Supporting the Detection of Sexual Content in Videos by Listening for Characteristic Sounds

XAVER HIMMELSBACH, Hochschule Darmstadt, Germany

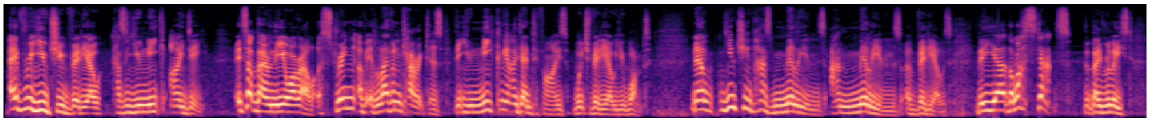

Fig. 1. Example spectrogram showing speech and noises from a YouTube video[13].

Detecting sexual content is necessary for any online service that handles user-generated media.

In videos, the basic approach is slicing frames out of the video file and rating them using neural networks.

Multi-modal approaches combine multiple file data sources, usually video frames and audio tracks, to create a more exact content rating.

Sexual content in videos has some characteristic sounds which facilitate its detection. The goal is to train a neural network to detect these sounds and thus support the visual rating component.

The implemented model is trained with a custom data set containing the audio tracks of pornographic videos.

While the model's performance is mixed, the fact that some characteristic noises are detected with a limited data set shows that the approach could work very well with an expanded data set.

CCS Concepts: • **Social and professional topics → Pornography**; • **Computing methodologies → Machine learning**; **Object detection**; *Image representations.*

Additional Key Words and Phrases: Audio/Video, Machine Learning, Text/Speech/Language, Dataset, Content Analysis, Prototyping/Implementation

ACM Reference Format:

Xaver Himmelsbach. 2023. Supporting the Detection of Sexual Content in Videos by Listening for Characteristic Sounds. In . ACM, New York, NY, USA, 13 pages.

1 INTRODUCTION

Nudity detection is an umbrella term for programmatically detecting nudity and sexual content in images and videos[2, p. 2]. Sexual content detection is an extension of this term, intended to cover nudity and explicit sexual acts, even if they involve little nudity.

Sexual content detection is often used by social media platforms to automatically detect and block illegal media that violates minors' protection or the platforms' guidelines[15, p. 1 f.]. Some markets are more heavily regulated than others, for example, the German one, where violations of youth protection can result in heavy fines or imprisonment[9, p. 23 ff.]. Therefore, it is crucial to closely examine the uploaded media for inappropriate content.

Manuscript submitted to ACM

However, sexual content detection can not only be used to block media but also to analyze content automatically. Adult entertainment site operators, especially in markets with strict parental restrictions, must assign the correct age ratings to uploaded files so that only users of legal age can access them. Manual checks of these ratings can be partially automated using sexual content detection. Keywords and content tags can also be set automatically as machine learning applications become more sophisticated.

The main problem with these systems is the recognition of context. Distinguishing between swimwear on the beach, lingerie in the bedroom, medical images and erotic films, or nude art and pornographic illustrations poses problems for the recognition process by algorithms and artificial intelligence. Situations in which the persons depicted cannot be recognized appropriately but perform sexual acts are also challenging to analyze.

This paper presents an approach to improving systems for recognizing sexual acts in videos, which is not based on image recognition, as most contemporary solutions are, but on sound recognition. Recognition should thus be improved in situations where sexual acts are difficult to classify in video frames. If sexual sounds occur on the audio track simultaneously, the recognition rate can be increased.

2 STATE OF RESEARCH

Sexual content detection is traditionally done visually. This approach is appropriate for examining single images where no other data is available[8, p. 1 ff.]. However, other data sources, most notably the audio track, are frequently neglected when examining videos.

2.1 Still Image Recognition

Older approaches to sexual content detection are based on analytical computer vision. The simplest way is to use skin detection algorithms that check the amount of skin-colored pixels in an image. Hence they detect nudity when it is above a set threshold[2, p. 1 ff.]. More advanced approaches use feature detection algorithms. Here, the corners and edges of characteristic structures are extracted from the image using convolution functions and compared with known structures, such as genitals[8, p. 2].

Modern approaches to nudity detection are often implemented using machine learning[8, p. 2]. Images are processed by neural networks that return either classification of the input, i.e., whether it contains nudity or a list of the objects present in the input with their position. In the context of sexual content detection, this mainly includes genitals.

Videos can also be checked with such systems. For this purpose, the individual images are extracted from a video and examined by the nudity detection system like a regular image[8, p. 2]. The disadvantage here is that other possible data sources are neglected. Multi-modal approaches can be used to inspect other data sources as well.

2.2 Multi-Modal Approaches

In multi-modal solutions, other data sources are examined in addition to the individual images of the video. Examples include the differences between individual images, the subtitles provided, or the soundtrack[5, p. 4 ff.]. For example, some implementations apply speech recognition to the audio track of a video and match the recognized words with a bad-word list. Due to the more extensive data basis, the systems can recognize sexual content more accurately and possibly even classify situations that would not have been recognized on a visual basis alone.

