Acoustic Pornography Recognition Using Recurrent Neural Network

Abdulaziz Saleh Ba Wazir  
Faculty of Engineering  
Multimedia University  
Cyberjaya, Selangor, Malaysia  
zezu9512@gmail.com

Hezerul Abdul Karim  
Faculty of Engineering  
Multimedia University  
Cyberjaya, Selangor, Malaysia  
hezerul@mmu.edu.my

Mohd Haris Lye Abdullah  
Faculty of Engineering  
Multimedia University  
Cyberjaya, Selangor, Malaysia  
haris.lye@mmu.edu.my

Sarina Mansor  
Faculty of Engineering  
Multimedia University  
Cyberjaya, Selangor, Malaysia  
sarina.mansor@mmu.edu.my

Abstract—This research proposes a pornography recognition model using audio features utilizing recurrent neural network. The proposed work is totally different from most of the previous attempts on pornography recognition using visual contents of the nudity images or videos. The importance of using sounds for pornography recognition arises in the cases in which the visual features are not adequately informative of the nudity contents such as dark scenes, and scenes with the covered bodies. Acoustic recognition model selects the finest signal representation by feature extraction of Mel-Frequency Cepstrum Coefficients (MFCCs) after been pre-processed for noise reduction. Extracted features are fed into a network with long short-term memory (LSTM) cells. The LSTM cells have the capability to solve problems associated with temporal dependencies and require learning long-term and solve the vanishing gradient problems associated with Recurrent Neural Network (RNN). A dataset of porno-sound samples has been created based on an existing pornography video dataset. 800 data samples were used including positive and negative samples. The experimental results confirm the feasibility of the proposed acoustic-driven approach by demonstrating an accuracy of 86.83%, and F-score of 86.83%, in the task of pornography recognition.

Keywords— pornography recognition, RNN, MFCC features, Long-shot Term Memory, acoustic recognition.

I. INTRODUCTION

Audio has become an effective tool for shaping one’s personality and character and almost has similar effect as videos as well. The audio materials could be an independent audio files or an audio of an internet videos, television programs, and movies. The ease of accessibility to a huge audio files raised the necessity of filtering the audio content. With this in view, some local and foreign broadcasting companies should obtain suitability approval before distribution and public viewing.

Most of the broadcasting companies hire a lot of manpower to detect the sexual scenes prior to censorship. This process is a big use of manpower, time and cost. In addition to inaccurate detection of improper voices due to fatigue of manpower and weakness of human system in long time tasks.

There are several previous attempts that addressed the problem of pornography recognition [1]–[7] that utilized the visual contents to automatically detect and classify the nudity scenes from the normal ones. However, this study motivated to recognize porno-scenes based on the sounds that clearly represent sexual scenes.

The major purpose of this study is to facilitate the task of nudity scenes detection by proposing to utilize the power of acoustic features of MFCC as these features represent the porno-sounds that encompass the real content of the pornography scenes. The MFCCs is playing a good role in the drastic improvement on speech and acoustic recognition from the implementations of basic word recognition to automatic speech and different sounds recognition that uses continuous speech in the era of Deep Neural Networks, as Deep Learning utilizes artificial neurons to serve high dimensional data [8].

Deep network types differ in architecture that is associated with determining the accuracy and speed of the result based on different types of dataset. Various architectures have been developed to carry out certain applications such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) [9]. RNN works based on the principle of saving layers output to feed it back to the previous layers. This process helps to predict the layer output where the first layer in the network formed as feed forward networks with the product of weight’s sum and features. Therefore, each neuron will remember the information on its previous time steps. This feature of RNN implies that each of the neurons perform like a memory cell while performing computations [10].

Long short-term Memory (LSTM) is a type of RNNs which solves the issue that is generally common in RNN as it can not solve the type of problems that require the learning of long-term temporal dependencies. This is due to the exponential decay of the gradient of loss function that is called the vanishing gradient problem. LSTM is not a standard RNN where it uses special units in addition to the standard ones. LSTM units contains memory cells that can maintain information in memory for long period of time. LSTM has become the most common solving method for vanishing gradient problem and it has the ability to remember values for a short of long time period [11].

