Signal classification by similarity and feature extraction with application in automatic insect identification

Diego F. Silva[1], **Gustavo E. A. P. A. Batista**[1]

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo

{diegofsilva,gbatista}@icmc.usp.br

Abstract. *Insects have a strong relationship with the human-beings. For example, some species of mosquito transmit diseases that kill millions of people around the world. At the same time, the presence of certain insects is essential for the ecological balance and food production. For this reason, we are developing a novel sensor as a tool to efficiently control disease vectors and agricultural pests without harming other species. In this paper, we demonstrate how we overtook the most important challenge to make this sensor practical: the creation of accurate classification systems. Despite the short duration and the very simple structure of the signal, we managed to successfully identify relevant features using speech and audio analysis techniques. We show that we can achieve an accuracy of 98% in the task of disease vector mosquitoes identification.*

1. Introduction

Insects have a strong relationship with the human-beings, in both positive and negative ways. For instance, mosquitoes may borne diseases that kill millions of people every year. 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 [W.H.O. 2012]. 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 [Benedict and Robinson 2003].

Due to such a complex relationship, many researchers have developed several methods of insect control [Walker 2002]. However, without the knowledge of the spatio-temporal distribution of the insects, the use of these techniques becomes costly and inefficient. Currently, studying the spatio-temporal distribution of insects is an expensive 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. 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 work, we demonstrate how we overtook the most important challenge to make this sensor practical: the creation of an accurate classification system. The data obtained by such sensor last tenths of a second and have a very simple structure. Nevertheless, we managed to successfully identify relevant features using speech and audio analysis techniques. This was the main focus of a MSc dissertation summarized in this paper [Silva 2014]. We show that we can achieve an accuracy close to 90% to identify nine

different species of insects. More important than that, we can achieve 98% of accuracy in the task of disease vector mosquitoes identification.

2. Laser Insect Sensor

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.

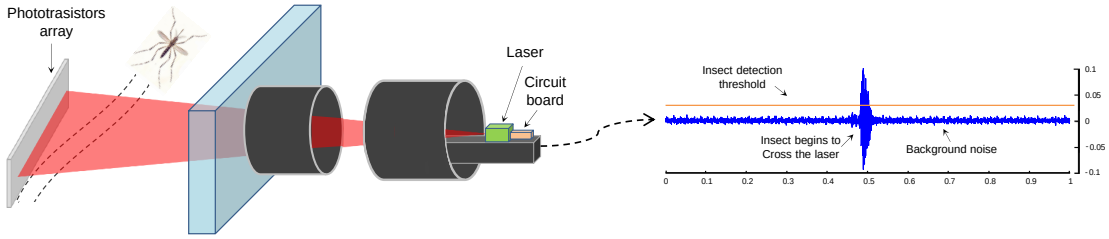

Figure 1. The logical design of the sensor used in this work

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 flies across the laser. For sake of space, we refer the reader interested in more details about the sensor's design, data collection and signal preprocessing to [Silva et al. 2014].

3. Signal Classification Approaches

Digital signals can be represented in several ways. We deeply explored the main three digital signal representations: temporal, spectral and cepstral. In addition, we used linear prediction-based features in our experiments.

In this section, we briefly describe the main strategies explored for classifying the signals obtained by the sensor.

3.1. Similarity-based Classification

The similarity-based classification depends on 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 [Silva et al. 2011]. We refer the reader to [Silva 2014, Silva et al. 2014] for a detailed description of the similarity measures used in this research.

3.2. Feature-based Classification

The second strategy for time series classification is the 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. The interested reader can find a detailed review of these features in [Silva 2014].

In this work, we use vectors with temporal and spectral features. The total number of features used in this work are 12 in the temporal and 17 in the spectral representations.

In addition to these representations, we also used features extracted from the cepstral representation. Specifically, we used Mel-Frequency Cepstrum Coefficients (MFCC), which are the most commonly used attributes in speech processing tasks, such as speaker and speech recognition. MFCC rescale the frequency spectrum before extract the cepstrum coefficients, 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 considered in this work the Linear-Frequency Cepstrum (LFC) and the Log-Linear Frequency Cepstrum (LLFC).