Noise analysis is one of the most promising areas as a large amount of data is available here which has not been used previously. While spoken words are already analyzed in some solutions[5, p. 6 ff.], little attention has been paid to sounds that do not represent speech[6, p. 1].

3 IDEA

A neural network will be trained to recognize sounds characteristic of sexual acts. This model can be combined with other systems for recognizing sexual content, especially ones working on image data. The model is intended to handle cases in which the other systems, especially the visual ones, miss explicit content.

3.1 Characteristic Sounds

Understanding the characteristic sounds of sexual acts is required to train a model for explicit sound recognition. Since this work is a proof of concept, no detailed analysis of the occurrence of specific sounds in pornographic content is done here. For this purpose, it is sufficient to pick several easily recognizable sounds on which the model is trained. However, selecting sounds with distinct sound patterns is crucial, as otherwise, there is a risk of confusing similar sounds that are not indicative of sexual activity[6, p. 2 f.]. In the context of pornographic films, from which the training material must come due to their availability, moaning, screaming, or wet sounds are promising candidates. These types of sounds rarely appear in videos with non-explicit content. They are, however, an integral part of sexually explicit videos.

3.2 Model

These sounds are detected in videos that can be of various lengths. In addition, sounds can appear several times in succession and superimposed over each other. For this reason, AI models for object recognition, also called object classification, are not applicable here[12, p. 4]. They can only decide whether a particular sound is on an audio track and need audio tracks of a normalized length as input. Although it is possible to split an audio track into individual elements of the length that an AI for object recognition can process, it implies additional avoidable work and hurts the model's efficiency. Further, sounds that extend across interfaces or continue over a longer period of time can only be recognized with unsightly solutions, such as starting the cuts from different starting positions or scaling the track.

A cleaner solution is to use object detection models[12, p. 4 f.]. They can detect sounds on audio tracks and return their position. Additionally, they can detect multiple occurrences of the same sound and overlaps of multiple sounds.

Object detection and object recognition models work primarily based on images[12, p. 2 ff.]. They cannot analyze audio files directly, so converting them to an image format is necessary before processing.

This conversion creates a frequency graph, an image whose X-axis represents the elapsed time on the audio track[6, p. 1 f.]. This time is not measured in seconds or milliseconds but samples. Using the sample rate of an audio file, which describes how many samples occur per second, it is thus also possible to convert the time points to seconds by dividing the index of the current sample by the sample rate. The Y-axis represents the frequency bands of the audio file. At each sample, certain frequency bands of the audio track are activated to different degrees, which can be seen in the frequency image by the color value of the respective pixel. Different sounds result from those frequencies, which are activated over specific periods of time. Since the algorithm is now based on images, conventional object detection models can also be used.

In the example spectrogram shown below, the excitation of the different frequency bands can be seen at successive sampling points.

The respective strength of the excitation in decibels can be read off the color bar. From the range of about 6000 Hertz upwards, a reduction in the activation of the frequencies can be observed. These high frequencies are hardly perceived by the human auditory system and are thus considered unimportant for noise detection tailored towards human noises.

