A Baseline for NSFW Video Detection in E-Learning Environments

Pedro V. A. de Freitas
TeleMídia/PUC-Rio
pedropva@telemidia.puc-rio.br

Gabriel N. P. dos Santos
TeleMídia/PUC-Rio
gabrielpereira@telemidia.puc-rio.br

Antonio J. G. Busson
TeleMídia/PUC-Rio
busson@telemidia.puc-rio.br

Álan L. V. Guedes
TeleMídia/PUC-Rio
alan@telemidia.puc-rio.br

Sérgio Colcher
Informatics Departament/PUC-Rio
colcher@inf.puc-rio.br

ABSTRACT

The broad use of video capture and services for its storage and transmission has enabled the production of a massive volume of video data. This usage presents a challenge in controlling the type of content that is loaded for these video storage services. The Internet slang NSFW (Not Safe For Work) is often used as a warning for media that contain inappropriate content, such as nudity, intense sexuality, violence, gore or other potentially disturbing subject matter. Convolutional Neural Network (CNNs) architectures, or ConvNets, have become the primary method used for audio-visual pattern recognition. In this work, we intend to: (1) create a CNN based model for video features extraction; And (2), validate these video features with baselines models for NSFW video classification using a multi-modal approach. In initial experimentation, our best model achieves a recall of 96.6% for NSFW class.

CCS CONCEPTS

• **Information systems → Multimedia information systems**;
• **Computing methodologies → Machine learning approaches**.

KEYWORDS

Video Classification, NSFW, Dataset, CNN

1 INTRODUCTION

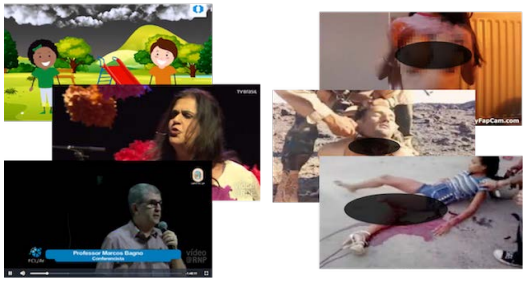
The broad use of video capture and services for its storage and transmission has enabled the production of a massive volume of video data. For example, in 2019, more than 500 hours of video are uploaded to YouTube every minute[1]. This usage presents a challenge in controlling the type of content that is loaded for these video storage services. For instance, we cite the that services like Facebook and Youtube are being sued for hosting videos from the Christchurch shootings[2].

[1]https://kinsta.com/blog/youtube-stats/
[2]https://www.bbc.com/news/technology-47705904

WebMedia '19, October 29-November 1, 2019, Rio de Janeiro , Brazil

ACM ISBN 978-1-4503-6763-9/19/10…$15.00
https://doi.org/10.1145/3323503.3360625

The Internet slang NSFW (Not Safe For Work) is often used as a warning for media that contain inappropriate content, such as nudity, intense sexuality, violence, gore or other potentially disturbing subject matter. On the other hand, SFW (Safe for Work) means that a content is suitable for most viewers. The Figure 1 illustrates these two categories. There are three scenes with appropriate (or SFW) content on the left, while three scenes with inappropriate (or NSFW) on the right.



(a) Safe educational videos **(b) Non safe videos.**

Figure 1: Examples of each category

Control the type of a content loaded to video storage service requires an automatic analysis in an efficient and quick way. Methods based in *Deep Learning* (DL) became the *state-of-the-art* in various segments related to automatic video analysis. Convolutional Neural Networks (CNNs) architectures, or ConvNets, have become the primary method used for audio-visual pattern recognition.

In our ongoing research, we intend to evaluate and develop methods of CNNs in order to detect NSFW videos In particular, we are interested in the video storage services that focus on host educational content. More precisely, we intend to address videos stored on video services from the Brazilian RNP (National Research Network), such as video@RNP[3](video repository) and RUTE[4] (Telemedicine University Network). Such services consist on video sharing networks that host videos from different Brazilian universities and have restrictions on inappropriate content. Uploading content that is inappropriate for these platforms by a malicious user can lead to legal issues.

Other papers also share our motivation, such as [12, 15, 17]. However, most of them don't use audio and image for classification,

[3]http://www.video.rnp.br
[4]http://www.rute.rnp.br

or use hand-crafted feature extraction methods, or don't use the latest feature extraction CNNs, which have been showing great potential in video recognition and classification. Our work uses two deep CNNS, one to extract image sequence features and other to extract audio features. We combine those features to create a single feature vector for the entire video, which then is used as input for the baseline classifiers. It is a rather simpler method for video classification and yet it still yields better results than the related work.

