TECNIACUSTICA ELCHE · 2022

53º CONGRESO ESPAÑOL DE ACÚSTICA
XII CONGRESO IBÉRICO DE ACÚSTICA

ANNOTATION AND RECOGNITION OF SOUND EVENTS IN THE PADEL/TENNIS SPORT ACTIVITY

Fernandes, C. [1], **Cabral, J.** [1], **Trigo, P.**[1], **Preto Paulo, J.**[1,2]

[1] ISEL-Instituto Superior de Engenharia de Lisboa, Portugal
[2] LAA-Lab. de Áudio e Acústica do ISEL, Portugal
{a45118@alunos.isel.pt, a46357@alunos.isel.pt, paulo.trigo@isel.pt, joel.paulo@isel.pt}

PACS: no. 43.60.-c, 07.05.Mh

Keywords: Sound Detection; Artificial Intelligence Algorithms, Machine Learning, audio datasets, Padel sport.

ABSTRACT

In a given sport, the performance of the athletes improves when they are supervised from an external perspective. In this scenario, the athlete can be supervised by a coach in order to achieve better results, or as a complement, it is possible to video record and analyze his performance afterwards. Thus, it is advantageous to develop a tool capable of performing this external analysis. The tool developed in this project, allows extracting and recognizing relevant events (i.e., periods when a greater exchange of balls occurs) based on the video of the athlete's sport activity, such as in the padel/tennis. The processing is done based on the audio extracted from the video, and is based on the extraction of patterns identifying the events, with the help of machine learning techniques, based on time series extracted from the sound waveform. At the end, statistics are performed, allowing a detailed summary of what was recorded in the video, giving a more comprehensive and objective perspective of the athlete's performance. The tool correctly identifies about 85% of the events under analysis, with some adjustments needed to improve the recognition process. The tests are carried out at ISEL's Audio and Acoustics Laboratory, LAA.

RESUMO

O desempenho de um atleta, num determinado desporto, melhora quando é acompanhado de uma perspetiva externa no decurso da sua atividade desportiva. Neste sentido, é vantajoso o desenvolvimento de uma ferramenta capaz de realizar essa análise externa. Trata-se de um problema de classificação binária em que o sistema deve fazer a distinção entre batidas de bola e ruído. Estes eventos têm uma duração típica pré-determinada e a distinção entre eles foi feita com base nas características *Onset Detect*, *Root Mean Square* e *Spectral Flux*) identificadas no áudio através de técnicas de aprendizagem automática. O *dataset* utilizado no processo de aprendizagem foi construído de raiz e o desequilíbrio verificado neste conjunto de dados foi abordado utilizando a técnica sub-amostragem. Entre os vários métodos de aprendizagem automática analisados, escolheu-se a rede neuronal do tipo MLP (*MultiLayer Percepton*) para realizar a discriminação entre os tipos de eventos mencionados. No final, através de uma aplicação *web*, são visualizados os resultados respeitantes aos instantes em que ocorrem batidas de bola, dando uma perspetiva mais abrangente e objetiva do desempenho do atleta. A ferramenta identifica corretamente cerca de 80% dos eventos em análise. Os ensaios são realizados no Laboratório de Áudio e Acústica do ISEL, LAA.

1. INTRODUCTION

Among the various sports practiced, this work focuses on the analysis of the sport component of padel, in order to provide athletes and/or coaches with a complement in the course of training.

The annotation of sound events in video refers to the analysis of audio extracted from a previously recorded video, with the objective of identifying each instant in which a tennis ball strike occurs. The identification of this type of event is accomplished through machine learning algorithms. This work analyzes only videos in indoor environment and all the processing on the audio will be performed in offline mode. The main contributions of this work are:

- Building a dataset where the events (ball hits and noise) are present;
- Building a model that allows the identification of the events referred to in the dataset;
- Building a web application that allows you to visualize the results returned by the classifier.

2. RELATED WORK

The system developed in [9], focuses on the acoustic component to distinguish between 6 types of events in tennis matches using a convolutional neural network (CNN). The system also provides a tool to perform event annotation, increase the dataset and improve the quality of the model. The audio signal is split into 20 millisecond frames from which MFCC coefficients are obtained to recognize events in a human-like manner. The work presented in [10] performs event detection by combining visual and acoustic information obtained from basketball videos. The acoustic component focuses on events that could be indicators of the most important moments in the game, such as the sound emitted by spectators or the enthusiasm in a commentator's voice. This author also considers the energy level associated with the segments in the process of building the feature vectors.

