uc3m | Universidad Carlos III de Madrid

University Degree in Audiovisual Systems
Academic Year (e.g. 2014-2019)

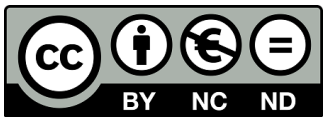
Bachelor Thesis

"Violent event detection from acoustic signals"

Óscar Otero Martínez

Carmen Peláez Moreno

Madrid, January 29, 2020



SUMMARY

Keywords:

DEDICATION

CONTENTS

1. INTRODUCTION 1
1.1. Context 1
1.2. Objectives 1
1.3. Regulatory framework 2
1.4. Socio-economic environment 2
2. STATE-OF-THE-ART 3
2.1. Acoustic Scene Classification and Acoustic Event Detection and Classification 3
2.1.1. Features and methods 3
2.2. Violent Event Detection 5
2.2.1. Gender-based violence 6
2.2.2. Our point of view 7
2.3. Databases 7
3. METHODS 10
3.1. Database: AudioSet 10
3.1.1. What is AudioSet 10
3.1.2. Ontology 10
3.1.3. Dataset 12
3.1.4. Format available 13
3.1.5. VGGish model 13
BIBLIOGRAPHY 14

LIST OF FIGURES

3.1 First two layers of Audio Set Ontology 11

LIST OF TABLES

2.1 Table of studied databases 9

3.1 Fields per category in the ontology 12

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

Tips?

1.4. Socio-economic environment

Tips?

2. STATE-OF-THE-ART

2.1. Acoustic Scene Classification and Acoustic Event Detection and Classification

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, i.e. understanding how the human being perceives 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 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.

notion here refers to ASC. Clear?

While all the advances in the ASC field took place, 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 is set. 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].

Review acronyms list

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 Support Vector Machine Classifier, also known as SVM, 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].

When the target is the acoustic event detection and classification, the working method used 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 techniques 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].

Include picture of the pipeline?

Go deeper? HMM, GMM, MFCC

2.2. Violent Event Detection

All the multimedia information available can be applied to many fields and in different connotations. One of the slopes that has appeared in the acoustic scenes and events sphere is the one applied to violence. For this case, an essential point before addressing any problem is to decide what kind of definition the word violence is going to adopt since it is a really subjective concept. An objective perspective has been given by the World Health Organization as "The intentional use of physical force or power, threatened or actual, against oneself, another person, or against a group or community, that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation" [19]. There exists other definitions found in different works as "physical violence or accident resulting in human injury or pain" [20] and "any situation or action that may cause physical or mental harm to one or more persons" [21].

Recent studies have treated this problem in different ways due to all types of conditions that this may take place in. During the last years, the possibility of creating and providing audiovisual content has grown widely which has led to an enormous amount of multimedia data. Within this content, the variety of topics is uncountable and some of them may be considered unappropriated for certain parts of the audience. This is the reason why there have just been done works related to the field of video content analysis and detection of violence. In some cases, audio and image features have been combined to address these problem [22]. However, it has been found that sound information could be really useful and a more efficient way of working compared to image, since it is easier to process and the cost is lower. Related works have utilized audio features in the time-domain and in the frequency-domain, similar to the ones explained for ASC, then combined with a normal SVM classifier [21]. Other researches have tried more complicated models with the intention of improving the classification task. It is the case of using

DNN, i.e., Deep Neural Networks, fed with both image and audio data, which performs the task more efficiently [23]. Violence detection has been used for other applications such as video surveillance. For example, one of the scenarios for this purpose consists on preventing violent acts inside elevators [24]. For this case, the considered dangerous situations are composed of anti-social actions that are likely to happen in this kind of places, concretely, urinating, vandalism and attacks on vulnerable victims, such as women, children or elderly. The framework proposed is based on audio-visual data, but the master classifier will be driven by audio, due to the possible subtleness of the scenes that are desired to detect. So, first the audio incident detector will trigger the process when a non-silent event takes place. Then, the image processing will begin in order to extract information related to who is involved in the action and how aggressive is it. Another utilization of the surveillance approach is its use for the evolution of smart cities [25]. For this goal, since the system will be implemented in real-life environments, one of the advantages about working with data coming from sounds is the respect for privacy, that, otherwise, using video recordings it would be violated.

