty} b_j \sin(\frac{2\pi jt}{T}) + c \quad (1)$$

The Fourier series can be extended to non-periodic signals. This extension is called Fourier transform, the result describes a mapping of frequency components that form a signal involving frequency and amplitude of each harmonic. The transform calculation of a signal in continuous time requires this signal to be generated by a given equation. Since this is not possible in many contexts, we can estimate the frequency spectrum using the Discrete Fourier Transform (DFT). The most widely used method to calculate the DFT is the Fast Fourier Transform [23]. The application of such an algorithm on the signal shown in Figure 4 generates the frequency spectrum shown in Figure 5.

Finally, the cepstrum is the result of applying the Fourier transform of a spectrum in logarithmic form. Its independent variable is known as quefrency and, despite being a measure of time, has no direct relationship with the temporal representation of the signal, but the period, inverse of frequency. Originally, the cepstrum was proposed for analyzing seismic echoes of earthquakes and bombs [24]. Currently, the cepstral

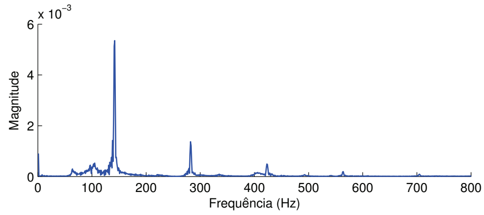

Figure 5. Spectral representation of the signal shown in Figure 4

features are also used in the field of audio analysis, achieving excellent results in areas such as speech and music analysis. Figure 6 shows the cepstrum obtained from the acoustic bass signal, shown in Figure 4.

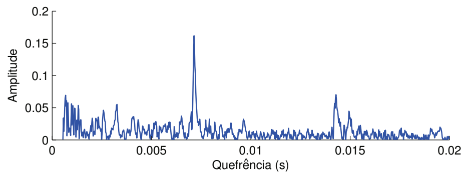

Figure 6. Cepstral representation of the signal shown in Figure 4

3.2.Similarity-based Classification

This classification approach assumes that similar series are more likely to belong to the same class. Given a query with unknown class, a distance measure is used to determine the similarity among the query and each labeled example in a training set. The label assigned to the query is the label of the most similar example (or the most frequent class among the most similar examples). Besides its simplicity, the similarity-based classification has demonstrated to be competitive with more complex classification methods.

The similarity-based classification depends of a distance measure and a data representation. There are dozens of distance measures in the literature which can be applied to signal comparison under the temporal, spectral and cepstral representations. In this research, we evaluated thirteen distance measures applied to the spectrum and the cepstrum of the signals. The time domain was not included here because the signals have different lengths and also because the results are very sensitive to the alignment of the signals. A more detailed discussion of this issue can be found in [2]. We refer the reader to [1], [11] for a detailed description of the similarity measures used in this research.

3.3. Temporal and Spectral Features

The second strategy for time series classification is use of machine learning classifiers with features extracted from the signals. Due to the similarity of the sensor signal with audio, we explored the most used features from audio and signal processing.

In this work, we use temporal and spectral features. The interested reader can find a detailed review of these features in [1], [25]. We use *temporal features* and *spectral features* to refer to feature vectors extracted from time and frequency domains, respectively. Table 1 lists the features that compose each of these vectors.

3.4. Mel-Frequency Cepstrum Coefficients

The Mel-Frequency Cepstrum Coefficients are probably the most commonly used attributes in speech processing tasks, such as speaker and speech recognition [27]. Briefly, to calculate those coefficients, we first take the magnitudes of frequency components using an acoustically-defined scale called *mel*, originated from the study of Stevens et al. [28], which relates physical frequencies to the frequencies perceived by the human auditory system. Next, we apply a Discrete Cosine Transform, widely used in data compression [29]. The MFCC are the cepstrum coefficients obtained from this operation. Equation 2 shows the conversion from frequency (f) to mel-frequency (m).

$$m = 2595 \times log_{10}(1 + \frac{f}{700}) \tag{2}$$

3.5. Linear Prediction and Line Spectral Frequencies

Linear Prediction (LP) is a technique used in many speech applications, such as recognition, compression and modeling for a long time [30]. LP is based on the fact that a speech signal can be described by Equation 3.

$$\hat{x}_k = \sum_{i=1}^{p} a_i x_{k-i} \tag{3}$$

where k is the time index and p is the order of LP – i.e., the number of employed LP coefficients (LPC). The a_i coefficients are calculated in order to minimize the prediction error using a covariance or auto-correlation method.