3. FUNDAMENTAL CONCEPTS

This chapter discusses the theoretical concepts that support the work implemented in this project.

3.1. Signal Processing Concepts

In this work, the following features are extracted from audio: onset detection and root mean square (RMS).

The Onset is divided into two components: the Onset Detect and the Spectral Flux. The first component corresponds to detecting the instant at which a given sound starts [1, 2]. The second component is to verify variations in the signal spectrum, in order to find differences between consecutive frames, which also allows detecting the beginning of a sound. Spectral flux can also be defined as a measure of how fast the spectrum of a signal varies [3].

The root mean square (RMS) refers to the average energy (intensity) concentrated in a frame [4].

Figure 1 is a representation of the three characteristics listed.

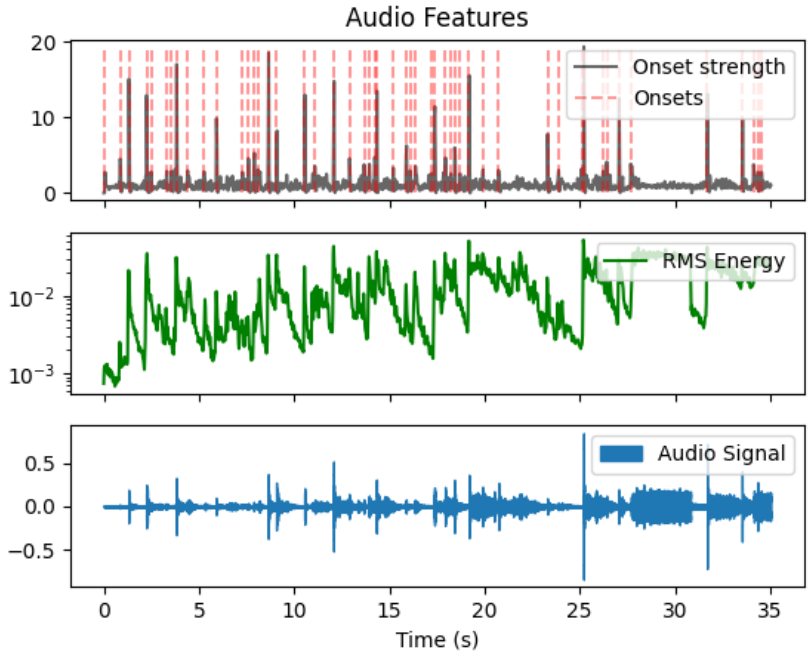

Figure 1 – Features (top and middle) and corresponding audio (bottom).

Observing the Figure 1, we see that the peaks of the two Onset components, as well as the RMS correspond to the moments in the original audio where the amplitude is highest.

3.2. Machine Learning Concepts

Machine Learning (ML) is a branch of artificial intelligence that uses the analysis of (large volumes of) data to extract patterns and represent them in a model [5].

Dataset

A dataset is a set of data used in the process of building the models generated through machine learning methods.
The dataset is structured in columns and rows (table format). In the case of supervised learning the columns represent the features and the class. The rows always represent instances (examples or observations). Considering a matrix of features, X, and the corresponding vector of class values, y, the classification problem in question aims to obtain the function, f, such that:

$$f(X) = y \quad (1)$$

The function $f(X)$ corresponds to a model that must receive the feature matrix and correctly classify each of its examples.

Imbalance in the dataset occurs when the distribution of examples across classes is uneven. This imbalance can result in the construction of a biased model that better recognizes certain classes. This problem can be addressed by increasing the number of examples from the minority class (oversampling) or reducing the number from the majority class (undersampling) [6].

In padel games, the number of ball hits is much lower than the background noise, so in this work we will use the subsampling technique to balance the dataset.

Neural Networks

An artificial neural network (ANN) is a framework used in machine learning methods that takes inspiration from human biology and the way neurons communicate with each other to understand perceived data [7].

A neural network is organized into layers, which are made up of units (network nodes).
Figure 2 illustrates a neural network, where the connections between the various constituent nodes are delineated. A net such as the one in this figure is called a Multi Layer Perceptron (MLP). Between the input layer and the output layer there are one or more hidden layers. Each node receives the output values from the nodes of the previous layer and associates a weight to them that is updated throughout the learning process [8].