The difference in these two applications, apart from the task they are addressing, resides on the data they are working with. For violent content analysis, the data usually comes from fictional audio sources as movies or video-games. However, for real-environment systems, the data is extracted straight from actual day-to-day life situations. In this second case, some disadvantages can be appreciated. For example, the signals are not preprocessed, which means the original properties of the sound are not modified so the processing part before classification becomes tougher. Also, the presence of background noise is more common and loudness of some events, as speech, may vary with time [26]. Makes sense?

2.2.1. Gender-based violence

Throughout history, women have been an object of abuse and suffering in many different situations even though in those that were considered as their familiar surroundings. They have been bashed, sexually harmed and psychologically maltreated by those who were supposed to be one of their closest intimates [27]. In the same way, in the recent times, late studies have shown that 35% of women from all over the world have been victims of physical or sexual damage [28], and 43% of women from Europe have declared going through some psychological or mental violence at least once in their lives [29]. In this context, it is necessary to define the concept of gender-based violence, which can be described as the multitude of harmful behaviours that are focused on women and girls just because of their sex, such as female children and wife abuse, sexual assault, dowry-related murder and marital rape, among others. Particularly, violence against women involves any act of verbal or physical force, extortion or lethal denial which has a woman or girl as a target and provokes the physical or psychological hurt, humiliation or irrational privation of liberty and contributes to continue women subordination [30]. Within this definition, it can be considered that most of the times that these

violent situations take place, they are originated due to persons that are supposed to be part of the victims' closest circle of trust, i.e., their husbands or boyfriends. This is called Intimate Partner Violence (IPV) intimate partner violence (IPV) and it is recognized as a public health problem affecting women across their life span and may resulting in different undesirable unhealthy outcomes, such as depression, chronic pain and even dead [31].

Glossary?

2.2.2. Our point of view

Explicación EMPATIA?

As a contribution to the EMPATIA-TC project developed by Universidad Carlos III de Madrid, the main goal in this work is to make progress in detecting gender-based violence situations, specifically applied to day-to-day scenes, in which IPV is likely to be present. One of the parts from the proposed system is composed by wearable devices that the victim can carry to collect diverse types of information and process them to obtain conclusions and increase the efficiency. Among these accessories, we can find a pendant that pays attention to the user's voice and the surrounding audio to analyse what is happening at a certain moment. For our purpose, the interesting part resides on achieving auditory data so as to detect violent incidents that are formed by sounds already known for characterizing these episodes considered dangerous by the victim.

The definition that is assigned to violence is really important in order to define which audio events should be taken into account. However, considering the subjectiveness of this concept, categorizing violence for every type of user is an extremely difficult task. For this reason, the final idea to answer this question is to make the victim able to decide which kind of hearing events the system must be aware of. In the complete project, this can be carried out by a phone user interface which displays a list of sound events and she has the labour of picking up those that are violent according to her criteria. Since the development of this tool is out of the scope of this work, we have decided to implement a simpler mechanism which will be explained **further on**.

2.3. Databases

A fundamental objective was to find a database that allows for building a system with the desired characteristics, so a rich variety of acoustic events is needed with an essential big representation of violent sounds. In the table 2.1 is represented a relation of the different databases that have been considered for the realization of this work.

The last three options shown in table 2.1 are the ones that better adapted to the problem of the work. *VSD Benchmark* was the first option we were happy with. Within the two ways of working given, the movies and the YouTube videos, the former was the easiest to use since the annotation specified exactly what kind of violent events were present in the scene and the onset and offset within the whole film. However, to access the data it was necessary to pay. The latter was alright but it just indicated the presence of violence,

without determining the type of event. Another choice was *Freesound dataset* because it is formed with all type of videos so we could extract those classes that are more interesting for us. However, it is still in an annotation process and it is not ready to download yet. As a final conclusion, we decided to go for *AudioSet*, which **will be explained further on**.