Finally, we also explored features based on Linear Prediction (LP). LP is based on the fact that a speech signal can be described by a simple polynomial. The coefficients of such a polynomial are calculated in order to minimize the prediction error using a covariance or auto-correlation method.

The Line Spectral Frequencies (LSF) representation, introduced by [Itakura 1975], is an alternative way to represent LP coefficients. 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 largest experiment performed during so far, we used a dataset containing four species of mosquitoes: *Aedes aegypti*, *Anopheles gambiae*, *Culex quinquefasciatus* and *Culex tarsalis*; three species of flies: *Drosophila melanogaster*, *Musca domestica* and *Psychodidae diptera*; the beetle *Cotinis 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.

In all of the performed experiments, the dataset 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.

4.2. Results Summary

We first investigated the influence of different distance measures in our data. Thirteen distance measures were used in a nearest neighbor classification. In frequency domain, the accuracy ranged from 71.20% to 81.54%. In the cepstral domain, the results varied between 26.79% and 80.34% of accuracy. This result shows the robustness of the similarity in the spectral domain.

In other experiment, we showed 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. We observed that, in different configurations of features and classifiers, the feature extraction approach is more accurate than the classification based on similarity. Specifically, the Support Vector Machine algorithm with RBF kernel trained with MFCC achieved an 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. By training different classifiers with the same feature set, we did not achieve improvements in terms of accuracy. In the other hand, the combination of different feature vectors as input to the same learning algorithm usually improved the results. In this case, the best accuracy was 88.70%. Finally, using all the feature evaluated in the same attribute-value table, we evaluated the use of feature subset selection techniques. In this case, the best accuracy rate was 89.55%.

Many applications of the sensor will require a simpler binary-class setting. For instance, in public health and agriculture, frequently the main goal is to estimate the density of a disease vector or pest of interest. In this context, we analyzed the performance of classifiers that consider disease-carrying mosquitoes as positive class and other species as negative class. This setting leads to a considerable change in the classes' distribution. For this reason, we considered the area under the ROC curve (AUC) as an additional performance measure.

In this binary classification scenario, we used 40 MFCC with a SVM-RBF classifier, since this configuration achieved the best result for a combination of classifier and feature extraction technique. By considering all the four species of mosquito as positive class, we achieved an accuracy of 97.82% and an AUC of 96.60%. If each of species of mosquito is separately considered as positive class, the results varied between 94.41% and 96.91% of accuracy and 86.10% and 94.20% of AUC. In these scenarios, the combination of classifiers did not presented significant improvements in the results.

5. Conclusion

The sensor presented in this paper is important for a range of applications. For its effective operation, we investigated techniques for signal classification that can be used in this domain.

A relevant discussion is about the embedment of these methods in the 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. Nevertheless, 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 considering disease-carrying mosquitoes as positive class.

There are several applications that require real-time estimation of spatio-temporal distributions of important insects. Therefore, we argue that the presented research has

great impact in different areas. For instance, the sensor for automatic insect classification may contribute to public health and agronomy.

A. Publications During the Development of the MSc Project

This paper summarized a MSc dissertation [Silva 2014], as well as the publication directly related to it [Silva et al. 2011, Silva et al. 2013c, Souza et al. 2013b, Silva et al. 2014]. In this appendix, we summarize other results obtained in this research, which extrapolated the limits of signal analysis for insect classification.

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 may interfere in the metabolism of insects. In [Souza et al. 2013a] we worked on the initial advances of insect classification considering the data acquisition as a non-stationary data stream. In this scenario, we evaluated several strategies to adapt to drifts in the stream without actual labels.

The research for feature extraction approaches leaded to other contributions. Particularly, we carried out an investigation to understand how the mosquitoes results could be generalized to other domains. In [Silva et al. 2012], we demonstrated that the LSF, overlooked in speech processing tasks, can create more robust speech recognition systems than the commonly used MFCC. In [Silva et al. 2013d], this analysis was extended to different scenarios, including different languages, number of coefficients and sampling quality. This study showed that both feature sets have similar behavior upon changing the sample rate of the signal 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, we proposed a novel distance measure for time series comparison that consists of two steps: (i) transforming a time series into a visual representation, the unthresholded recurrence plot; (ii) on this representation, the application of the CK-1 [Campana and Keogh 2010], a video compression-based distance. Our proposal has been successfully used in time series classification [Silva et al. 2013b] and in music information retrieval [Silva et al. 2013a]. The proposal of using unthresholded recurrence plots for classification of time series, instead of extracting features of the binary recurrence matrix, opened a new path for other methods such as the use of image texture descriptors [Souza et al. 2014].