The paper is organized as follows. The next section (II) briefly overviews the recent works in the domain. It is followed by Section III which details the experimental setup and procedures followed in our research. In Section IV, results of the experiment are presented and discussed; this is followed by Section V which concludes the study.
A previous pornography recognition study has used audio features for the classification purpose. Two types of audio features have been used that are fused Pitch, and Mel-Frequency Cepstrum Coefficients (MFCC) in order to train five different variations of the k-nearest neighbor (KNN) supervised classification models based on the fusion of these features. The study has used porno-sound dataset created based from an existing pornography video dataset. A total of 169 audio dataset has been used containing 89 positive samples and 80 as negative samples. The proposed method has achieved an accuracy of 88.40%, and F-score of 85.20% in the task of pornography recognition [21].

III. METHODOLOGY

This section details the experimental procedure, the choice of the datasets, signal processing, feature extraction, performance metrics as well as the system specifications. Fig.1 presents the overall designed system. Firstly the video dataset was converted into soundtracks. Then, MFCCs were used to extract the features for both training dataset and testing dataset. The training dataset features is then fed into the RNN, LSTM model. Lastly, the testing dataset features are fed into the trained model for the output prediction.

A. Video and Sound Dataset

This study used an NDPI video dataset presented in [22] which includes 800 samples of videos. 800 soundtracks also were extracted using an open third-party software. Overall, the newly created soundtracks included 400 data points representing positive samples of porno-sounds, and 400 instances representing negative samples of normal voices, and normal action sounds. Sample image frames of the used video dataset are presented in Fig. 2 and Fig. 3.

![Fig. 2. Positive samples representation of the used video dataset](image1)

![Fig. 3. Negative samples representation of the used video dataset](image2)

![Fig. 1. The overall system design for acoustic pornography recognition](image3)
B. Samples Pre-processing

Data samples were required to be out of noise as possible, then trimmed into small audio files where each file consist of ten seconds prior to labelling using a free open source software named Audacity. Audacity is used in manually pre-processing the porno-sound samples as it provides the user with ability to select a window of the voice sample to process each small audio file separately. The extracted voice samples were with different length. Therefore, a sample of ten seconds of each sample has been taken and trimmed using a window of ten seconds on Audacity. The trimmed portion is then processed for noise production and labelling using Audacity features.

C. Feature Extraction

Mel-Frequency Cepstrum Coefficients (MFCCs) are features of acoustic and speech signals that can be used in the tasks of recognition and classification of voices [23]. The MFCC segments the input signal into various overlapped frames and then computes the cepstral features for each individual frame. More particularly, the algorithm outputs delta, the difference between the previous coefficient and the current coefficient over the defined frame length. The algorithm, as in [24], calculated the cepstral values by fitting the coefficients of the adjacent frames of an input signal. Prior to model development and training, feature extraction of the audio files is required. Firstly, the files format of the all separated files of 800 audio file is converted into wav format to execute the feature extraction algorithm. This project used MFCCs for feature extraction purpose. Each of the positives and negative sounds features are extracted with 20 MFCCs. Each of the 20 MFCCs are computed based on the following conditions as shown in Table I.

| TABLE I. MFCCs computation parameters |
|------------------|-----------------|
| Sampling Rate    | 44100 Hz, 16 bits |
| Window applied   | Hamming         |
| alpha            | 0.97            |
| Filter H(z)      | pre-emphasized  |
|                  | 1 - 0.95 z^-1   |

D. Training RNN Model

Recurrent Neural Network (RNN); Long Short-Term Memory (LSTM) acoustic pornography recognition model was developed using Python along with the use of Google’s TensorFlow library for functions. These functions are primarily essential for the implementation. The algorithm is written in python script, as the Tensorflow library is written for Python and periodically updated. The developed algorithm is then used to be trained for more than a trial to optimize the final model training parameters. The accuracy of the model and loss is checked through a function of TensorFlow named Tensorboard when tuning the parameters for optimization purpose. Different learning rate has been used in the model (0.0005, 0.0001, 0.002, 0.001, 0.01, and 0.1) to check which one is more efficient for the developed model.