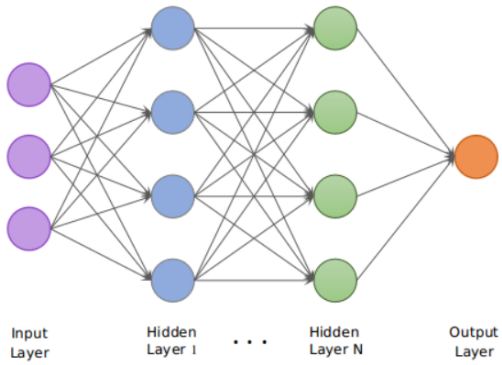

Figure 2 - Typical representation of a neural network.

The output values are calculated based on an activation function that defines how the weights to be updated are calculated in a given layer of the network.

In this work, it is intended that the neural network performs the distinction between ball hits (class of positives) and noise (class of negatives).

4. APPROACH

The developed system recognizes ball hits, allows to assist in the dataset building process and consequently in the process of improving the classifier performance. The operation of the system can be represented as shown in Figure 3.

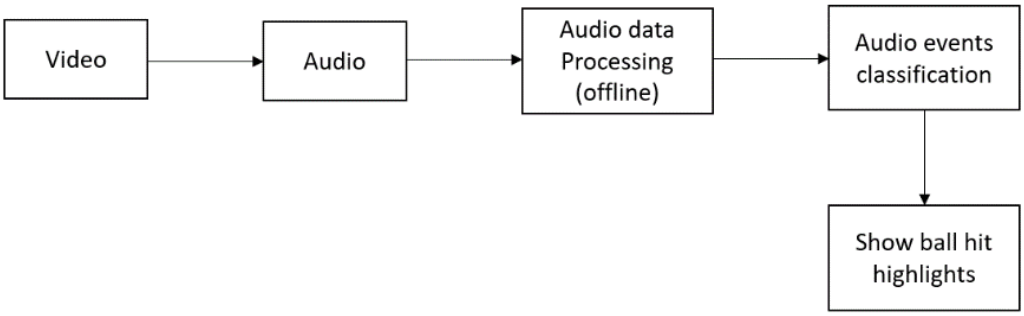

Figure 3 – Proposed event annotation system.

4.1. Build of *Dataset*

The process of building the dataset is illustrated in Figure 4.

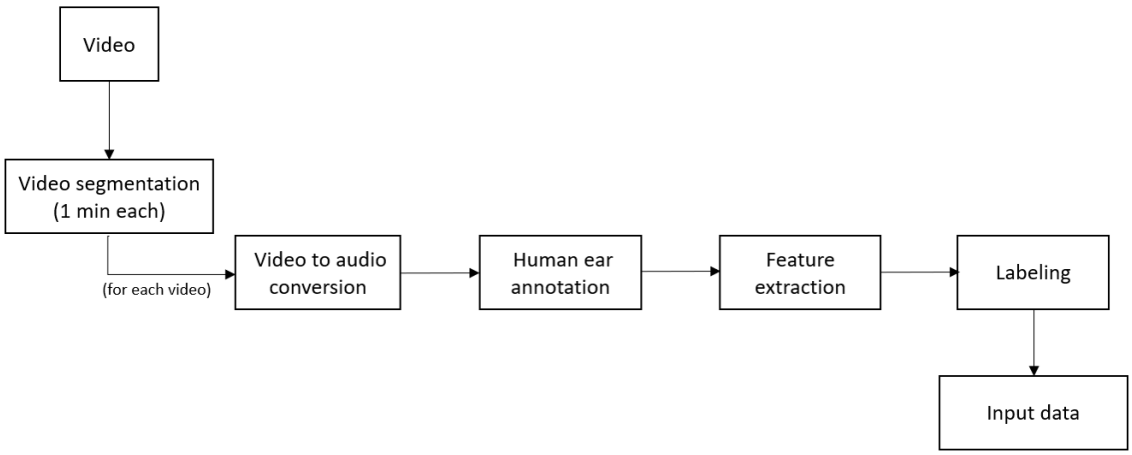

Figura 2 – *Dataset* construction process.

