uc3m | Universidad Carlos III de Madrid

University Degree in Audiovisual Systems

Academic Year (e.g. 2014-2019)

Bachelor Thesis

“Violent event detection from acoustic signals”

Óscar Otero Martínez

Carmen Peláez Moreno

Madrid, January 15, 2020

SUMMARY

Keywords:

DEDICATION

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Context

Violence against women remains an invisible phenomenon, deeply within the victim's private life in most cases. It is based on deep social and cultural roots and it is undoubtedly linked to unbalanced relationships between men and women in different situations and contexts, such as economics, politics and religion. In order to prevent these conflicts, the related legislation has achieved important improvements for the last years. According to the results of most of the studies, victims can be usually defined as women who endured violence during their childhood and felt socially isolated. They are also characterized for a considerable economic dependency and a low educational level.

With the purpose of making a difference when identifying situations showing this kind of violence and apply all the knowledge and technological advances acquired during this information era, machine learning and deep learning models can collect all the available data to protect eventual victims.

The main goal is to get to know how the victim is feeling, for example, if she is scared or nervous, and combine this with other variables which may play an important role in the scene and might be helpful in making a decision about the characterization of the ambiance. There are several factors that can be considered to achieve this task. One of them is the audio, either the victim's voice or the environmental sounds.

Plenty of useful information can be extracted from the acoustic scene of a certain place. The detection of audio events is an equally good way to define what is happening in a certain moment. Once these data are collected, they can be classified in different categories and thus describe the scene. Based either on an objective definition of gender violence or in an explanation previously obtained from a particular/specific victim, this acoustic knowledge can be interpreted as dangerous for the user.

1.2. Objectives

The utilization of learning models to extract useful information from the worlds data has become a very common practice in most of the fields. One type of habits that have gained a lot of popularity in the scientific community is the use of multimedia data. In many cases, the samples used to train the models consist in images that belong to a certain kind of problem, such as medical imaging or object recognition. This field is known as computer vision (CV). Many world well known architectures and enormous data bases have been born during the study of this kind of problems.

In the same way, audio data have been used to get conclusions from a lot of real world problems. In order to tackle the task of violent event detection it is important to decide what perspective is going to be taken into account when defining a violent event, whether an objective point of view or a more personalized standpoint according to the victim criteria. Apart from this, it is also necessary to extract the required features, that is, information from the audio signals that will allow to train the models so to get the results. However, the main work will be characterized by classifying a whole scene depending on the events this is built by. Once an action sound is categorized, it can be identified as violent by checking if it belongs to the violence definition previously defined.

The different acoustic scenes that may be considered for the problem can be composed by events of different nature or those that belong to just one class. This difference may cause that the techniques utilized to address the problem can differ. As a further approach, it is interesting to find a method that can distinguish among events that come from different sources of audio.

1.3. Regulatory framework

1.4. Socio-economic environment

2. STATE-OF-THE-ART

2.1. ASC and AED/C

Acoustic scene classification, also known as ASC, refers to the association of an audio sequence to a certain semantic label that describes the environment in which it took place [1]. With this idea in mind, the classification of acoustic sceneries have been attacked with two different kinds of concepts: soundscape cognition, this is, understanding how the human beings perceive the sounds in a subjective way from the physical environment that surrounds them [2] and working on new computational methods that may help and allow to perform this task in an automatic way by using machine learning and processing signal techniques, which is also called, computational auditory scene analysis (CASA) [3]. In many applications this notion can be found, as in context recognition, based on allowing devices to achieve benefits and information from the situation it is placed in [4], also for medical utilizations [5], as a tool for musical recognition [6] or for a complement to computer vision.

At the same time that advances have been taken place in the ASC field, another related area has evolved during last years. Some computational work has been deployed for the tasks of acoustic event detection and classification, also known as AED/C. It can be described as the processing or treatment of sound signals in order to convert them into significant descriptions that match a listener's sensing of the events and sources that compose the acoustic environment [7]. The detection part consists on identifying the events in a temporal stream of audio and assign them a label. The result is usually accompanied by the time interval in which the occurrence can be found. However, the classification is a task that acts directly on the event that has been already isolated and has the purpose of designating a label or class to the sound [8]. There exist plenty of applications in which these techniques have been used for, as in the medical field [9], in biological topics such as bird noise detection [10], and for multimedia information retrieval from video sources in social media [11].

2.1.1. Features and methods

In the literature, a bunch of works have been published related to ASC field. These can be sorted into two different currents in regard to how the problem is addressed. One of them considers the scene as a single instance with the purpose of representing it through a long-term statistical distribution that models a set of low-level features[12]. There exist different ways of characterizing an acoustic event or scene for this type of method. In previous works, some of the common habits usually utilized for speech recognition had the main role in the extraction of features, such as the fundamental frequency, or F0, F0 envelope and the probability of voicing. Apart from these, also spectral features, as Mel-Spectrum bins, zero crossing rate (ZCR) and spectral flux (SF), and energy features, such as the energy in bands or the logarithmic-energy [13] had an important job on this task. However, the best results have been achieved with what is called Mel-frequency cepstrum coefficients (MFCC) which is defined as a cepstral feature, which will be explained further on. This kind of characteristics extracted from the audio can be called low-level descriptors and they are usually combined with algorithms and methods to address the classification task. In this "bag-of-frames" approach, in which the scene is considered as a single object, a typical technique was to model the samples features into global statistical characteristics from the local descriptors by using Gaussian Mixture Models (GMM) [14].

Explain more MFCC?