Equation 3 can be rewritten in the frequency domain with a $z-$transform [31]. In this way, a short segment of speech is assumed to be generated as the output of a filter $H(z) = 1/A(z)$, where $A(z)$ is the inverse filter such that:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4}$$

The Line Spectral Frequencies (LSF) representation, introduced by Itakura [32], is an alternative way to represent LP coefficients. In order to calculate LSF coefficients, the inverse filter polynomial is decomposed into two polynomials $P(z)$ and $Q(z)$:

$$P(z) = A(z) + z^{p+1}A(z^{-1}) \text{ and}$$
$$Q(z) = A(z) - z^{p+1}A(z^{-1})$$

where $P(z)$ is a symmetric polynomial and $Q(z)$ is an antisymmetric polynomial. The roots of $P(z)$ and $Q(z)$ determine the LSF coefficients.

LSF is well suited for quantization and interpolation [33]. Therefore, LSF can represent the speech signal, mapping a large signal to a small number of coefficients, better than other LP representations.

4. Experimental Results

In this section, we present experimental classification results using the strategies of similarity comparison and feature extraction.

4.1. Dataset description

In the experiments presented in this paper, we included four species of mosquitoes: *Aedes aegypti* (vector of filariasis, dengue, yellow fever, and West Nile virus), *Anopheles gambiae* (vector of malaria), *Culex quinquefasciatus* (vector of lymphatic filariasis) and *Culex tarsalis* (vector of St. Louis Encephalitis and Western Equine Encephalitis); three species of flies: *Drosophila melanogaster* also known as fruit fly, *Musca domestica* or house fly and *Psychodidae diptera* popularly known as moth fly; the beetle *Cotinis*

Table 1. List of features that compose temporal and spectral feature vectors

Domain	Feature
Temporal	Mean amplitude, Root mean square, Short-time energy, Interval, Temporal centroid, Zero-crossing rate Complexity estimate [26], Variance, Standard deviation, Skewness, Kurtosis, Duration
Spectral	Fundamental frequency, Inharmonicity, Tristimulus 1, Tristimulus 2, Tristimulus 3, Flux, Spectral centroid, Spectral irregularity Modified spectral irregularity, Variance, Standard deviation, Skewness, Kurtosis, Mean magnitude, Energy, Roll-off, Flatness

mutabilis and the bee *Apis mellifera*. The number of examples of each species varies between 172 (0.95%) and $5,309$ (29.31%), for the species *Cotinis mutabilis* and *Culex tarsalis*, respectively.

The data set was divided into standard training and test partitions. This division was performed in a stratified approach, leaving 33% of the examples in the training set and the remaining in the test set. Table 2 summarize this dataset.

Table 2. Summary of the data used in the experimental evaluation

Species	Instances	Distribution (%)
Aedes aegypti	4756	26.25
Anopheles gambiae	1411	7.79
Apis mellifera	511	2.82
Cotinis mutabilis	172	0.95
Culex quinquefasciatus	3137	17.32
Culex tarsalis	5309	29.31
Drosophila melanogaster	777	4.29
Musca domestica	1343	7.41
Psychodidae diptera	699	3.86
Total	18151	100.00

4.2. Similarity-based Classification

We start our analysis by comparing the use of four widely known distances, Euclidean, Manhattan, Cosine and Correlation, applied to cepstrum and spectrum [11]. Table 3 present the results.

Table 3. Classification results for similarity over the spectrum and the cepstrum

Distance	Accuracy (%)	
Measure	Spectrum	Cepstrum
Euclidean	76.14	78.66
Manhattan	80.09	67.24
Cosine	77.25	76.29
Correlation	76.60	75.34

The results achieved by similarity on the spectrum were slightly superior then the ones obtained by the cepstrum. Given these results, we decided to extend the evaluation of the classification by similarity only to the spectral domain. Table 4 presents the results for nine additional distance measures, Canberra, Chebyshev, Jaccard, Topsoe, Clark, Average L_1 L_∞, Squared χ^2, Additive Symmetric χ^2 and DTW with band-width of 5 observations.