Name	Description	Considerations
URBAN-SED [32]	10,000 soundscapes with sound events. Every soundscape contains 1 to 9 sound events with strong annotations.	Events are completely specified but it just contains three interesting types of classes.
UPC-TALP [33]	It belongs to CHIL project, for the AED task. Isolated acoustic events that occur in a meeting room environment.	Payment is needed to achieve the data and the classes are a little out of our topic.
MIVIA: Audio Events Data Set for Surveillance Applications [34]	6,000 events with background noise.	The classes included belong to our topic, but they are just three: glass breaking, gun shots and screams.
TUT rare sound events [35]	Source files for creating mixtures of rare sound events (classes baby cry, gun shot, glass break) with background audio.	Similar problem to MIVIA: just from three interesting classes.
IEEE AASP Challenge [36]	Composed by ASC and AED. It is formed by two subtasks: OL (Office-live) and OS (Office Synthetic)	Labels for both subtasks are out of our scope since they are likely to happen in an office environment: keyboard clicks, hitting table, etc.
TUT-SED Synthetic 2016 [37]	Isolated sound event samples were selected from commercial sound effects	The variety of classes is large enough but for our purpose just four of them are useful.
VSD benchmark [38]	Violent events from 32 Hollywood movies and 86 YouTube web videos, together with high-level audio and video concepts.	Payment is needed to purchase the movies and the videos do not specify the type of violent event
AudioSet [39]	An ontology of 632 audio event classes and a collection of 2,084,320 human labeled 10-seconds sound clips from YouTube videos.	The final pick. Plenty of the videos have more than one audio label but we were able to adapt the data to the problem because of the huge amount of clips.
Freesound dataset (FSD) [40]	Filling AudioSet ontology with 297,144 audio samples from Freesound.	This may seem a very good option as well but it is not available yet.

Table 2.1. TABLE OF STUDIED DATABASES

3. METHODS

3.1. Database: AudioSet

3.1.1. What is AudioSet

How cite?

Audio Set can be described as a sound large-scale dataset that has the intention of putting the availability of audio and image data on the same level. It is composed by a huge variety of manually-annotated audio events and is organized by following an important ontology formed by 632 different audio classes. The data has been extracted from YouTube videos and the labelling process has been based on diverse factors such as metadata, context and content analysis. It has been developed by Google with the purpose of producing an audio event recognizer that can be applied to plenty of acoustic situations coming from the real world [39].

3.1.2. Ontology

In order to put this dataset together the events have been organized in an abstract hierarchy. This is composed by higher-level classes which describe a certain type of sound and also acts as parents of other labels that refer to more specific events. With this purpose, the relationship among different classes needed to be non-exclusive, so labelling similar audio events may result into a more general class, the parent, if there existed ambiguity. This is also helpful for labellers due to group the clips in an easier and faster way.

The Audio Set Ontology has been made considering some fundamental guidelines as the ones explained below:

- **A complete collection of all labels must be prepared** so that it can be used to define sound events from real-world aural data.
- When labelling an audio event the result must match the criteria of a common listener.
- Different categories should be easy to distinguish by an ordinary listener. In the case that two different labels do not satisfy this requirement, these should be merged. With this condition, the spectrum of possible labels remains limited.
- The distinction of two different classes must be done by relying just on the audio, it cannot be accompanied by image or visual information.
- The hierarchy should not be very deep by keeping the number of children per parent class to no more than 10. This also eases the annotation labour.

It is easy that an ontology of this volume gets leaned or biased in a particular direction due to several factors, such as the subjectiveness of its creators or the selection of the initial set of classes used when starting to work. With the intention of generating a primary list that covers a wide range of audio events in an objective way, the researchers decided to apply an impartial, web-scale text analysis from the very beginning. They agreed on detecting hyponyms of the word "sound" by utilizing a modified version of the famous technique called *Hearst patterns* [41], a method proposed to automatically acquire hyponymy lexical relations from unrestricted text. As a result, an enormous collection of terms came up. This was filtered by considering how well these terms represented audio, i.e. by combining together the global frequency of occurrence and how exclusively these are recognized as hyponyms of "sound" instead of other terms. As an output of the final process, a list of 3000 terms was obtained.

With respect to the hierarchical relation among categories, it was constructed by the authors with the main intention that this satisfied their human comprehension of the sounds. Event though this is a subjective manner, it also makes sense since this is how the hierarchy best performs its labour on helping human labelling. After all the organization process, the model is not based on a strict behaviour as a singular node can appear in many different locations, i.e., a single node can be child of different parent nodes. The final result is composed by 632 audio event labels and, in the hierarchy, there is no deeper case than 6 levels. Figure 3.1 shows the nodes that belong to the two top layers of the ontology.