In an initial phase, the training video is segmented into one-minute-long videos. From each of these videos the corresponding audio is extracted. In the "Human ear annotation" phase, the audios are listened to and the instants in which the ball hits occur (initial sample and final sample of the event) are annotated (.csv files). All unannotated information is considered noise.

The impact of a ball strike has an average duration of 0.5 seconds (for the set of sounds analysed). Therefore, in the feature extraction process, a sweep over the audio is performed with a time window of fixed size equal to half a second. This process is performed over several iterations, where in each iteration the window swipes a specific number of samples and the samples covered by the window are used to calculate the desired features. Also at each iteration the window is divided into *N* sample segments according to the following expression:

$$N = \frac{\text{eventLength} \times \text{samplingRate}}{\text{hopLength}} \quad (2)$$

where in the equation, $eventLength$ is the length of an event, $samplingRate$ is the sampling frequency (at which the audios are obtained), and $hopLength$ refers to the number of samples scrolled at each iteration of the sweep, as well as the number of samples in each segment of the window. Table 1 shows the values of *N* obtained for the various combinations of $eventLength$ and $hopLength$.

Table 1 - Values of duration (in seconds), samples and N to consider, for a $samplingRate$ value equal to 44100Hz.

$eventLength$	$hopLength$	N
0.5	1024	22
0.5	2048	11
0.5	4096	5

Figure 5 is a representation of the process of scanning over the audio to build the feature matrix, X. In this figure the $hopLength$ value is equal to 2048, so at each iteration the window skips 2048 samples and is divided into 11 groups of 2048 samples. However, throughout the model building process multiples of this value can be used to see what impact they have on the performance of the model.

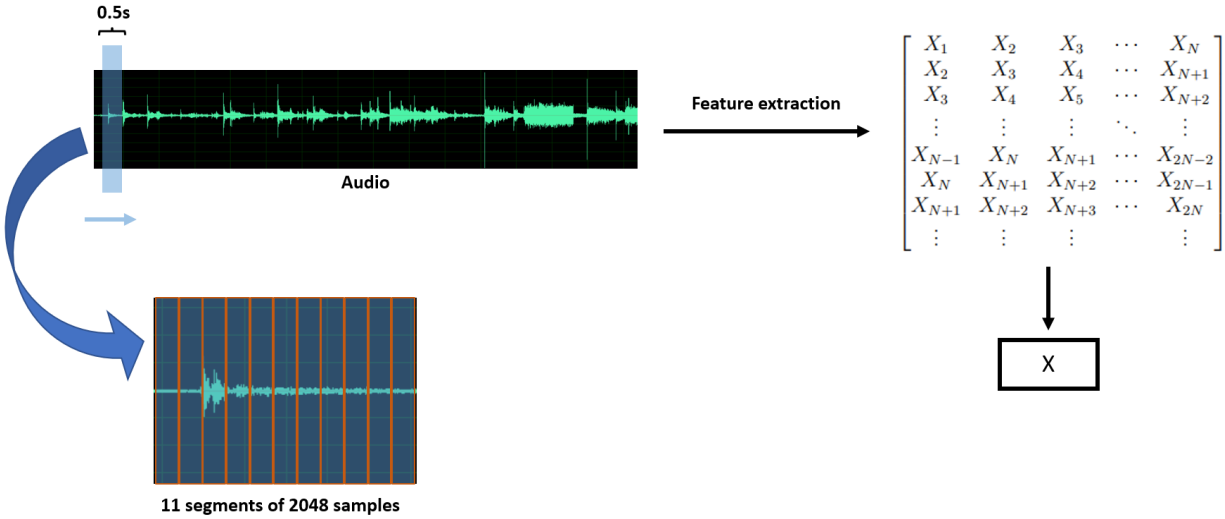

Figure 5 - Feature extraction process.

After the feature extraction process, the labeling phase is performed, where the annotated information (classes) is joined with the information referring to the feature extraction (features). This step corresponds to assigning each of the examples in the feature matrix a class (ball hit or noise).

4.2. Construction of the Classifier

The construction of the model can be described using Figure 6, which broadly represents the phases involved in this process.