The developed model uses loss function to calculate the error and optimizer to optimize the result by changing the weights in the network. The optimized parameters are then used to train the final model of the recognizer. The training dataset used is 600 data samples containing a mix of positive and negative instances.

E. Testing of RNN Model

The trained network is then applied to verification to test the developed model accuracy. Testing dataset consist of 200 data points. These voices are unlabeled to calculate the accuracy of the developed acoustic pornography recognition model as well as understanding the structure of the neural network. Having a good structure primarily affects the output. If the system is not organized properly the nodes may not be connected and therefore develop an inaccurate neural network.

F. Performance Metric

Four different evaluation metrics were used to assess the performance of the proposed recognition model (porn vs. non-porn) as Accuracy, Precision, Recall, and F-score formulated in (1), (2), (3), and (4).

\[
\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)
\]

\[
\text{Precision} = \frac{TP}{(TP+FP)} \quad (2)
\]

\[
\text{Recall} = \frac{TP}{(TP+FN)} \quad (3)
\]

\[
F\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)
\]

Where TP is the number of true positive samples (correctly classified as pornography), TN is the number of true negatives (correctly classified as normal), FP is the number of false positive instances (incorrectly classified as pornography), and FN is the number of false negatives (incorrectly classified as normal).

G. Software and Hardware

All the implementations were carried out using Python 3.7 The experiments were conducted on a desktop computer with Ubuntu Linux 18.04.2 (64-bit), 16 GB RAM, Intel Core i7-4770 CPU @ 3.40 GHz, and an NVIDIA GPU with 4 GB internal memory (GeForce GT 640).

IV. RESULTS

Learning rate size effects the accuracy and time of the developed model. Therefore, It was optimized with different values. The learning rates that is associated with time and accuracy convergence were tested for several rates. For 0.1 of learning rate, the accuracy of the network training algorithm per iterations is not varying and almost swinging around very low value of accuracy that is less than 20% because 0.1 learning rate is considered fast and does not allow training accuracy to increase to better value.

For 0.01 learning rate, the accuracy is changing a lot more. However, it reaches a constant value of accuracy around 36.5% after several iterations as 0.01 is relatively fast learning rate for acoustic recognition using LSTM. For 0.001 learning rate, the accuracy had risen much higher than all the previous rates as it reaches the 87.25% accuracy within 320 iterations which make this rate to be acceptable for training the model. A lower learning rate of 0.0001 the accuracy dropped around 50% difference than 0.001.

Therefore, 0.001 is a good learning rate for the developed network. Other rates within that range were tested to ensure
that 0.001 was the correct value to be used. Two values were tested 0.002 and 0.0005. For 0.002, accuracy graph shows a similar change to the 0.001 learning rate model. However, it does not reach the same accuracy of the same iteration with 0.001 as it is dropped around 10%. For the 0.0005 learning rate, the training accuracy was good however it reaches 86.75% accuracy at 410 iterations. Thus, 0.001 learning rate is ensured to be the most suitable for model training. Table II summarize the learning rate that has been tuned prior to final model training parameters decision.

Optimized learning and hyper parameters are obtained by the previous result trials. The most parameters found to be suitable is used to train the final model. The parameters used are in Table III.

<table>
<thead>
<tr>
<th>Learning rate</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>15.72</td>
</tr>
<tr>
<td>0.01</td>
<td>36.50</td>
</tr>
<tr>
<td>0.001</td>
<td>87.25</td>
</tr>
<tr>
<td>0.0001</td>
<td>37.00</td>
</tr>
<tr>
<td>0.002</td>
<td>78.25</td>
</tr>
<tr>
<td>0.0005</td>
<td>86.75</td>
</tr>
</tbody>
</table>

The parameters in Table III, are used to train the network. The input to the network is converted from MFCCs. The Adaptive Moment Estimation (Adam) optimizer allow the network to go for higher accuracy by adaptive moment gradient change by changing the weights in the network to achieve better accuracy, while cross-entropy loss function calculate the error in exponential decay way. Learning rate used is 0.001. It is modified after 640 iterations to 0.0001. The learning rate is slowed down to increase the chance of accuracy convergence. Consequently, accuracy kept at its optimal point, in addition to loss kept decreasing to its minimal point. The model achieved high training accuracy of almost 89%.