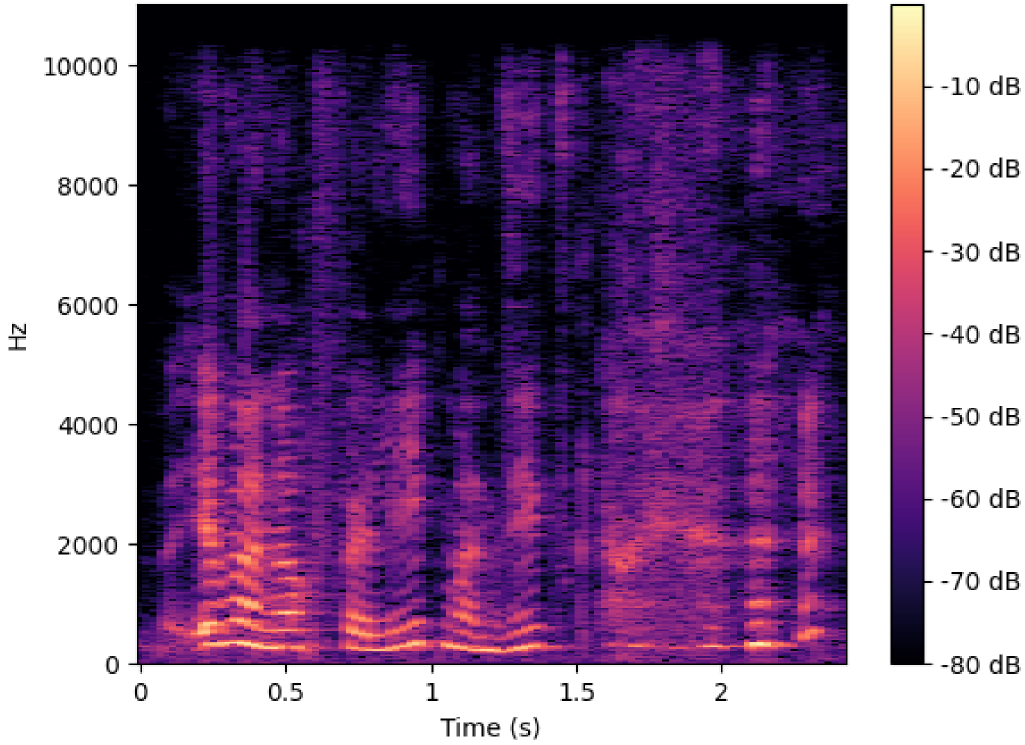

Fig. 2. Spectrogram showing the frequency distribution for a few seconds of speech and laughing.

This circumstance of human hearing will also be relevant for further processing spectrograms before passing them to the AI model.

4 IMPLEMENTATION

A data set must be created to implement the noise detection model. The selected AI model is then trained and validated with this data set.

4.1 Training Data Set

Spectrograms have to be additionally processed before being added to the data set. The sounds of interest are naturally those to which the human auditory system reacts particularly strongly. Mel spectrograms are used to represent this circumstance. These are not linear but logarithmic spectrograms based on the mel scale that was developed to approximate human hearing in signal processing. On the mel scale, low-frequency bands take up significantly more space than the high-frequency bands, just like the human auditory system can process differences between low frequencies much better than high ones[15, p. 2]. By focusing on the frequency bands relevant to human hearing, the AI model should also be able to identify sounds similar to humans.

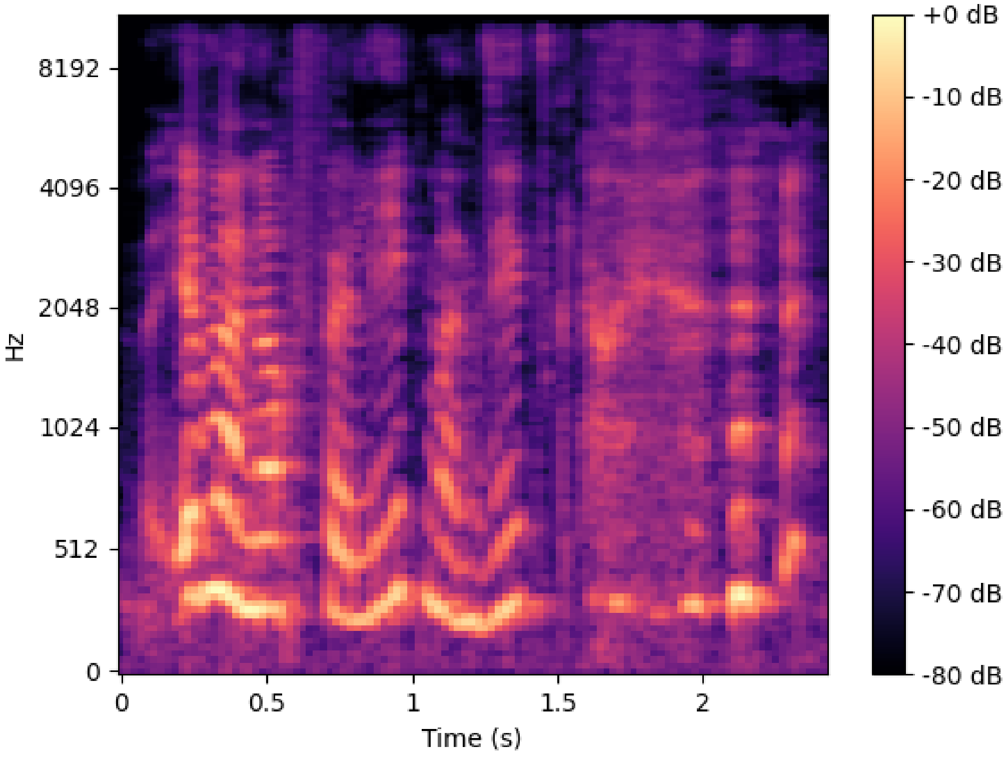

Fig. 3. Mel spectrogram showing the frequency distribution for a few seconds of speech and laughing.

Comparing the Mel spectrogram with the linear spectrogram of the same sound section from page 4 shows how the lower frequencies are stretched and take up a more significant part of the diagram. On the other hand, the relevance of the high-frequency ranges has shrunk considerably. This representation approximates the functionality of human hearing. High-frequency band compression is also applied in other areas of signal processing, especially in image, audio, and video compression[1, p. 206 ff.]. Here, too, the high frequencies are irrelevant for human comprehension, hence can be reduced without significant disadvantages.