Finally, during the development of the MSc project, it was possible to collaborate in the area of class imbalance [Batista et al. 2012, Prati et al. 2014].

References

Batista, G. E. A. P. A., Silva, D. F., and Prati, R. C. (2012). An experimental design to evaluate class imbalance treatment methods. In *International Conference on Machine Learning and Applications*, pages 95–101.

Benedict, M. Q. and Robinson, A. S. (2003). The first releases of transgenic mosquitoes: an argument for the sterile insect technique. *Trends in parasitology*, 19(8):349–355.

Campana, B. J. and Keogh, E. J. (2010). A compression-based distance measure for texture. *Statistical Analysis and Data Mining*, 3(6):381–398.

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57:S35.

Prati, R. C., Silva, D. F., and Batista, G. E. A. P. A. (2014). Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Journal of Knowledge and Information Systems*, pages 1–24.

Silva, D. F. (2014). Classificação de séries temporais por similaridade e extração de atributos com aplicação na identificação automática de insetos. Master's thesis, Universidade de São Paulo.

Silva, D. F., Batista, G. E. A. P. A., Keogh, E. J., and Mafra-Neto, A. (2011). Resultados preliminares na classificação de insetos utilizando sensores ópticos. In *Encontro Nacional de Inteligência Artificial*, pages 1–12.

Silva, D. F., Papadopoulos, H., Batista, G. E. A. P. A., and Ellis, D. P. W. (2013a). A video compression-based approach to measure music structural similarity. In *International Society for Music Information Retrieval Conference*, pages 95–100.

Silva, D. F., Souza, V. M. A., and Batista, G. E. A. P. A. (2013b). Time series classification using compression distance of recurrence plots. In *International Conference on Data Mining*, pages 687–696.

Silva, D. F., Souza, V. M. A., and Batista, G. E. A. P. A. (2014). Exploring low cost laser sensors to identify flying insect species. *Journal of Intelligent and Robotic Systems*, pages 1–18.

Silva, D. F., Souza, V. M. A., Batista, G. E. A. P. A., and Giusti, R. (2012). Spoken digit recognition in Portuguese using Line Spectral Frequencies. In *Ibero-American Conference on Artificial Intelligence*, pages 241–250.

Silva, D. F., Souza, V. M. A., Batista, G. E. A. P. A., Keogh, E. J., and Ellis, D. P. W. (2013c). Applying machine learning and audio analysis techniques to insect recognition in intelligent traps. In *International Conference on Machine Learning and Applications*, pages 99–104.

Silva, D. F., Souza, V. M. A. d., and Batista, G. E. A. P. A. (2013d). A comparative study between MFCC and LSF coefficients in automatic recognition of isolated digits pronounced in Portuguese and English. *Acta Scientiarum. Technology*, 35(4):621–628.

Souza, V. M. A., Silva, D. F., and Batista, G. E. A. P. A. (2013a). Classification of data streams applied to insect recognition: Initial results. In *Brazilian Conference on Intelligent Systems*, pages 76–81.

Souza, V. M. A., Silva, D. F., and Batista, G. E. A. P. A. (2014). Extracting texture features for time series classification. In *International Conference on Pattern Recognition*, pages 1425–1430.

Souza, V. M. A., Silva, D. F., Garcia, P. R., and Batista, G. E. A. P. A. (2013b). Avaliação de classificadores para o reconhecimento automático de insetos. In *Encontro Nacional de Inteligência Artificial e Computacional*, pages 1–12.

Walker, K. (2002). A review of control methods for African malaria vector. Technical Report 108, Bureau for Global Health.

W.H.O. (2012). The world malaria report. Technical report, World Health Organization.