Table 4. Classification results for nine additional similarity measures in the frequency domain

Distance Measure	Accuracy (%)
Canberra	72.28
Chebyshev	71.20
Jaccard	77.26
Topsoe	81.54
Clark	75.59
Average L_1 L_∞	80.09
Squared χ^2	81.38
Additive Symmetric χ^2	81.01
DTW (band-width = 5 observations)	81.04

4.3. Feature Extraction

The feature extraction approach uses different representations of signals to identify features, which are used as input to machine learning algorithms [6], [5].

We use Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Coefficients (LPC) and Line-Spectral Frequencies (LSF). Certain feature sets, such as MFCC, use a scale based on the human perception of sound. However, there is no *a priori* reason to limit our approach to the limited frequency range and resolution of human hearing. To circumvent this issue, we also evaluated the Linear-Frequency Cepstrum (LFC) and the Log-Linear Frequency Cepstrum (LLFC).

We evaluate several machine learning techniques using these features. Most learning algorithms have parameters that can significantly influence their performance. Our first experiment consists of a search for the parameters that maximize classification accuracy. Since the use of test data is restricted to the final classifiers evaluation, we used 10-fold cross-validation on the training data to search the parameter values. For each possible combination of parameter values, the accuracy of the classifier was measured in the "internal" cross-validation test sets. We use the best combination of parameter values for a given learning algorithm as the final setting, then use this combination to learn over the entire training set and evaluate the resulting classifier on the test set.

In the case of Support Vector Machine, we use grid search [34] to vary the parameters of the base algorithm and of the kernel. Given values of minimum, maximum and step size, we evaluate the cross-validation accuracy of each combination of parameters. This search is performed with coarse estimate, using 2-fold cross-validation. The search is then refined in regions with better results.

The learning algorithms, as well as parameter ranges, are described in Table 5.

Table 5. Learning algorithms with their respective parameter ranges

Algorithm	Parameters range (initial:step:final)
Decision Tree (J48 implementation)	Pruning factor P = 0.1:0.1:0.5
Gaussian Mixture Models (GMM)	Number of components N = 3:2:21
K-Nearest Neighbors (KNN)	Number of neighbors K = 1:2:25
Naïve Bayes (NB)	-
Random Forest (RF)	Number of trees N = 5:2:75
Support Vector Machine Poly. kernel (SVM Poly)	Complexity C = 10^i, i = -7:1:5 Poly. Degree D = 1:1:3
Support Vector Machine RBF kernel (SVM RBF)	Complexity C = 10^i, i = -7:1:5 $\gamma = 10^i$, i = -4:1:0

Table 6 presents the results of the first experiment. For reasons of readability, we omit results obtained by Naïve Bayes and J48 classifiers, since they achieved the worst results across all feature sets. Additionally, we only show the results for SVM RBF since SVM Poly had inferior results.

The best results were obtained with MFCC, being that LFC and LSF achieved slightly lower accuracy rates, and the spectral feature set and LLFC also slightly lower. The results obtained with temporal features and LPC were substantially lower than the other features. The best single classifier performance, 87.33%, was obtained with the SVM RBF classifier applied to MFCC, and seems to be a respectable accuracy rate given the complexity of the application. The best result obtained by similarity search was 81.87%.

The second experiment investigates how to combine the output of distinct classifiers. Classifiers in-

Table 6. Accuracy results per classifier and feature set with the optimal parameter values. The best result in each feature set is highlighted

Feature Set	Algorithm	Selected Parameter Configuration	Acc. (%)
LFC	KNN	#c= 75. $k = 7$	81.71
	RF	#c= 80. $T = 75$	83.49
	SVM RBF	#c= 95. $c = 10.\gamma = 1$	**86.93**
	GMM	#c= 100. $G = 9$	83.17
LLFC	KNN	#c= 15. $k = 7$	74.70
	RF	#c= 20. $T = 60$	76.30
	SVM RBF	#c= 70. $c = 10^4.\gamma = 0.01$	**79.05**
	GMM	#c= 20. $G = 17$	74.03
MFCC	KNN	#c= 30. $k = 5$	83.61
	RF	#c= 35. $T = 75$	85.39
	SVM RBF	#c= 40. $c = 10.\gamma = 1$	**87.33**
	GMM	#c= 45. $G = 13$	82.42
LPC	KNN	#c= 45. $k = 21$	56.18
	RF	#c= 65. $T = 75$	60.90
	SVM RBF	#c= 45. $c = 10^5.\gamma = 0.1$	**66.85**
	GMM	#c= 40. $G = 19$	54.15
LSF	KNN	#c= 95. $k = 5$	80.23
	RF	#c= 95. $T = 75$	84.25
	SVM RBF	#c= 100. $c = 10.\gamma = 1$	**84.97**
	GMM	#c= 75. $G = 17$	75.28
Temporal	KNN	$k = 11$	50.91
	RF	$T = 75$	60.13
	SVM RBF	$c = 10^5.\gamma = 0.1$	**60.62**
	GMM	$G = 19$	42.76
Spectral	KNN	$k = 5$	70.51
	RF	$T = 50$	**79.38**
	SVM RBF	$c = 10^5.\gamma = 0.1$	76.24
	GMM	$G = 21$	63.73