No existing data sets contain audio tracks labeled for the occurrence of sounds characteristic of sexual acts. The training data set must be explicitly created for this paper. The annotation process will focus on one type of sound: moans, in particular, "porn moans." They refer to the loud and exaggerated moaning sounds generally associated with pornographic videos. Such moans should be easily identified in the annotation process and have a characteristic pattern in the spectrogram that the model can recognize.

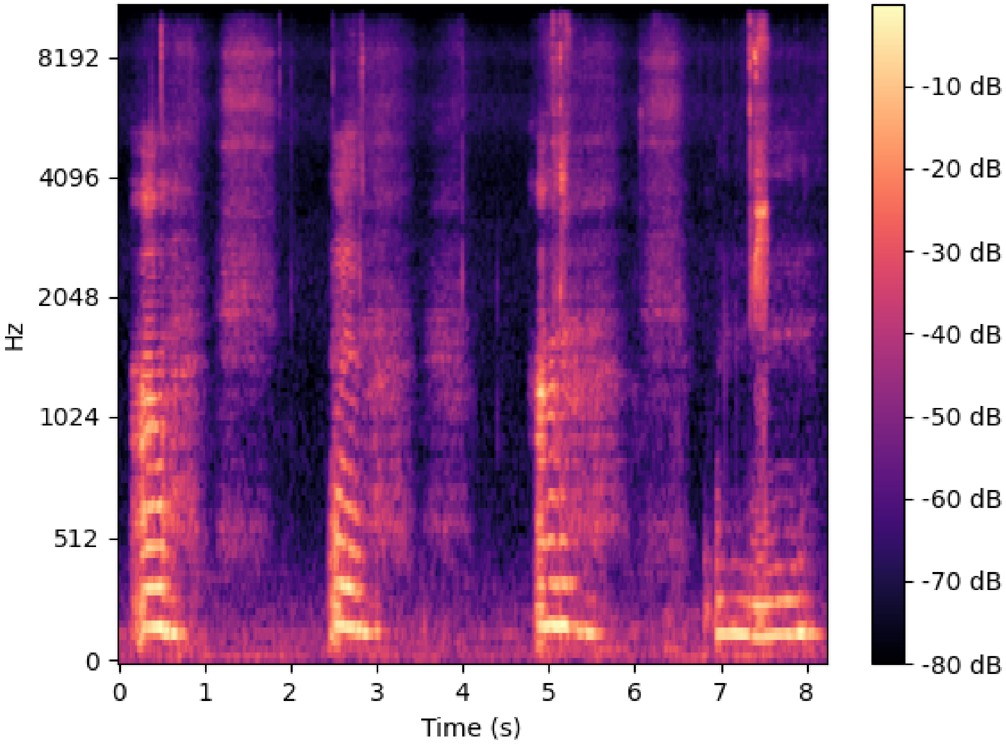

Fig. 4. Mel spectrogram of four successive moans that meet the criteria to be detected.

The diagram above shows successive moans occurring in an audio file at seconds 0, 2.5, 5, and 7. The first three moans are very similar in structure. First, all frequency bands are intensely excited for a short time, especially the low ones. Second, comes a decay in which the low frequencies are more weakly excited, while there is no more excitation in the high frequencies. A kissing noise coincides with the fourth moan, characterized by a continuous excitation of all frequencies at 7.4 seconds. Such mixing and impurities of the sounds are to be expected, for example, when several people make noises or music is involved. The model should still detect moans correctly in these situations.

4.1.1 Data Set Workflow.

A script was written to automate the data set's annotation as much as possible. The script can be found at src/create_dataset_script.py in the provided source code. The directory src/create_dataset/ contains the additional scripts that handle downloading and converting files.

This script reads in a list of URLs. The corresponding video is downloaded for each URL in this list using YT-DLP[16]. The next step is to annotate the video, where manual intervention is required. The annotation program "Praat" is used for annotation[4].