There is another path to dig for acoustic scene classification, which consists on including a representation of data previous to the classification which is based on transforming the scene by using a set of high level features normally obtained with a vocabulary or dictionary formed by acoustic atoms. These are usually a depiction of events or streams within the scene and do not need to be known a priori [12]. Apart from the typical well-known audio features, the ones named above as low-level descriptors, there exists other acoustic characteristics which may seem to be hidden in the data but can be found by using unsupervised-learning methods. This is the way to act when dealing with the acoustic atoms mentioned above. One of the approaches that can be found in the literature about this idea is based on the use of a previously learned overcomplete dictionary that is utilized to sparsely decomposed the spectrogram of audio. This dictionary will be used by an encoder which has the labour of mapping new input data to real similar version of their own sparse representation in a fast and efficient way. Finally, the obtained codes will feed a SVM classifier used for the task of music genre prediction [15].

Include results?

Another job done in the sparse-feature representation framework presents a way of mixing high feature learning techniques with a pooling method for the objective of music information retrieval and annotation. After some preprocessing of the audio signals data, three feature-learning algorithms are trained finding that sparse restricted Boltzmann machine (sparse-RBM) gets better results than K-means and Sparse Coding. Once the features are obtained, an extra step takes place before performing the classification task, the one called pooling and aggregation. The goal of this procedure is to achieve a feature representation for a long sequence as a song is. Since when joining short-term features that belong to small segments inside the song may result in a loss of their local meaning, a max-pooling operation is computed over each subsegment in order just to consider the maximum value for each feature dimension. After that, these are aggregated by computing the average. The max-pooling contribution resides on reducing the smoothing effect when averaging the values [16]. This approach is feasible because of the homogeneity in music data. However, this technique could be a bit risky when dealing with acoustic scenes. For this case, a modified version of this method has been proposed. Taking into account that the presence of events is less frequent, instead of considering the whole long sequence to apply the max-pooling for, it will just be used in those segments that had been already detected as significant events by establishing a threshold value and setting an onset and offset that allow to know the start and end time[17].

Include picture of the pipeline?

The classification of acoustic scenes can go with the hand of event detection. The working method used for this task is really similar to the one used for ASC. Then, it is not surprising that most of the works found in the literature address this task with the use of MFCC as features and with such as HMM or GMM. For the purpose of finding the desired events, the whole detection process can be split in two parts. Firstly, a classification of already isolated events should be executed in order to build a vocabulary of acoustic actions. In this case, the data used belong to short-term sequences that must strongly show the semantic meaning of the corresponding event. This is important because there may be more acoustic representations in the same short segment than the one that is desired to detect, but this must stand out among the others. Then, for the detection part, the input data will be composed by long tracks so time allocation of the events will be implemented. So, after obtaining the different short segments from breaking the long sequence up, they will be classify taking into account the results from the first step [18].

Aquí no estoy entrando en materia de como hacen la detección porque me da la sensación de que igual debería hablar antes sobre HMM, GMM e incluso MFCC

2.2. Violent Event Detection

BIBLIOGRAPHY

[1] D. Barchiesi, D. D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce", *IEEE Signal Processing Magazine*, 2015. DOI: 10.1109/MSP.2014.2326181.

[2] D. Dubois, C. Guastavino, and M. Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories", *Acta Acustica united with Acustica*, 2006.

[3] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. 2006. DOI: 10.1109/9780470043387.

[4] A. J. Eronen *et al.*, "Audio-based context recognition", in *IEEE Transactions on Audio, Speech and Language Processing*, 2006. DOI: 10.1109/TSA.2005.854103.

[5] M. Bahoura, "Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes", *Computers in Biology and Medicine*, 2009. DOI: 10.1016/j.compbiomed.2009.06.011.

[6] D. Van Nort, P. Oliveros, and J. Braasch, "Electro/acoustic improvisation and deeply listening machines", *Journal of New Music Research*, vol. 42, no. 4, pp. 303–324, 2013.

[7] A. Temko *et al.*, "Acoustic Event Detection and Classification", in *Computers in the Human Interaction Loop*, 2009, ch. Part II, 7. DOI: 10.1007/978-1-84882-054-8_7.

[8] A. Temko *et al.*, "CLEAR evaluation of acoustic event detection and classification systems", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007. DOI: 10.1007/978-3-540-69568-4_29.

[9] E. S. Sazonov *et al.*, "Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior", *IEEE Transactions on Biomedical Engineering*, 2010. DOI: 10.1109/TBME.2009.2033037.

[10] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools", *Applied Acoustics*, 2014. DOI: 10.1016/j.apacoust.2014.01.001.

[11] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks", in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2016. DOI: 10.1109/ICASSP.2016.7472176.

[12] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and Classification of Acoustic Scenes and Events”, *IEEE Transactions on Multimedia*, 2015. DOI: 10.1109/TMM.2015.2428998.

[13] J. T. Geiger, B. Schuller, and G. Rigoll, “Large-scale audio feature extraction and SVM for acoustic scene classification”, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013. DOI: 10.1109/WASPAA.2013.6701857.

[14] J.-J. Aucouturier, B. Defreville, and F. Pachet, “The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music”, *The Journal of the Acoustical Society of America*, 2007. DOI: 10.1121/1.2750160.

[15] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. Lecun, “Unsupervised learning of sparse features for scalable audio classification”, in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, 2011.

[16] J. Nam, J. Herrera, M. Slaney, and J. Smith, “Learning sparse feature representations for music annotation and retrieval”, in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012*, 2012.

[17] K. Lee, Z. Hyung, and J. Nam, “Acoustic scene classification using sparse feature learning and event-based pooling”, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013. DOI: 10.1109/WASPAA.2013.6701893.

[18] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, “Acoustic event detection in real life recordings”, in *European Signal Processing Conference*, 2010.