duced by different algorithms or different feature sets may make errors on different examples. Thus, we can construct classifier ensembles to explore this diversity. We evaluated three different strategies to combine the results. The first and simplest is voting: each classifier votes for the predicted class and the final answer is given by the class with the highest number of votes. In case of a tie, the class with highest prior probability is chosen.

The other two strategies use sum and product functions on the output score of each classifier. One possible advantage of these strategies in relation to voting is that they consider the fact that classifiers can assign similar score values to different classes when an object is close to borderline regions. Therefore, the classification of borderline cases can potentially benefit from the classifiers combination.

Table 7 presents the results of ensembles of different algorithms using the same feature set. We show two results for each feature set: the first one obtained with the combination of all the four classifiers in Table 6 for a given feature set; and the second obtained by combining only the three best classifiers.

Table 7. Results achieved by the combination of different classifiers on the same feature set. The highlighted results represent the accuracy gain over the best base classifier

Feature Set	Best Acc.	Combined Algorithms	Accuracy (%) Sum	Prod.	Voting
LFC	86.93	SVM RBF, KNN, GMM, RF	84.86	84.70	86.07
		SVM RBF, GMM, RF	83.58	83.94	86.29
LLFC	79.05	SVM RBF, KNN, GMM, RF	77.94	77.75	**79.12**
		SVM RBF, GMM, RF	77.48	77.88	78.68
MFCC	87.33	SVM RBF, KNN, GMM, RF	85.48	85.22	86.69
		SVM RBF, KNN, RF	85.30	85.80	86.59
LSF	84.97	SVM RBF, KNN, GMM, RF	81.72	80.49	84.64
		SVM RBF, KNN, RF	83.78	84.15	84.84
Spectral	79.38	SVM RBF, KNN, GMM, RF	73.82	72.55	77.02
		SVM RBF, GMM, RF	77.22	77.51	78.41

The results clearly show that the combination of different classifiers using the same feature set does not improve classification accuracy systematically. The accuracy rates obtained by the ensembles were higher than the best base classifier in only one (3.33%) of the analyzed cases. Even in this case, the gain was not significant.

We also evaluated the hypothesis that the combination of different representations can provide enough diversity to improve the classification accuracy. We performed experiments with different combinations of feature sets using the same induction algorithm.

First, we checked if different frequency scales used to extract cepstral coefficients can be complementary. So, we created combinations of LFC, LLFC and MFCC. We also used LSF and spectral features in combination with MFCC, since they are the best known and most used cepstral features and achieved some of the best results in our first experiment, and LFC, which obtained competitive results in comparison to MFCC. In addition, we also evaluated the combination of all feature sets (LFC, LLFC, MFCC, LSF and spectral). Table 8 shows the results.

The combination of different feature sets provided a significant number of accuracy improvements. In total, 31 (64.58%) of the analyzed cases showed some improvement. It is worth noting that the combination of all feature sets improves the accuracy over the base classifiers in all cases. The best result, 88.70%, was achieved by combining the five feature sets using the sum of SVM RBF outputs.

This result may lead the reader to questions about the real contribution of each feature in classifier combinations. So far, we only used combinations of classifiers outputs, obtained by using different features. To know the real contribution of the different types of features, we built a data set with all features with the largest number of coefficients used previously. In other words, we built a dataset with 529 features: 100 LFC, 100 LLFC, 100 MFCC, 100 LSF, 100 LPC, 12 temporal features and 17 spectral features.