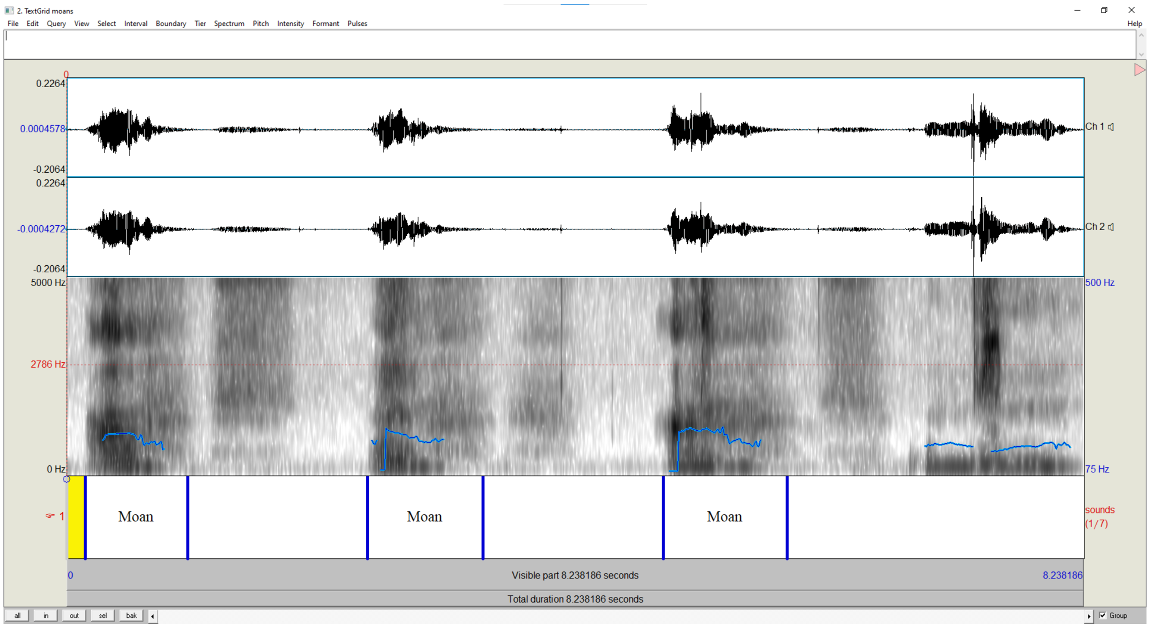

Fig. 5. Annotation interface of Praat.

Praat allows loading an audio file and annotating labels on different tiers. When creating a new annotation file, the default tiers must be replaced. Only one tier called "sounds" is used for this application. In this tier, moaning sounds are marked with the label "Moan".

Praat creates files in TextGrid format that contain the annotations. The format contains the length and structure of the audio tracks. TextGrid files also list all annotated labels by position and name[3]. Since the Object Detection Framework does not support the TextGrid format, it must first be converted.

The TextGrid file gets converted to an object with a TextGrid parser and written into the XML structure required by the AI framework. In addition, the corresponding Mel spectrogram for the audio track is generated and saved as a PNG file. Finally, the XML and the PNG file are copied into the folder structure required for training, and the next URL is processed.

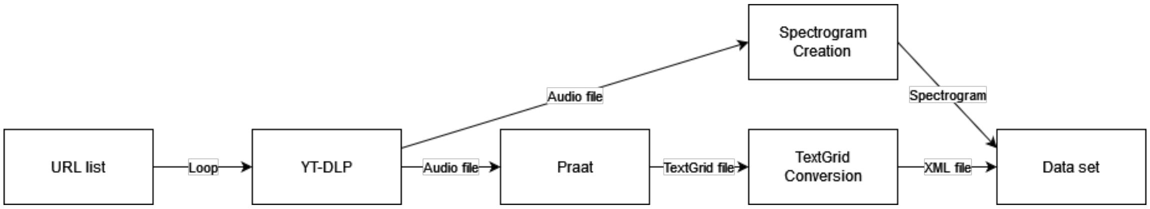

Fig. 6. Data set workflow.

4.1.2 Data. For training, 100 videos from Pornhub's top-rated category were used. These videos were processed with the data set workflow; half were used as training and half as validation data. The videos were all about 5 to 7 minutes long. The total time was 510 minutes and 20 seconds, spread over 83 videos, as a significant number of videos had to be removed from the data set due to insufficient GPU VRAM. A complete list of used videos can be found in "log.csv" in the provided data set.

4.2 Model

Since this paper is mainly concerned with detecting characteristic sounds, a pre-built framework has been used to implement the machine learning model. ImageAI is an AI framework that allows the creation of an object detection model based on existing models. It facilitates training, validation, and prediction of image data[10].

The model used as a base is YOLOv3, where YOLO stands for You Only Look Once[7]. YOLOv3 is a relatively fast and lightweight model, mainly because it funnels the examined image through the network only one time to detect the objects. The integrated Darknet-53 feature extraction model generates feature maps at different scales on which the actual object detection is performed[11, p. 2 f.]. For each pixel on the feature maps, bounding boxes are created These boxes are checked for overlap with the searched objects. If a bounding box overlaps an object better than a previous one, it is adopted as the new best box[11, p. 1 f.]. With ImageAI and YOLOv3 as a base, the trained model outputs all best-fitting bounding boxes for the objects it has been trained to detect.

4.3 Training, Validation, and Prediction

Training, validation, and prediction of the model are implemented via the scripts train.py, validate.py, and predict.py in the src directory.

The first step in each of these scripts is to call an external script that activates memory growth if a graphics card is available. If memory growth is not enabled, graphics card acceleration is not available. The training process involves configuring ImageAI's Detection Model Trainer appropriately, selecting the YOLOv3 model, and starting the training process.