Table IV presents the pornography recognition confusion matrix of the of 200 datapoints of which 100 are positive samples. 89 samples were classified correctly as porn, while 11 are classified wrongly as non-porn.

<table>
<thead>
<tr>
<th></th>
<th>Porn</th>
<th>Non-Porn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Porn</td>
<td>89</td>
<td>11</td>
</tr>
<tr>
<td>Non-Porn</td>
<td>16</td>
<td>84</td>
</tr>
</tbody>
</table>

Table V presents the experimental results on the proposed acoustic pornography model and the newly constructed porno-sound dataset in this research; It presents the performance evaluation of the model with respect to the four-performance metrics.

Table V also presents a comparison between the proposed RNN model of this research and previous study on acoustic recognition using KNN model proposed in [21] in terms of evaluation metrics. The study compares two methods only as both studies used acoustic features concerning pornography recognition.

As it can be observed, high results achieved by the proposed RNN classifier in terms of all the performance metrics. The system achieved an average accuracy of 86.50% for 200 data points. F-score achieved while testing the model was 86.83% that comprises the percentages of precision, and recall that are 89.00%, and 84.76% respectively.

The type of dataset chosen has high effect on the good results achieved as the data used are almost comprehensive including clear positive and negative samples, as well as close to positive samples to facilitate the network to easily recognize the porno-sounds from normal ones. Additionally, some similarity of the features for each number extracted by MFCCs that cause confusion when testing the network that cause a slight loss of the model accuracy.

In comparison between the current and previous study as per Table IV, the KNN model uses a small number of dataset for training purpose, meanwhile RNN is a deep network that requires a very large number of data to achieve a high results.
for a comprehensive model that indicates that the proposed RNN model could be improved in the future with adding a number of dataset to train the model as well as network structure. The Deep RNN is kind of comprehensive model than KNN that is limited to small number of data samples as KNN with large number of data samples will produces lower accuracy and F-score than RNN model. Additionally, our proposed model has been trained using python where KNN model uses Matlab which produces a difference in the computation methods and time as well.

The proposed model was to use acoustic information extracted from video clips in order to train RNN classification model of LSTM specifically and test the feasibility of acoustic-driven features in the task of pornography recognition. More particularly, the features of MFCC were employed to construct acoustic representations of the audio tracks. The model has proven that pornography can be recognized with the use of acoustic features as it results a high accuracy that will facilitate the pornography detection models by enhancing the detection of visual and acoustic content of pornography side by side.

V. CONCLUSION

This research has successfully developed an acoustic pornography recognition solution using Recurrent Neural Network. Long Short-Term Memory (LSTM) has been chosen to carry out this experiment and utilize MFCCs to extract features of voice files that has been converted from existing video dataset. Although the time taken to train this detector is optimal, it takes quite a long period of time to train the network.

This research constructed the audio dataset comprising training and testing dataset. Model training was based on 600 data points out of total 800 data points of positive and negative samples that has been benchmarked with previous pornography recognition that utilized a bag of visual features. The learning parameters has been optimized for training accuracy optimization purpose. The developed model has achieved an accuracy of 89% during training phase and minimal loss.

Final purpose of this research is to test the model using a new test data set. 200 data points for all the positive and negative type of samples. The result seems promising with an 86.50% recognition accuracy as well as 86.83% as F-score comprising precision and recall percentages. These numbers show that this model could help to enhance the pornography recognition models that solely depends on the visual features as well as there could be enhancement on the design of the RNN layers to further improve the accuracy.

ACKNOWLEDGMENT

This research was fully funded by TM R&D, Malaysia project number [MMUE/180029].

REFERENCES