Table 8. Results achieved by the combination of different feature sets with the same learning algorithm. The highlighted results represent an accuracy gain over the base classifier

Algorithm	Best Acc. (%)	Combined Feature Sets	Accuracy (%) Sum	Product	Voting
SVM RBF	87.33	LFC, LLFC, MFCC	**87.46**	87.27	**87.91**
		LFC, LSF, Spectral	86.83	86.44	87.09
		MFCC, LSF, Spectral	86.85	86.35	87.14
		All five feature sets	**88.70**	**88.47**	**88.44**
KNN	83.61	LFC, LLFC, MFCC	**85.48**	**85.57**	**84.57**
		LFC, LSF, Spectral	**83.94**	83.56	82.46
		MFCC, LSF, Spectral	**84.82**	**84.45**	83.05
		All five feature sets	**86.15**	**86.00**	**85.18**
GMM	83.17	LFC, LLFC, MFCC	**85.50**	**86.35**	**84.72**
		LFC, LSF, Spectral	83.17	**84.16**	81.49
		MFCC, LSF, Spectral	82.86	82.68	81.18
		All five feature sets	**86.20**	**86.01**	**85.50**
RF	85.39	LFC, LLFC, MFCC	**86.69**	**86.93**	84.82
		LFC, LSF, Spectral	**86.50**	**86.36**	84.76
		MFCC, LSF, Spectral	**86.99**	**86.89**	**85.44**
		All five feature sets	**87.83**	**87.97**	**86.14**

Due to the high dimensionality of this dataset, feature selection techniques were applied on it. Specifically, we used the Correlation-Based Feature Selection (CFS) [35] and the Relief [36] algorithms. In the case of Relief, the algorithm just creates a ranking of features according to their quality. We must then choose how many features will be used and select them according to the order established by the algorithm. To do this, we used 27, 53, 106 and 159 features (5%, 10%, 20% and 30% of total). The CFS algorithm does require have this parameter, and this algorithm automatically selected 74 features.

Interestingly, the MFCC are always selected in a large quantity. In all cases, CFS and variations of Relief, the feature vector with larger number of selected coefficients was always the MFCC. LFC and LSF were also taken in large numbers by the feature selection algorithms. The same happened for the spectral attributes. In contrast, the LPC and temporal features were mostly discarded.

The learning algorithms used in this phase were the KNN, SVM with RBF and Random Forest. This choice was made because these algorithms have provided the best results in previous experiments. The results are shown in Table 9.

Table 9. Result of the classification using feature selection techniques. The highlighted values are relative to those with better performance than the base classifier considering the best feature set for it

	KNN	RF	SVM RBF
All Features	83.51	**86.98**	**89.14**
CFS	**86.19**	**85.63**	**88.78**
Relief 5%	83.07	**85.63**	85.88
Relief 10%	82.76	**86.16**	86.96
Relief 20%	**83.85**	**86.86**	**87.38**
Relief 30%	**85.23**	**87.54**	**89.55**
Individual Acc.	83.61	85.39	87.33

The use of all features does not systematically improve the performance of classifiers. In one of the analyzed classifiers, this strategy achieved a lower performance than the classifier trained with only one feature vector. The same does not happen when a feature selection strategy is used. In the case of CFS, its application improved classification performance in all cases. The same happened for the algorithm Relief with certain number of selected features. In this case, 20% (106) and 30% (159) of the total.

4.4. Binary Classification

So far, we have evaluated our classifiers in a multiclass setting. Although this setting provides an overall assessment of our classifiers, not all classes are equally important in most applications. For instance, the sensor can be adapted in an intelligent trap to capture only insects of interest, such as a disease vector or an agricultural pest [5]. All other species can be regarded as a negative class, and are set free by the trap. Therefore, many practical applications require just a binary classifier. In this context, we analyze the performance of classifiers that consider disease vector mosquitoes (*Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus* and *Culex tarsalis*) as positive class and other species (*Drosophila melanogaster*, *Musca domestica*, *Psychodidae diptera*, *Cotinis mutabilis* and *Apis mellifera*) as negative class.

The classification with such division causes considerable variation in the class distribution of the data. In the complete data set (training and test) the number of samples generated by insect disease vectors is 14613, compared to only 3502 examples of the other species. In this experiment we do not evaluate only the accuracy, but also the area under the ROC curve (AUC), recommended for evaluating the classification performance in problems involving imbalanced classes [37]. In the present scenario, we chose to not use treatment algorithms for imbalanced classes, since the ratio between positive and negative classes is not considered extreme and presents no expressive loss in performance when using the SVM algorithm [4], [12].