Once the training is completed, validation can be started. Here, all models generated during training are examined and validated. The model with the lowest error rate can then be used for prediction.

The selected model receives an audio track as input and outputs all detected moaning sounds with their position in the image on the command line.

The libraries cuda and cudnn must be installed to access graphics card acceleration when training the model. One problem encountered during the training process on several PCs was that the shared library "libcusolver.so.10" could not be found, as only "libcusolver.so.11" is supplied with the current versions of cuda. Creating a symbolic link from "libcusolver.so.10" to "libcusolver.so.11" on the host system could not fix this problem. What helped was to install the cuda libraries with an Anaconda virtual environment and create the symbolic link inside of that.

5 RESULTS

After implementing and training the model, the model's performance can be analyzed. However, several challenges arose during the implementation of the model, which affected its results.

5.1 Challenges

The reason for the annotated videos all being only 5 to 7 minutes long is that ImageAI crashes when processing longer inputs. The Pillow library, used to manipulate the spectrograms, can no longer process them above a specific image size. The cutoff size seems to be around 1.3 trillion pixels on the used hardware. The selected videos thus have to be shorter than that. There are possible solutions to this problem discussed in 6, but these could not be implemented within the scope of this paper.

A significant challenge was annotating the data. Labeling the 510 minutes of audio material took over two weeks. The resulting amount of training material is, nevertheless, only sufficient for mixed prediction results.

Furthermore, selecting moans as the primary noise had consequences for classification. Exaggerated moaning sounds did indeed appear on some audio tracks, especially those belonging to videos produced by big studios, which is to be expected. Such sounds were much harder to find in content produced by amateurs, especially by individuals. Since the audio tracks would not have contained any labels otherwise, less explicit noises were labeled as moans during a first annotation pass. These included more subdued moans, groaning, or heavy panting. Testing the model with these annotations revealed that the variance in training data confused it heavily. It began to detect all kinds of minuscule frequency patterns as moans. Ultimately, it was decided that it was more reasonable for a proof of concept to focus on the few clear examples found within the training data, even if that meant ignoring potentially viable moans.

5.2 Test Usage

Two other videos that the model had not yet seen during training and validation were selected to explore its performance. The first video shows pornographic content; the second does not. For each moaning sound detected by the model, a check is made to see whether an actual moan occurred at that timestamp. In addition, it is examined whether sexual content can be seen simultaneously, which would result in a higher age rating of the video at this point.

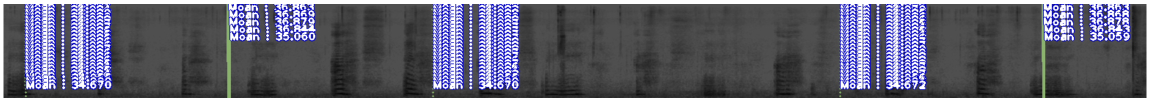

Fig. 7. Moans detected in a sample audio file.

If multiple moans occur in succession over a short period of time, they are combined into a single sound to facilitate the analysis and presentation of the data. Aggregation is necessary to receive satisfactory results because the AI detects many tiny moans in the spectrogram that are very strongly locally concentrated. Although this distribution is not the expected result, there is hope that moans can be detected by aggregating these local detections. In the following examples, sounds were aggregated over one second. If there was a bigger gap between two noises, they were considered separate detections.

In the example above, the green lines represent the bounding boxes of the detected moaning sounds, while the white text represents the labels. Due to the number of detected sounds, about 100 in the shown section, overlaps occur. The image was created with the default image output of ImageAI.

In the first video, 23 out of 31 detected moans were correct. In the second video, all 33 detected moans were incorrect. The result is a prediction accuracy of 74.2 % for the first video and 32.8 % for both videos.

The quality of detecting moaning sounds in videos that do not show pornographic content is very mixed. In some cases, moaning sounds are even detected in near-silence. However, they are often discovered at points on the track where actors take a breath or emphasize short words. Here, false detections are expected because such sounds' spectrograms are similar to moans.

In the first video, a significant amount of moans were not recognized, for example, at 107, 189, or 269 seconds. These moans are not noticeably different from the correctly detected ones. It may be noted that moans produced by performers

Table 1. Detected moans in the explicit test video https://pornhub.com/view_video.php?viewkey=ph5d828dc686b12