It is important to note that this short paper just presents our baseline, in order to validate our feature extraction method, the final model development is still ongoing. Because of that, we still don't have a benchmark to present and evaluate. To present our proposal, this paper is organized as follows. 2 discuss related work. We presents our used *dataset* in section 3. Then we discuss the used model to classify the NSFW video. Then, we present experiments and results using such model in section 5. Finally, the section 6 presents our final remarks and future work.

2 RELATED WORKS

Song and Kin [12] create a scheme for detecting pornography videos using multimodal features, those features being, image descriptor features of the frame sequence, extracted using the VGG-16 CNN[11], motion features extracted using optical flow[6] and the VGG-16, and audio features extracted using a Mel-scaled spectrogram. The final features for each model are obtained by an average pooling of each of the features by sample in the video. Each of those kind of features are used in a single SVM classifier per type of feature, resulting in an image sequence based detector, a motion based detector, and an audio based detector. The final decision making is done by model stacking all detectors. The authors used a modified dataset based on the 2k-pornography dataset[9] for training and testing. The results of their method are an average of 63.4% with 100% true positive rate for porn, and an average of 23.5% of false positive rate.

Our work also uses multimodal features, but we only use image sequence features and audio features, furthermore, we use an Inception-V3 CNN instead of the VGG-16 CNN for extracting image sequence features and use an Audio VGG CNN for extracting audio features. Although the authors achieve 100% recall rate for pornography, which is the main goal of their task, their model also has 23.5% false positive rate, which means that normal videos would be occasionally classified as NSFW. Our aim is to also have a true positive rate as high as theirs, but reducing the false positive rate.

Castro [15] show an implementation of a SFW/NSFW video classifier using a convolutional neural network from Open NSFW[7]. The CNN does a logistic regression on each frame, resulting in a value from 0 to 1 at each frame. The higher the value is, the higher is the likelihood of the frame being NSFW. The dataset used contained 90 SFW video segments and 89 NSFW video segments extracted from 11 movies. The final score for the video the the max value from all frames of the video. The experiment showed accuracy of 81%, f1-score and Matthew's correlation coefficient(MCC) for the NSFW class of 0.8047 and 0.6343, respectively.

Although the work also approaches the NSFW content detection in videos problem with CNN like ours, it does not make use of audio features. The method is also different, it performs the regression first, then it takes the max value from all frames of the video, while ours combines features from all frames of the video into a single vector of features(by averaging) and then performs classification on the resulting features.

Wehrmann *et al.* [17] classifies adult content trained on the NPDI video dataset, which consists on 802 videos, totalling 80 hours of videos, half of them with adult content. Those videos were processed by keyframes, varying between 1 and 320 frames per video. The selected keyframes of each video were chosen by a scene segmentation algorithm, resulting in 16727 images. Their architecture consists of a Convolutional Network and a LSTM (Long-Short Term Memory). Those models were chosen for feature extraction with CNN and sequence learning with LSTM, taking in consideration modifications on the images such as scaling and distorting. Using this approach they achieved a score of 95.3% accuracy using a ROC curve as evaluation criterion.

In our model, we also approached the video analysis using frame by frame processing, but we chose evenly spaced frames by their timestamps, not a segmentation algorithm. We also processed the extracted sound and image embeddings from each frame using a pre-trained ConvNet, instead of an untrained one. Resulting in an accuracy of 98% and 97.97% f1-score for NSFW class.

3 DATASET

Our *dataset* is structured into videos of appropriate (SFW) and inappropriate content (NSFW). It is divided into 55.400 SFW videos, 50.400 NSFW videos. For *appropriate content*, we selected educational videos from public repositories. We extracted 49.920 videos from Yotube8M[5]. Because of the size of the dataset (8 million videos) and because of the video classification challenges it holds. One of the objectives of the project is to evaluate the presence of inappropriate videos in educational video repositories. Therefore, in selecting your own videos within YouTube8M, we have chosen videos from "Jobs & Education" and "News" top-categories. We made this choice because these videos generally have the "talking head" format. We believe that this format is common in an educational video repository such as video@RNP.

Included in our dataset, there are 5.000 videos from video@RNP. For further comparison with other works, The NDPI [2] dataset were integrated in our dataset, it contains 400 SFW videos and 400 NSFW videos. Those SFW videos being 200 labeled "hard" and 200 labeled "easy" to represent likelyhood of misclassification, example of "hard" videos are videos with high amounts of skin exposed, such in swimming and sumo fighting.

We also included the Cholec80 [16] in our dataset, it contains 80 videos of cholecystectomy surgeries performed by 13 surgeons. All videos from the Cholec80 dataset were labeled as SFW, since videos of surgery are usual in educational context.