In this task, we use 40 MFCC to train a SVM RBF, since we found that this configuration achieved the best result. We also evaluate the combination (ensemble) of all cepstrum scales, LSF, and spectral features, using the best classifier for each and combined by the sum of scores. In the first case, we achieve an accuracy of 97.82% and an AUC of 96.60%. In the second case, combining classifiers, we achieve an accuracy of 98% and AUC of 96.80%. The gain in accuracy is due to the observation that the errors in the multiclass classifiers were concentrated among species with similar characteristics, such as mosquito species. These species, in a binary classification scenario, belong to the same class.

5. Results Summary

The experiments and results presented show that we can obtain highly expressive features from the sensor data. Even though the sensor provides very brief signal events with an apparently simple structure. In particular, the MFCC form the feature set that provides the best overall results. We also show that a combination of different features can provide an improvement in the classification accuracy. The feature extraction procedure can also provide excellent results in scenarios that binary classes are considered, even in the presence of class imbalance. We believe the binary class is the most practical scenario, since most applications will require the identification of a single species of interest.

A relevant discussion is related to the computational complexity of the feature extraction procedures evaluated in this work. We note that these methods perform a fast post-computation over the spectrum or cepstrum of the signal. Therefore, they inherit the time complexity of the fast Fourier transform, $O(n \log n)$.

A more practical discussion is about the embedment of these methods in sensor, for instance, using a microcontroller. With the current technology, low-powered embedded devices can certainly handle the time complexity of the feature extraction procedures previously mentioned. However, the complexities of the feature selection procedures and ensembles of classifiers are far more challenging. However, we note that even our simplest approaches can provide results that support a practical application. For instance, the use of 40 MFCC and a SVM RBF classifier provided an accuracy of 87.33% for the multi-class classification and 97.82% (96.80%AUC) for the binary classification.

6. Conclusion

The sensor presented in this paper is important for a range of applications. For the effective operation of the sensor, it is necessary to investigate techniques for signal classification that can be used in this application. Thus, the aim of this study was to conduct and present a comprehensive investigation on these methods. We conducted our research with two approaches for time series classification: similarity search and feature extraction. Both approaches were applied using different representations.

We demonstrated the influence of different distance measures in our data. Thirteen distance measures were evaluated with classification by similarity in frequency domain and the accuracy ranged from 71.20% to 81.54%.

With the feature extraction approach, we evaluated features from temporal, spectral and cepstral representations, as well as features based on linear predictions coefficients and its variant LSF. We observed that, in different configurations of features and classifiers, the feature extraction approach is more accurate than the classification based on similarity search. More specifically, the Support Vector Machine algorithm with RBF kernel trained with MFCC achieved accuracy of 87.33%. This result represents an improvement of nearly 7% compared to the best classifier based on similarity search.

We also evaluated different ways to combine classifiers and features. The combination of different feature vectors as input to the same learning algorithm usually improves the results. In this case, the best accuracy was 88.70%. In the case of the features being used together, creating a new data set with feature subset selection techniques, the accuracy achieved 89.55%.

Acknowledgments

The authors would like to thank the funding by São Paulo Research Foundation (FAPESP). Grants #2011/04054-2 and #2012/50714-7.

Appendix A. Publications During the Development of the MSc Project

This paper summarized the publications directly related to the objectives of the candidate's MSc project [2], [5], [6], [11]. In this appendix, we summarize other the results obtained during the project development that extrapolated the limits of signal analysis for insect classification. In fact, the candidate also presents significant contributions in the areas of stream mining (with sensor classification), music retrieval, speech recognition, time series classification and learning with class imbalance.

In most applications involving intelligent sensors, it is not possible to assume that the data is generated by a stationary stochastic process. In the case of the sensor for automatic classification of insects, environmental changes, such as temperature and humidity may interfere in the metabolism of insects. In [7] the candidate worked on the initial advances of insect classification considering the data acquisition as a non-stationary data stream. In this scenario, we considered that the actual class of each insect crossing by the laser can not be given by an expert. We evaluated several strategies to adapt to drifts in the stream without actual labels, including learning with all predicted labels and with predicted labels with high confidence.