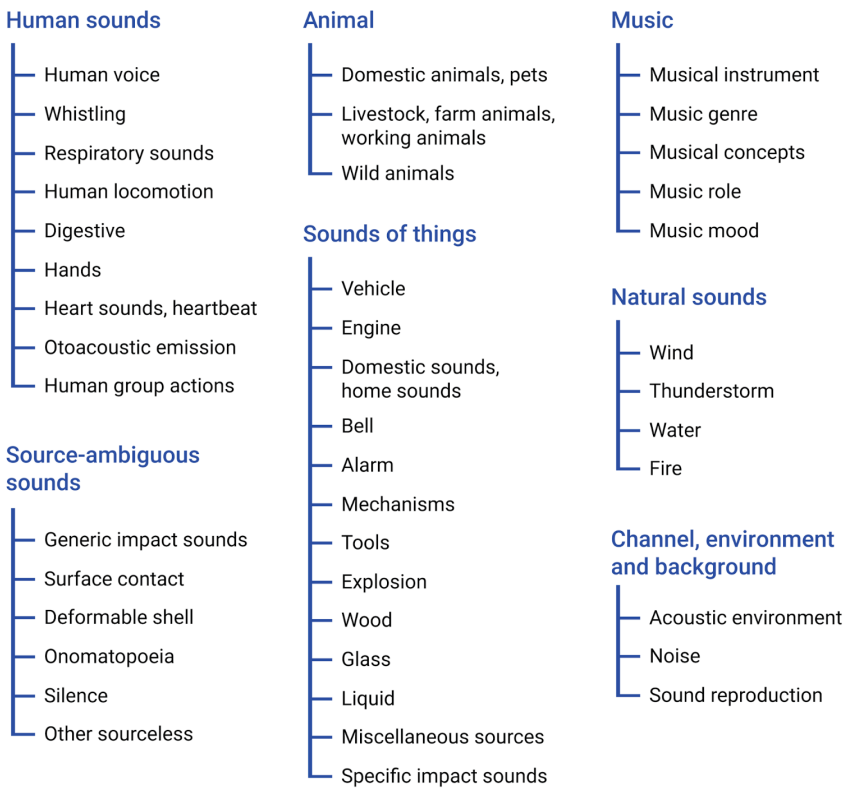


Fig. 3.1. First two layers of Audio Set Ontology

This whole structure has been given to the user as a file in JSON format. A couple of fields have been included for each label in order to describe its meaning and make clear its position within the hierarchy. A description for all of these can be found in table 3.1.

ID	This field includes the Knowledge Graph Machine ID (MID) that best describes the sound or its source. It is used as a primary identifier for the class.
Display name	Short name formed by one or two words that identifies the audio class. It sometimes includes a small alternative separated by a comma so it does not feel ambiguous.
Description	One or two explaining sentences so the meaning of the category is more defined. These can be extracted from Wikipedia or WordNet.
Examples	At least one example of the label is provided as a URL of a YouTube video.
Children	An array filled by the Machine ID (MID)s from all immediate children of the class.
Restrictions	It specifies if the category in question either has been discarded or there are no audio clips under it.

Table 3.1. FIELDS PER CATEGORY IN THE ONTOLOGY

For the field *ID*, the identifiers are known as Machine ID (MID) and belong to the *Knowledge Graph* designed by Google [42]. This is a knowledge base that Google services use to improve the quality of its search results and it is composed with information extracted from a wide variety of sources. The MIDs are the identifiers of the different elements that belong to this huge dictionary. For instance, the MID of the word "Speech" is "/m/09x0r". Another field that deserves a special explanation is the one corresponding to *Restrictions*. Within all the categories of sound events, there are two flags that indicate an exclusive behaviour that differ from a typical label: "blacklisted" and "abstract". The former refers to a class that has been hidden from labellers due to its confusing meaning. The latter has been used for those classes that are just utilized as intermediate nodes in order to provide a better grouping inside the organization, and are not expected to be used in the implementation tasks. In total, out of the 632 categories, 56 have been categorized as "blacklisted" and 22, as "abstract".

3.1.3. Dataset

The different YouTube videos that constitute the dataset are included in a CSV file in which each row is formed by the video identifier, the start and end time of the audio event within the video and the ontology labels that the certain clip belongs to. All the video

segments have a longitude of 10 seconds maximum, except from those that the original video is shorter.

The final release of the dataset is composed by 1,789,621 segments, with a duration of 4,971 hours of video and audio. After executing the selection process in which the different labels were populated with the final corresponding segments, a total of 527 classes were gathered, out of which 485 counted with at least 100 samples.

Include any explanation about ratings?