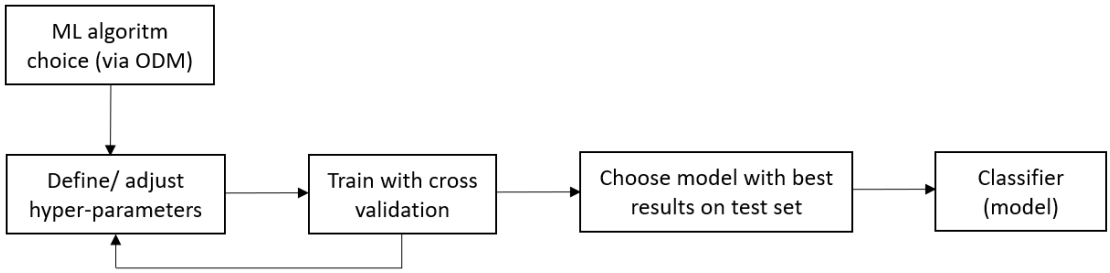

Figure 6 - Process of building the classifier.

In the phase of choosing the machine learning method, the multi-layer neural network (MLP) was chosen.

To obtain a model that correctly identifies racket ball hits, several experiments were performed, changing the parameters of the neural network. Table 2 contains the results obtained for these experiments. A number of batches equal to 128, the binary cross-entropy error function and the Adam optimization algorithm were considered. All other hyper-parameters are shown in the table.

The duration of the events (eventLength) was preset to half a second (0.5 s), so the value of *N* (in expression 2) and the number of examples in the dataset in ("Dataset Observations") depend on the value of hopLength ("Hop"). Each of the experiments (1 to 18) was performed by applying the Stratified KFold technique, which allowed to obtain 5 different models.

Throughout the various experiments it was found that simpler models obtain better performances for data never before analyzed.

Of the experiments under analysis, experiment 12 was chosen, since the corresponding neural network is the simplest compared to the others and the hopLength value (4096) generates fewer clips in the web application.

Table 2 - Results obtained for the various hyperparameters of the neural network and different datasets.

Experiment	Hop	Layers	Neurons		Regularization			Epochs	Dataset Observations			
									Train		Validation	
			1st Layer	Hidden Layers	Dropout	L1	L2		Ball Hits	Noise	Ball Hits	Noise
1	1024	4	66	64 32	0.4 0.2	10-3	10-3	100	15780	43716	3944	10930
2	1024	4	66	64 32	0.4 0.2	10-3	10-3	100	12653	12739	3163	3185
3	2048	3	33	256	0.45	N/A	10-5	100	6328	6367	1582	1592
4	2048	3	33	256	0.45	N/A	10-5	300	7897	21863	1974	5466
5	2048	3	33	256	0.45	N/A	10-5	300	6328	6367	1582	1592
6	4096	4	15	32 16	0.4 0.2	N/A	10-5	300	3068	3213	767	804
7	4096	3	15	64	0.4	N/A	10-4	200	3068	3213	767	804
8	4096	3	15	64	0.4	N/A	10-4	200	11076	3828	958	2768
9	1024	3	66	64	0.4	N/A	10-6	200	12653	12739	3163	3185
10	1024	3	66	64	0.4	N/A	N/A	200				
11	2048	3	33	64	0.4	N/A	N/A	200	6328	6367	1582	1592
12	4096	3	15	64	0.4	N/A	N/A	200	3068	3213	767	804
13	1024	4	66	64 32	N/A	10-2	N/A	200	12653	12739	3163	3185
14	1024	4	66	64 32	N/A	10-4	N/A	200				
15	1024	3	66	128	N/A	10-3	10-3	200				
16	1024	4	66	32 32	N/A	N/A	10-5	200				
17	1024	3	66	128 32	N/A	N/A	10-5	200				
18	1024	3	66	64	N/A	N/A	10-5	200				

Experiment	Learning Rate	Accuracy		Loss		Activation Function
		Train	Validation	Train	Validation	
1	10-3	92%	92%	0.25	0.24	sigmoid
2		92%	93%	0.25	0.23	
3		93%	92%	0.18	0.21	
4		95%	94%	0.16	0.18	
5		94%	92%	0.16	0.22	
6		91%	90%	0.25	0.26	
7	10-2	90%	90%	0.27	0.28	
8		92%	93%	0.23	0.22	
9	10-4	93%	93%	0.19	0.19	
10	10-3	94%	93%	0.16	0.18	
11		93%	92%	0.19	0.21	
12		91%	90%	0.25	0.26	
13		89%	89%	0.32	0.32	
14		94%	93%	0.18	0.22	
15		95%	93%	0.15	0.22	
16		95%	93%	0.13	0.21	
17		97%	93%	0.09	0.25	
18		96%	93%	0.11	0.23	

4.3. Web Application Development

As mentioned in the introduction, the application aims to allow the user to visualize the results returned by the chosen model and contribute to the increase of the dataset with validated information.