The research for feature extraction approaches for insect data leaded to other contributions. Particularly, the candidate evaluated the approaches in different applications, in an investigation to understand how the mosquitoes results could be generalized to other research areas. In [3], the candidate demonstrated that the LSF, a feature extraction approach overlooked in speech processing tasks, can create more robust speech recognition systems than the commonly used MFCC. Specifically, these features were analyzed in the task of recognizing digits spoken in Portuguese. In [8] this analysis was extended to different scenarios, including different languages, number of extracted coefficients and quality of sampling. This study showed that both feature sets have similar behavior upon changing the sample rate of the sound or the language in which speech is produced. However, the LSF were much more robust when the user makes a poor choice of the number of coefficients.

In similarity-based classification, the candidate proposed a novel distance measure that consists of two steps: (i) transforming a time series into a representation that reveals their patterns of recurrence, the unthresholded recurrence plot; (ii) on this representation, the application of the CK-1 [38], a distance measure based on video compression. Our proposal has been successfully used in time series classification [9] and in music information retrieval [10]. The proposal of using unthresholded recurrence plots as a visual representation for classification of time series, instead of extracting attributes of the binary recurrence matrix, opened a new path for other methods such as the use of image texture descriptors to classify time series [13].

Finally, during the development of the MSc project, it was possible to collaborate in the area of class imbalance [4], [12]. That was the subject of the undergraduate research project conducted by the candidate.

References

[1] D. F. Silva, "Classificação de séries temporais por similaridade e extração de atributos com aplicação na identificação automática de insetos," Master's thesis, Universidade de São Paulo, 2014. [Online]. Available: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-28042014-142456/en.php

[2] D. F. Silva, G. E. A. P. A. Batista, E. J. Keogh, and A. Mafra-Neto, "Resultados preliminares na classificação de insetos utilizando sensores ópticos," in *Encontro Nacional de Inteligência Artificial*, 2011, pp. 1–12. [Online]. Available: http://www.dimap.ufrn.br/csbc2011/anais/imagens/acrobat.png

[3] D. F. Silva, V. M. A. de Souza, G. E. A. P. A. Batista, and R. Giusti, "Spoken digit recognition in Portuguese using Line Spectral Frequencies," in *Advances in Artificial Intelligence – IBERAMIA 2012*, ser. Lecture Notes in Computer Science, 2012, vol. 7637, pp. 241–250. [Online]. Available: http://link.springer.com/chapter/10.1007%2F978-3-642-34654-5_25

[4] G. E. A. P. A. Batista, D. F. Silva, and R. C. Prati, "An experimental design to evaluate class imbalance treatment methods," in *International Conference on Machine Learning and Applications*, 2012, pp. 95–101. [Online]. Available: http://www.icmc.usp.br/pessoas/gbatista/files/icmla2012.pdf

[5] D. F. Silva, V. M. A. de Souza, G. E. A. P. A. Batista, E. J. Keogh, and D. P. W. Ellis, "Applying machine learning and audio analysis techniques to insect recognition in intelligent traps," in *Proceedings of the 12th International Conference on Machine Learning and Applications*, 2013, pp. 99–104. [Online]. Available: http://www.icmc.usp.br/pessoas/gbatista/files/icmla2013.pdf

[6] V. M. A. de Souza, D. F. Silva, P. R. Garcia, and G. E. A. P. A. Batista, "Avaliação de classificadores para o reconhecimento automático de insetos," in *Encontro Nacional de Inteligência Artificial e Computacional*, 2013, pp. 1–12. [Online]. Available: http://www.icmc.usp.br/pessoas/gbatista/files/eniac2013.pdf

[7] V. M. A. de Souza, D. F. Silva, and G. E. A. P. A. Batista, "Classification of data streams applied to insect recognition: Initial results," in *Brazilian Conference on Intelligent Systems*, 2013, pp. 76–81. [Online]. Available: http://www.icmc.usp.br/pessoas/gbatista/files/bracis2013_2.pdf

[8] D. F. Silva, V. M. A. d. Souza, and G. E. A. P. A. Batista, "A comparative study between MFCC and LSF coefficients in automatic recognition of isolated digits pronounced in Portuguese and English," *Acta Scientiarum. Technology*, vol. 35, no. 4, pp. 621–628, 2013. [Online]. Available: http://periodicos.uem.br/ojs/index.php/ActaSciTechnol/article/view/19825/pdf