For *inappropriate content*, we selected porn and violence (hereafter referred to as *gore*) from specialized websites. For the porn type, we extracted 47.758 videos from XVideos[6]. Because of the dataset size (7 million videos) and because of the amount and variety of annotations(tags and main tags). Each video has one main tag,

[5]https://research.google.com/youtube8m
[6]https://info.xvideos.com/db

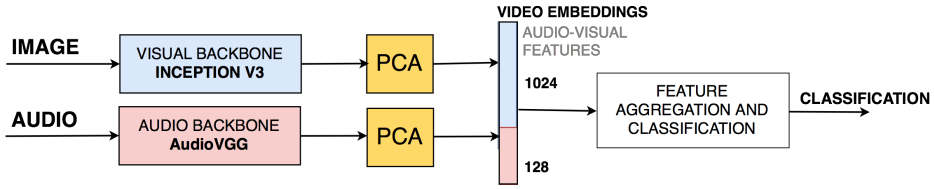


Figure 2: Bimodal architecture for NSFW video classification.

totalling 70 tags. To select the videos in this dataset, we distributed videos in these tags to maintain a proportion equal to the original XVideos dataset. In particular, to prevent lower-quantity tags from disappearing, we have defined a minimum of 10 porn videos for each tag. For the gore content we used a web crawler to extract 2.242 gore videos from various websites dedicated to gore media, such as Gore[7], BestGore[8] and GoreBrasil[9].

4 SFW/NSFW CLASSIFIER MODEL

Our CNN-based SFW/NSFW classifier is composed by two modules. The first module is what the researchers call the *backbone*, that acts as the feature extractor from which the model draws its discriminating power. The second module, the *classifier*, operates over the extracted features by the backbone to aggregate and classify it. The architecture of our NSFW classifier is illustrated in Fig. 2. We opt to a bi-modal approach that uses two backbones to extract the audio and image features form videos. Once we have extracted the features from the video, we then use a shallow model to perform the video classification. In the remainder of this section, we detail the embeddings extractor and the algorithms used for classification.

4.1 Video Embeddings Extractor

CNNs when trained tend to learn at the first layers the low level features (e.g. in visual domain: edges, corner, contours). At the intermediate and final layers, the combination of these filters helps to extract more complexes features, resulting in a vector of continuous numbers called *embeddings*. In this work, we use two bachmark CNNs to extract both image and audio *embeddings* from videos by using transfer learning technique [14].

Based in the work of Abu-El-Haija *et al.* [1], we decode each video at 1 frame-per-second up to the first 360 seconds and feed an InceptionV3 [13] with the network weights pre-trained in the ImageNet[10] to extract the image *embeddings*. We also feed the AudioVGG [5] with the network weights pre-trained in the Audioset[11] to extract the audio *embeddings*. Next, we apply PCA (+whitening) to reduce the dimensions of the image *embeddings* to 1024 and audio *embeddings* to 128. Finally, we concatenate both image and audio embeddings to compose the final video embeddings with 1152 dimensions.

4.2 Classifiers

(1) **Support Vector Machine (SVM)** [3] in which the data is mapped into a higher dimension input space where an optimal separating hyper-plane is constructed. These decisions surfaces are found by solving a linearly constrained quadratic programming problem.
(2) **K-Nearest Neighbors (KNN)** [10] uses distance measure between training samples so that the k-nearest neighbors always belong to the same class, while samples from different classes are separated by a large margin.
(3) **Multilayer-perceptron (MLP)** [4] contains layers of nodes: input layer, output layer and various hidden layers in between. The number of layers used is problem dependent, as is the number of nodes in each hidden layer. The weights are adjusted by local optimization using a set of feature vectors so that the network produces the optimal expected output.

5 VIDEO EMBEDDINGS VALIDATION

In this experiment, our objective is to attest the quality of our video *embeddings*. We evaluate the performances of popular baseline classifiers over the video *embeddings* that were extracted from a subset of the original dataset described in Section 3. In next subsections we discuss the experiment setup, used metrics and our empirical findings.

5.1 Setup

We selected a subset of 3300 videos from our original dataset. Then we split it into approximately 80%, 10% and 10% for the training, validation and testing sets, respectively. We opted to select an "easy" subset to check if our video *embeddings* are good. That is why we selected only YouTube videos for the SFW class and porn videos for the NSFW class. Then, we balanced these three sets with 50% of each class. Each set was structured as a collection of batches and each batch has 100 samples.

For the classifiers, we use the following hyper-parameters.

(1) **SVM** hyper-parameters: C: 1.0, decision function shape: 'ovr', degree: 3, gamma: 'scale', kernel: 'rbf', max iterations: -1, random state: None, shrinking: True and tolerance: 0.001.
(2) **KNN** hyper-parameters: leaf size 30, used the minkowski metric, k = 5 , p = 2 and uniform weights.
(3) **MLP** hyper-parameters: two hidden layers, the first one with 2000 neurons, and the second one with 3200, RelU activation, xavier initialization, adam optimizer, 0.001 learning rate, and softmax cross entropy loss function.