Moan Detected (from - to in s)	Correct Detection?	Sexual Content Visible?
41.23 - 41.25	no	yes
68.81 - 68.84	no	yes
96.47 - 96.49	no	yes
114.25 - 114.25	yes	yes
121.17 - 121.17	yes	yes
124.10 - 124.12	yes	yes
128.07 - 128.07	no	yes
151.73 - 151.82	yes	yes
169.49 - 169.49	no	yes
176.39 - 176.39	no	yes
179.36 - 179.43	yes	yes
197.14 - 197.14	yes	yes
203.95 - 204.09	yes	yes
206.99 - 206.99	yes	yes
210.86 - 211.00	yes	yes
217.78 - 217.90	yes	yes
224.70 - 224.80	yes	yes
231.62 - 231.69	yes	yes
234.57 - 234.62	yes	yes
238.54 - 238.63	yes	yes
245.44 - 245.55	yes	yes
252.36 - 252.50	yes	yes
259.28 - 259.41	yes	yes
262.20 - 262.25	yes	yes
266.15 - 266.29	yes	yes
273.04 - 273.18	yes	yes
279.99 - 280.13	yes	yes
286.93 - 287.07	yes	yes
289.83 - 289.88	yes	yes
293.85 - 293.92	no	yes
317.53 - 317.58	no	yes

with deep voices were hardly recognized, even though they did occur in the video. However, this is probably due to inaccuracies in the annotation process, which now show their consequences as emerging biases.

5.3 Possible Biases

Biases are an issue to be aware of when using artificial intelligence. They distort the model's results and can result from erroneous measurements, incorrect annotation, distorted patterns in the training data, or incorrect interpretation of the predictions[14, p. 1 f.]. Complete elimination of biases is unrealistic because this is just a proof of concept. However, it is essential to mention the biases that may occur here and keep them in mind when analyzing the concept.

The most likely bias is a sampling biases due to the focus on moaning sounds from high-pitched voices[14, p. 3]. Since the majority of videos in Pornhub's top-rated category feature heterosexual or lesbian performers, and there is a much greater focus on the performance of female performers in these videos, there are hardly any moans from people with deep voices in the data set. The latter are thus less easily recognized by the AI.

Table 2. Detected moans in the test video https://www.youtube.com/watch?v=gocwRvLhDf8

Moan Detected (from - to in s)	Correct Detection?	Sexual Content Visible?
33.06 - 33.06	no	no
39.05 - 39.05	no	no
45.02 - 45.02	no	no
50.98 - 50.98	no	no
56.90 - 56.90	no	no
62.87 - 62.87	no	no
68.86 - 68.86	no	no
74.80 - 74.80	no	no
80.77 - 80.77	no	no
86.72 - 86.72	no	no
92.68 - 92.68	no	no
146.34 - 146.34	no	no
152.33 - 152.33	no	no
158.25 - 158.25	no	no
164.24 - 164.24	no	no
170.21 - 170.21	no	no
176.15 - 176.15	no	no
182.12 - 182.12	no	no
188.04 - 188.04	no	no
194.01 - 194.01	no	no
199.97 - 199.97	no	no
205.92 - 205.92	no	no
211.88 - 211.88	no	no
217.87 - 217.87	no	no
223.82 - 223.82	no	no
229.81 - 229.81	no	no
235.73 - 235.73	no	no
241.70 - 241.70	no	no
247.66 - 247.66	no	no
253.63 - 253.63	no	no
259.57 - 259.57	no	no
265.40 - 265.54	no	no
271.51 - 271.51	no	no

Another manifestation of sampling bias is focusing on one type of sound, moans. Other types of sounds, such as licking, groaning, or slapping, were not considered, even though these would also indicate the existence of sexual activity. Negative examples were also not annotated[14, p. 4 f.]. For example, if parts of the audio tracks in which speech is heard had been consistently labeled as speech. Possibly no moans would have been detected in these sections, only speech. When evaluating the results, these sections could then have been ignored.

Label bias probably also occurred during the annotation process[14, p. 4]. This is caused by inattention and inconsistencies when creating the labels. Again, the effect is inaccurate predictions after training the AI on the labeled data.

In principle, however, these biases do not detract from the validity of the proof of concept.

6 FURTHER RESEARCH

Although the implemented approach produced mixed results, the underlying concept shows potential.

The approach can be used as a complementary application to AI models for visually detecting sexual content. By checking the audio track, the auditory model can mark areas of the video with a high probability of sexual activity. If the visual AI has not detected explicit content in these areas, it can examine them again with stricter rules. Marking the areas for investigation by a human could also be a suitable measure. The model for detecting characteristic sounds does not have to be the only component that analyses the audio track. It can also be used with AIs that read the conversations in a video and compare them with bad-word lists to achieve even more accurate results[5].

Several optimizations could improve the models' predictions.

The first possibility is to increase the amount of training data. If a dedicated data set were to be created systematically, the results of the AI would undoubtedly improve. Publishing this data set as open source would be an excellent step to drive forward innovation.

In the course of this, other sounds could also be annotated. There are many other indicators of sexual activity, such as groans, clapping, or gagging sounds. These must map to clear patterns in the spectrogram to make them recognizable for the model. For example, claps show up as single, loud spikes across all frequency bands, while groans activate the low frequency ranges over an extended period of time.

