Neurocomputing 272 (2018) 432–438

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Brief papers

Adult content detection in videos with convolutional and recurrent neural networks

Jônatas Wehrmann[a], Gabriel S. Simões[a], Rodrigo C. Barros[a,*], Victor F. Cavalcante[b]

[a] Pontifícia Universidade Católica do Rio Grande do Sul, Av. Ipiranga, 6681, Porto Alegre 90619-900, RS, Brazil
[b] Motorola Mobility, R&D-Brazil, Rodovia SP 340 Km 128.7, Jaguariuna 13820-000, SP, Brazil

ARTICLE INFO

Article history:
Received 6 September 2016
Revised 3 March 2017
Accepted 6 July 2017
Available online 13 July 2017

Communicated by Dr Xiang Xiang Bai

Keywords:
Adult content detection
Deep learning
Convolutional neural networks
Recurrent neural networks

ABSTRACT

The amount of adult content on the Internet grows daily. Much of the pornographic content is unconstrained and freely-available for all users, requiring parents to make use of parental control strategies for protecting their children. Current parental control devices depend on human intervention, and hence there is the need of computational approaches for automatically detecting and blocking pornographic content. Toward that goal, this paper proposes *ACORDE*, a novel deep learning architecture that comprises both convolutional neural networks and LSTM recurrent networks for adult content detection in videos. Experiments over the freely-available NPDI dataset show that *ACORDE* significantly outperforms the previous state-of-the-art approaches for this task, decreasing by half the number of false positives and by a third the number of false negatives.

1. Introduction

The automatic detection of adult (pornographic) content in images and videos is an important and challenging task, especially due to the huge amount of freely-available adult content on the web, whose spread has significantly increased with the massive adoption of mobile devices across the globe. A recent report[1] indicates that the Internet traffic to porn websites accounted for 8.5% of the total in the UK in June 2013, surpassing the traffic for shopping, news, business, and social networks.

Even though organizations such as MPAA[2] have developed rating systems to protect viewers from adult scenes in motion pictures, content available on the web is practically unconstrained and easy-to-access, motivating the development of computational approaches that are capable of automatically detecting pornography with the final goal of protecting sensitive populations (e.g., children under 18). The task of automatically identifying adult content, however, poses a greater challenge than other classification problems due to the degree of subjectivity and uncertainty surrounding the problem. For instance, it is hard even for human beings to properly assess degrees of sensuality in scenes where people wear swimsuits or underwear. Indeed, sometimes more than one image/frame is needed for contextualizing the scene in order to define whether it should be classified as adult content or not.

Earlier work on pornography identification focused on human skin detection [1–4], in which the idea is that greater amounts of detected skin would lead to higher probabilities of nudity within the image or video, hence characterizing the content as pornographic. Nevertheless, these approaches suffer with a high rate of false positives, especially in the context of beaches or practice of aquatic sports. More recent studies [5–8] approached the problem under the perspective of *Bag of Visual Words* (BoW) and similar models (e.g., BossaNova [8,9]) for aggregating (quantizing) sophisticated image descriptors.

For benchmarking the proposed approaches in the area in terms of both video and image detection, researchers have used the NPDI dataset [8]. The best results achieved in NPDI are described by [10], where the authors propose a video descriptor based on binary features (*BinBoost* [11]) which is used with the BoW/BossaNova representations. However, the very same approach reaches only 44.6% of mean average precision (mAP) in the well-known PASCAL VOC dataset [12], while recent deep learning approaches reach about 60% of mAP in that same dataset [13]. This is a clear indication that deep learning based approaches could be a good option for pornography detection in both images and videos.

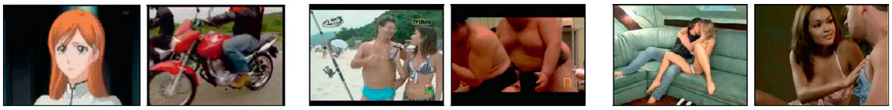
Therefore, in this paper we propose a novel approach for adult content detection in videos, namely *ACORDE* (Adult Content Recognition with Deep Neural Networks). Its architecture makes use of a convolutional neural network (ConvNet) as a feature extrac-

* Corresponding author.
E-mail addresses: jonatas.wehrmann@acad.pucrs.br (J. Wehrmann), gabriel.simoes.001@acad.pucrs.br (G.S. Simões), rodrigo.barros@pucrs.br (R.C. Barros), victorfc@motorola.com (V.F. Cavalcante).