3.1.4. Format available

The data can be obtained through the website [43] in two different formats:

- Files in csv format that include for each video segment its YouTube video ID, start time, end time and the one or more labels it belongs to.
- Instead of the audio files themselves, they provide already extracted audio features for each segment in compressed files that can be easily downloaded.

For our purpose, we worked with the dataset in both different ways. However, we get further with the already extracted features. These are obtained by using a model called *VGGish* [44] which is inspired on the well known VGG neural network [45]. The released model has been pre trained on a preliminary version of the database YouTube-8M [46]. The features are available to the user in TensorFlow record files and the VGGish model is also included in a public repository.

cite for TensorFlow

3.1.5. VGGish model

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.

[19] E. G. Krug, J. A. Mercy, L. L. Dahlberg, and A. B. Zwi, "The world report on violence and health", *Lancet*, 2002. DOI: 10.1016/S0140-6736(02)11133-0.

[20] C. H. Demarty *et al.*, "The MediaEval 2013 affect task: Violent Scenes Detection", in *CEUR Workshop Proceedings*, 2013.

[21] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006. DOI: 10.1007/11752912_55.

[22] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010. DOI: 10.1007/978-3-642-12842-4_13.

[23] A. Ali and N. Senan, "Violence video classification performance using deep neural networks", *Advances in Intelligent Systems and Computing*, vol. 700, pp. 225–233, 2018. DOI: 10.1007/978-3-319-72550-5_22.

[24] T. W. Chua, K. Leman, and F. Gao, "Hierarchical audio-visual surveillance for passenger elevators", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014. DOI: 10.1007/978-3-319-04117-9_5.

[25] J. García-Gómez, M. Bautista-Durán, R. Gil-Pita, I. Mohino-Herranz, and M. Rosa-Zurera, "Violence detection in real environments for smart cities", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016. DOI: 10.1007/978-3-319-48799-1_52.

[26] M. Bautista-Duran *et al.*, "Acoustic detection of violence in real and fictional environments", in *ICPRAM 2017 - Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, 2017. DOI: 10.5220/0006195004560462.

[27] United Nations, *No Title*, 1989.

[28] WHO. Department of Reproductive Health Research. London School of Hygiene and Tropical Medicine. South African Medical Research Council., *WHO | Global and regional estimates of violence against women*. 2013.

[29] European Union Agency for Fundamental Rights, *Violence against women : An EU-wide survey*. 2014. DOI: 10.2811/62230.

[30] L. Heise, M. Ellsberg, and M. Gottemoeller, *Ending violence against women.* 1999. DOI: 10.4324/9780429269516-5.

[31] K. Beyer, A. B. Wallis, and L. K. Hamberger, "Neighborhood Environment and Intimate Partner Violence: A Systematic Review", *Trauma, Violence, and Abuse*, 2015. DOI: 10.1177/1524838013515758.

[32] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation", in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017. DOI: 10.1109/WASPAA.2017.8170052.

[33] V. Mapell, *UPC-TALP database of isolated meeting-room acoustic events*, 2012. [Online]. Available: http://catalog.elra.info/en-us/repository/browse/ELRA-S0268/.

[34] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments", *Pattern Recognition Letters*, 2015. DOI: 10.1016/j.patrec.2015.06.026.

[35] E. Fagerlund and A. Hiltunen, *TUT Rare sound events*, 2017.

[36] D. Stowell and E. Benetos, *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events*, 2013.

[37] E. Cakir and T. Heittola, *TUT-SED Synthetic*, 2016.

[38] C. H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "VSD, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation", *Multimedia Tools and Applications*, 2015. DOI: 10.1007/s11042-014-1984-4.

[39] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events", in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017. DOI: 10.1109/ICASSP.2017.7952261.

[40] E. Fonseca *et al.*, "Freesound datasets: A platform for the creation of open audio datasets", in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017*, 2017.

[41] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora", 1992. DOI: 10.3115/992133.992154.

[42] A. Singhal, *Introducing the Knowledge Graph: things, not strings*, 2012.

[43] Sound Understanding group, *AudioSet*, 2017.

[44] S. Hershey *et al.*, "CNN architectures for large-scale audio classification", in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017. DOI: 10.1109/ICASSP.2017.7952132. arXiv: 1609.09430.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015. arXiv: 1409.1556.

[46] Video Understanding Group, *YouTube-8M*, 2017.

ACRONYMS