[9] D. F. Silva, V. M. A. de Souza, and G. E. A. P. A. Batista, "Time series classification using compression distance of recurrence plots," in *International Conference on Data Mining*, 2013, pp. 687–696. [Online]. Available: http://www.icmc.usp.br/pessoas/gbatista/files/icdm2013.pdf

[10] D. F. Silva, H. Papadopoulos, G. E. A. P. A. Batista, and D. P. W. Ellis, "A video compression-based approach to measure music structural similarity," in *International Society for Music Information Retrieval Conference*, 2013, pp. 95–100. [Online]. Available: http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/174_Paper.pdf

[11] D. F. Silva, V. M. A. de Souza, and G. E. A. P. A. Batista, "Exploring low cost laser sensors to identify flying insect species," *Submitted to the Journal of Intelligent and Robotic Systems*, pp. 1–20, 2014.

[12] R. C. Prati, D. F. Silva, and G. E. A. P. A. Batista, "Class imbalance revisited: a new experimental setup to assess the performance of treatment methods," *Submitted to the Journal of Knowledge and Information Systems*, pp. 1–20, 2014.

[13] V. M. A. de Souza, D. F. Silva, and G. E. A. P. A. Batista, "Extracting texture features for time series classification," in *Accepted to be presented in the International Conference on Pattern Recognition*, 2014, pp. 1–6.

[14] W.H.O., "Dengue: guidelines for diagnosis, treatment, prevention and control," World Health Organization, Tech. Rep., 2009.

[15] ——, "The world malaria report," World Health Organization, Tech. Rep., 2012.

[16] D. Pimentel, "Environmental and economic costs of the application of pesticides primarily in the United States," in *Integrated Pest Management: Innovation-Development Process*, 2009, pp. 89–111.

[17] M. Q. Benedict and A. S. Robinson, "The first releases of transgenic mosquitoes: an argument for the sterile insect technique," *Trends in parasitology*, vol. 19, no. 8, pp. 349–355, 2003.

[18] P. Kevan, "Pollinators as bioindicators of the state of the environment: species, activity and diversity," *Agriculture, Ecosystems & Environment*, vol. 74, no. 1-3, pp. 373–393, 1999.

[19] K. Walker, "A review of control methods for African malaria vector," Bureau for Global Health, Tech. Rep. 108, 2002.

[20] B. Walsh, "A world without bees," Time Magazine, August 2013.

[21] G. E. A. P. A. Batista, Y. Hao, E. J. Keogh, and A. Mafra-Neto, "Towards automatic classification on flying insects using inexpensive sensors," in *International Conference on Machine Learning and Applications Workshops*, vol. 1, 2011, pp. 364–369.

[22] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[23] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[24] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Symposium on Time Series Analysis*, 1963, pp. 209–243.

[25] T. H. Park, "Towards automatic musical instrument timbre recognition," Ph.D. dissertation, Princeton University, 2004.

[26] G. E. Batista, E. J. Keogh, O. M. Tataw, and V. M. de Souza, "CID: an efficient complexity-invariant distance for time series," *Data Mining and Knowledge Discovery*, pp. 1–36, 2013.

[27] B. Zhen, X. Wu, Z. Liu, and H. Chi, "On the importance of components of the MFCC in speech and speaker recognition," in *Annual Conference of the International Speech Communication Association*, 2000, pp. 487–490.

[28] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[29] A. Watson, "Image compression using the discrete cosine transform," *Mathematica journal*, vol. 4, no. 1, p. 81, 1994.

[30] L. Rabiner and R. Schafer, *Digital processing of speech signals*. Prentice-hall Englewood Cliffs, NJ, 1978, vol. 100.

[31] A.V. Oppenheim, R.W. Schafer, *Discrete-time signal processing*. Prentice-Hall, 1989.

[32] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, p. S35, 1975.

[33] K. Paliwal and W. Kleijn, "Quantization of LPC parameters," *Speech Coding and Synthesis*, pp. 433–466, 1995.

[34] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003.

[35] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[36] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *European Conference on Machine Learning*, 1994, pp. 171–182.

[37] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Exploration Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[38] B. J. Campana and E. J. Keogh, "A compression-based distance measure for texture," *Statistical Analysis and Data Mining*, vol. 3, no. 6, pp. 381–398, 2010.