In the application, the user is redirected to page 2 when selects a particular video (on the homepage). On this page the user can view each of the clips classified by the model, or make changes if he disagrees with the way the model classified them. Upon submitting the changes made, the user is redirected to page 3, where the changes made are visualized. These changes are saved in text files (.txt) and can later be added to the dataset by the user.

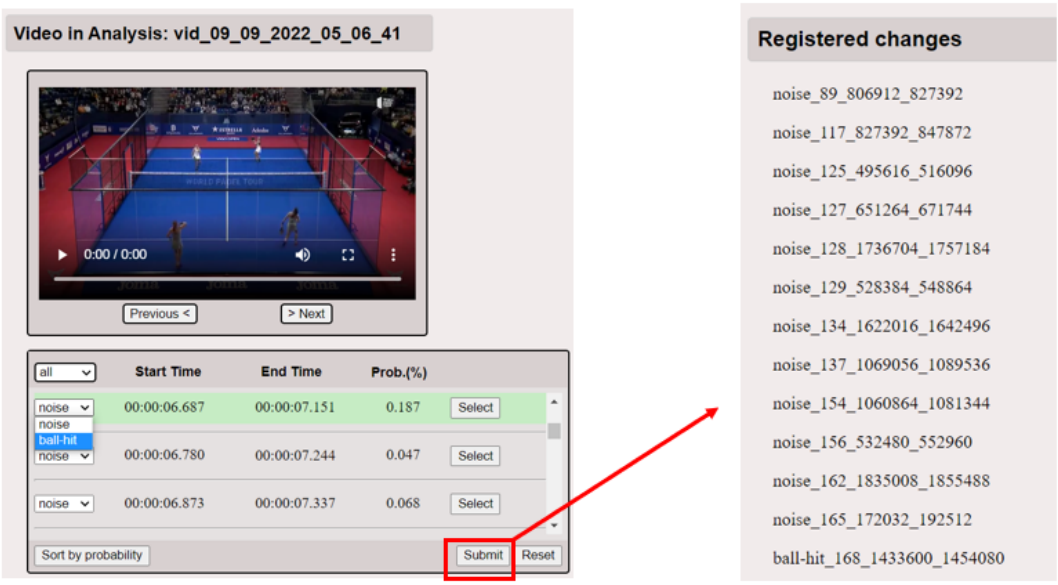

Figure 7 - Layouts of pages 2 (left) and 3 (right) of the web application.

5. TESTING ON NEW VIDEOS

To check how well the classifier performs when analyzing data with characteristics different from those observed in the dataset, 4 more videos (A, B, C and D) with durations between 25 and 45 seconds were tagged. The results obtained for these videos are shown in Figures 8 and Table 3.

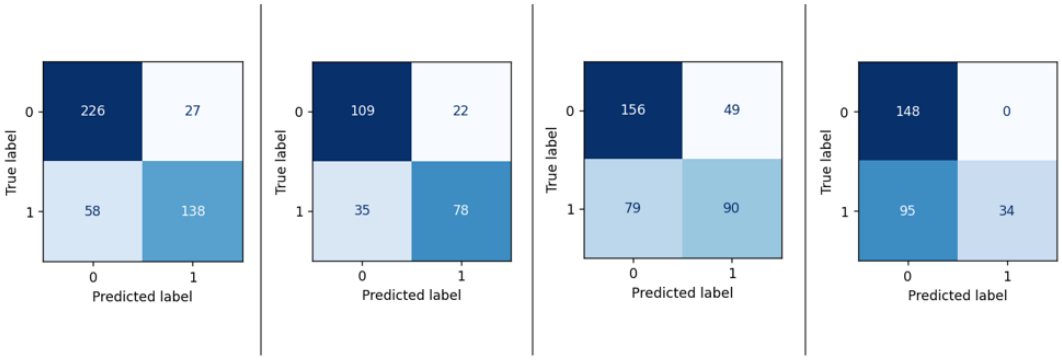

Figure 8 - Confusion matrices resulting from the classification of video A, B, C and D (from left to right).