As a further step, one could look beyond the spectrograms of individual sounds and examine the patterns in which they appear. If similar slapping sounds are heard at the same intervals, this could indicate rhythmic sexual intercourse, for example.

The crashes of ImageAIs Pillow library when working with huge spectrograms could be combatted in two ways.

On the one hand, spectrograms can be divided into several smaller image files examined individually by the AI. In this case, the intersections must be considered separately so as not to miss sounds beyond an intersection. A similar effect could also be achieved by reducing the audio file's sample rate. This reduction, however, would also impact detection quality.

On the other hand, the structure of the neural network can be optimized. For the examination of an audio track, the Y-axis is generally irrelevant. The goal is to find the points in time at which certain sounds occur. If one uses a convolutional neural network that successively reduces the Y-axis of the spectrogram to a height of one pixel, significantly longer audio files could be processed. In addition, this can minimize the amount of superfluous detected moans since only one pixel is relevant for noise detection at any given time. With such an architecture, one could also use a simpler neural network and directly examine the activation of the output neurons of the network, which then output one of several predefined values at a certain point in time that map to detected sounds.

Overall, with the appropriate improvements, the concept of supporting visual detection by listening for characteristic sounds could be used in the future. The approach might not only help detect sexual content in videos. A noise detection network could also assist in building surveillance or machine maintenance systems, wherever the occurrence of specific sounds may be relevant.

REFERENCES

[1] 1992. *Proceedings of SPIE*. Bellingham, Wash.

[2] Rigan Ap-apid. 2005. *An Algorithm for Nudity Detection*. Ph.D. Dissertation. De La Salle University, Manila. https://www.researchgate.net/profile/Rigan-Ap-Apid/publication/249767252_An_Algorithm_for_Nudity_Detection/links/5693b99108aeab58a9a2a57a/An-Algorithm-for-Nudity-Detection.pdf

[3] Paul Boersma and David Weenink. 2018. TextGrid file formats. https://www.fon.hum.uva.nl/praat/manual/TextGrid_file_formats.html

[4] Paul Boersma and David Weenink. 2022. Praat: Doing Phonetics by Computer. https://www.fon.hum.uva.nl/praat/

[5] M. Y. Chuttur and A. Nazurally. 2022. A multi-modal approach to detect inappropriate cartoon video contents using deep learning networks. *Multimedia Tools and Applications* (2022). https://doi.org/10.1007/s11042-022-12709-2

[6] Jonathan Dennis, Huy Dat Tran, and Haizhou Li. 2011. Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions. *IEEE Signal Processing Letters* 18, 2 (2011), 130–133. https://doi.org/10.1109/LSP.2010.2100380

[7] Ali Farhadi. 2022. YOLO: Real-Time Object Detection. https://pjreddie.com/darknet/yolo/

[8] Abnishek Gangwar, Eduardo Fidalgo, Enrique Alegre, and Victor González-Castro. [n.d.]. Pornography and child sexual abuse detection in image and video: A comparative evaluation. In *8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017)*. https://gvis.unileon.es/wp-content/uploads/2018/09/PAPERID_82_ICDP2017_Camera_Ready_Pornography-and-Child-Sexual-Abuse-Detection-in-Image-and-Video.pdf

[9] Kommission für Jugendmedienschutz. 2003. Staatsvertrag über den Schutz der Menschenwürde und den Jugendschutz in Rundfunk und Telemedien. https://www.kjm-online.de/fileadmin/user_upload/Rechtsgrundlagen/Gesetze_Staatsvertraege/JMStV_geaend._durch_19._RAEStV.pdf

[10] Moses Olafenwa and John Olafenwa. 2022. ImageAI. https://github.com/OlafenwaMoses/ImageAI

[11] Joseph Redmon and Ali Farhadi. [n.d.]. YOLOv3: An Incremental Improvement. https://doi.org/10.48550/arXiv.1804.02767

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. [n.d.]. ImageNet Large Scale Visual Recognition Challenge. https://doi.org/10.48550/arXiv.1409.0575

[13] Tom Scott. 2016. Will YouTube Ever Run Out Of Video IDs? https://www.youtube.com/watch?v=gocwRvLhDf8

[14] Ramya Srinivasan and Ajay Chander. 2021. Biases in AI systems. *Commun. ACM* 64, 8 (2021), 44–49. https://doi.org/10.1145/3464903

[15] Muhammad Uzair Tariq, Afsaneh Razi, Karla Badillo-Urquiola, and Pamela Wisniewski. 2019. A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. In *Human-Computer Interaction. Design Practice in Contemporary Societies*, Masaaki Kurosu (Ed.). Lecture Notes in Computer Science, Vol. 11568. Springer International Publishing, Cham, 90–108. https://doi.org/10.1007/978-3-030-22636-7{_}6

[16] yt dlp. 2022. YT-DLP. https://github.com/yt-dlp/yt-dlp