5.2 Metrics

We evaluate the models by the Precision (P), Recall (R) and F1-Score for SFW and NSFW classes:

[7]https://www.gore.com.br
[8]https://www.bestgore.com/
[9]https://www.gorebrasil.com
[10]http://www.image-net.org/
[11]https://research.google.com/audioset/

$$P = \frac{TP}{TP + FP} \quad (1) \qquad R = \frac{TP}{TP + FN} \quad (2) \qquad F1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

Where TP, TN, FP and FN denotes the examples that are true positives, true negatives, false positives and false negatives, respectively. The F1 score, defined in Equation 3, measures how precise the classifier by the harmonic mean between Precision (Equation 1) and Recall (Equation 2). The F1-score represents an overall performance metric, and the precision and recall metrics can give insights on where the classification model is doing better.

5.3 Results

In the Table 1, for the SVM model, we can observe that the recall metric value for the SFW class is 100%, that means that it correctly classified all SFW videos. The 100% value in the precision metric for the NSFW class means that it had no false negatives. The support columns represent the number of examples(videos) of each class for the validation.

	F1-score	Precision	Recall	Support
SFW	97.40%	94.93%	100.0%	150
NSFW	97.26%	100.0%	94.66%	150

Table 1: The evaluation metrics for the SVM model.

In the Table 2, representing the evaluation metrics for the KNN model, the recall value for the SFW class is higher than the one from NSFW class, meaning that, like the SVM model, it misclassified less SFW instances. The f1-score metric shows that the KNN model has overall worse performance than the SVM model in this task.

	F1-score	Precision	Recall	Support
SFW	94.53%	91.30%	98.00%	150
NSFW	94.11%	97.84%	90.66%	150

Table 2: The evaluation metrics for the KNN model.

Finally, in the Table 3, the precision and recall metrics are have closer values, meaning that the misclassified instances were better distributed between the classes. Still, the most valuable metric in this task is the recall for the NSFW class. The MLP model shows the biggest recall value for NSFW class of all models. The MLP model also showed the smaller error rate of all models, misclassifying only 6 out of 300 validation video instances. The Figure 3 shows the confusion matrix of the three models used in the experiment.

	F1-score	Precision	Recall	Support
SFW	98.02%	96.75%	99.33%	150
NSFW	97.97%	99.31%	96.66%	150

Table 3: The evaluation metrics for the MLP model.

The main drawback observed from using the MLP model, the best performing of our baselines, was its recall rate. Because, in the context of improper content detection, usually the recall metric for NSFW content is more important than the recall metric for SFW content.

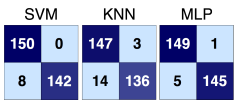


Figure 3: The confusion matrix for the SVM, KNN and MLP.

6 FINAL REMARKS

In this work we presented a comparison between common classification baseline models and established a baseline model for detection of NSFW content in E-Learning environments, our results show that the feature extraction using deep learning yield high accuracy even in linear or shallow models. This highlights that the feature extraction method is a solid base from which to build deeper classification models. The best performing baseline model is the MLP, with 98% of accuracy, 98.02% of f1-score for SFW and 97.97% of f1-score for NSFW classes.

The preliminary results shows that our experiment is easy because it had only YouTube videos and pornographic ones. Our goal at this point in the project was validate our video *embeddings* with common baselines. In the future we plan to redo the experiment by inserting more difficult videos, such as surgery and gore. In addition, we also plan to implement and try state-of-the-art models for video classification such as NetVlad and Context Gating [8].

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.
[2] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo De A AraúJo. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5):453–465, 2013.
[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
[4] Simon S Haykin et al. *Neural networks and learning machines/Simon Haykin*. New York: Prentice Hall„ 2009.
[5] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 131–135. IEEE, 2017.
[6] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.
[7] Jay Mahadeokar and Gerry Pesavento. Open sourcing a deep learning solution for detecting nsfw images. *Retrieved August*, 24:2018, 2016.
[8] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.
[9] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Pornography classification: The hidden clues in video space–time. *Forensic science international*, 268:46–61, 2016.
[10] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.
[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[12] Kwang Ho Song and Yoo-Sung Kim. Pornographic video detection scheme using multimodal features. *J. of Eng. and Applied Sciences*, 13(5):1174–1182, 2018.
[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[14] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, pages 270–279. Springer, 2018.
[15] Manuel Torres Castro. Automatic flagging of offensive video content using deep learning. Master's thesis, Universitat Politècnica de Catalunya, 2018.
[16] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. on medical imaging*, 36(1):86–97, 2016.
[17] Jônatas Wehrmann, Gabriel S Simões, Rodrigo C Barros, and Victor F Cavalcante. Adult content detection in videos with convolutional and recurrent neural networks. *Neurocomputing*, 272:432–438, 2018.