Table 3 - Results of the classification on videos with different characteristics.

Test Accuracy				Test Loss			
Video A	Video B	Video C	Video D	Video A	Video B	Video C	Video D
81%	77%	66%	66%	0.42	0.53	0.76	1.17

The results suggest that more examples with similar characteristics to the videos should be added in order to obtain better results in the classifier. In this sense, the tool (application) developed can be used to annotate more videos, add more information to the dataset, retrain the model and consequently improve its quality.

6. CONCLUSIONS

The manual annotation of events is a very time consuming process. In this project we developed an annotation tool that allows the identification of (racket) ball strikes in padel games in order to overcome this problem.

In the process of building the dataset, the task of creating the sliding window allowed us to define various values for the bounces that occur at each iteration in the scanning process. It was also found that different jumps produce different datasets, which allows for variability in the data used to train the neural network.

The tests performed by varying the hyper-parameters in the neural network allowed us to verify that unbalanced datasets lead to the construction of models that better recognize the examples of the noise class than the examples of the hit class. It was also found that the use of the Dropout regularization technique is very effective in reducing the complexity of the model, allowing it to generalize more easily.

The model was chosen for its simplicity and for reducing the time to be spent by the user during the automatic annotation of events. When classifying videos with characteristics different from those observed in the dataset, it was found that performance decreases. However, it is possible to perform annotations on these videos in order to improve the dataset and improve the quality of the model.

This project allows us to conclude that it is possible to use machine learning techniques to detect events in padel games.

FUTURE WORK

The developed application already identifies some ball hits, but at this point the dataset is quite incomplete in terms of variability. In this sense, it would be very advantageous to use the tool to automatically increase the dataset.

Depending on the environment where the padel sport is played (indoor or outdoor), the sound propagation are different. In this sense, as other approaches, it could be verified what impact changing the event length has on the model built, and a process of dynamically adjusting this parameter could be advantageous.

Adding other types of events, i.e. recognizing other types of events or other types of ball hits could also contribute to making the application richer.

The features used to extract information from the sound could be reviewed, in order to verify if there are features that make it easier to distinguish the events under analysis.

Regarding the construction of the model, other algorithms could be used, such as the logistic regressor, which according to the ODM application, also classifies the ball hits with some robustness and CNNs.

It is expected, in a future perspective, that this application will be useful in the recognition of ball hits.

ACKNOWLEDGMENTS

This work has been supported by ISEL's Audio and Acoustics Laboratory, LAA..
https://acusticaudiolab.isel.pt

REFERENCES

[1] Semmlow, J, Chapter 3 - fourier transform: Introduction. In Semmlow, J., editor, Signals and Systems for Bioengineers (Second Edition), Biomedical Engineering, p. 81–129. Academic Press, second edition, 2012
[2] Rosão, C. M. T., Onset detection in music signals. PhD thesis, 2012
[3] Meyda, Audio feature extraction for javascript. https://meyda.js.org/audio-features.html, (visitado em Jul.2022)
[4] Room, C., Audio feature extraction. machine learning, 16(17):51, 2021
[5] Janiesch, C., Zschech, P., e Heinrich, K. ,Machine learning and deep learning. Electronic Markets, 31(3):685–695, 2021
[6] Badr, W., Having an imbalanced dataset? Here is how you can fix it. Towards Data Science, 22, 2019
[7] Rauber, T. W. Redes neuronais artificiais. Documento de Apoio (visitado em Jul, 2022)
[8] Medium, A beginner intro to neural networks. https://purnasaigudikandula.medium.com/ a-beginner-intro-to-neural-networks-543267bda3c8, 2019
[9] Baughman, A., Morales, E., Reiss, G., Greco, N., Hammer, S., e Wang, S. Detection of tennis events from acoustic data. In Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports, p. 91–99, 2019
[10] Liu, S., Xu, M., Yi, H., Chia, L.-T., e Rajan, D. (2006). Multimodal semantic analysis and annotation for basketball video. EURASIP Journal on Advances in Signal Processing, 2006:1–13