[1] http://goo.gl/nG0s7n.

[2] http://www.mpaa.org/.

http://dx.doi.org/10.1016/j.neucom.2017.07.012
0925-2312/© 2017 Elsevier B.V. All rights reserved.



(a) Easy non-adult class. (b) Hard non-adult class. (c) Adult class.

Fig. 1. Frames from the NPDI dataset.

tor and of a long short-term memory (LSTM) to perform the final video classification. *ACORDE* extracts feature vectors from the video *keyframes* of NPDI, building a sorted set of semantic descriptors. This set is used to feed the LSTM that is responsible for analyzing the video in an end-to-end fashion. The proposed approach does not require fine-tuning nor re-training the ConvNet. Results show that *ACORDE* comfortably establishes the new state-of-the-art for adult content detection in NPDI, reducing by half the number of false positives and by a third the number of false negatives.

This paper is organized as follows. Section 2 briefly introduces the NPDI dataset as well as recent methods for pornographic classification of videos. Section 3 describes our proposed approach in detail. Section 4 presents how the experimental setup was organized for performing the empirical analysis, which is presented in Section 5. Finally, in Section 6 we detail our conclusions and future work directions.

2. Background

This section discusses earlier work that performs adult content detection, and also describes the NPDI dataset, which will be used to validate our novel approach.

2.1. NPDI dataset

Currently, the largest publicly-available pornographic dataset is NPDI [9], which comprises nearly 80 h from 802 videos (half of them with adult content), all downloaded from the Internet. The non-adult class is further sub-divided in 201 easy-to-classify videos and 200 hard-to-classify videos. The latter were selected based on textual search queries like *beach*, *wrestling*, and *swimming*, in order to verify the ability of the proposed classifiers in scenarios of high skin-exposure. The adult class comprises 401 videos selected from adult content web sites. Fig. 1 shows a sample of frames from the easy non-adult, hard non-adult, and adult classes.

As described in [9], a scene segmentation algorithm was employed to extract keyframes from the videos, resulting in a total of 16,727 images. Each video may contain 1–320 keyframes. The average amount of keyframes per class are: 15.6 for adult videos; 33.8 for easy non-adult videos; and 17.5 for hard non-adult videos. NPDI has a wide ethnic diversity with asian, black, white, and multi-ethnic videos. Issues like one-keyframe videos and *anime*-style content considerably increase the challenge of NPDI.

2.2. Related work

In the work of [9] and [10], the authors make use of both low and mid-level visual features extracted from the NPDI dataset. They use such features to build a final movie representation. The method is based in a low-complexity alternative for feature extraction using binary descriptors and a combination of mid-level representations. They aggregate the descriptors via the BoW model [14], generating the *BoW video descriptor* (BoW-VD). Also, they use the BossaNova method [15], which is an improved extension of the BoW model, generating the *BossaNova video descriptor* (BNVD). BNVD is a video descriptor that represents the median distance for each visual word of a given *codeword*[3] for a *codebook*.[4]

The work of [16] is the first to use deep neural networks to address the pornography detection problem. That work proposes a method that requires fine-tuning two distinct ConvNets, namely *AlexNet* [17] and *GoogLeNet* [18]. The author performs the training phase by reusing models pre-trained over the ImageNet dataset [19] and fine-tunes them over NPDI. That approach requires the training of ten distinct models: one model per training fold (5) and per network (2). Keyframes were rescaled to 256 × 256 to allow the data augmentation process with crops of the size 224 × 224 randomly sampled from each image in order to avoid overfitting. To normalize the data, the author subtracted the mean image from all instances.

Unfortunately, several methodological aspects are not clearly detailed in the paper, such as: (i) the stopping criteria adopted, (ii) the usage of a validation set, (iii) values of important hyper-parameters like learning rate, momentum, and regularization; and (iv) the updated layers in each model. Note that the absence of a proper validation set may compromise the reliability of the results. For the test phase, each network predicts *benign* (non-adult) and *adult* probabilities for each keyframe. The probabilities from both models are averaged, and a video is classified as adult (benign) when most of its keyframes are predicted as belonging to the *adult* (*benign*) class.

3. ACORDE

In this paper we propose a novel method for adult content detection in images and videos, namely *ACORDE* (Adult Content Recognition with Deep Neural Networks). The architecture of *ACORDE* comprises a convolutional neural network (ConvNet) [20] for feature extraction and a long short-term memory network (LSTM) for sequence learning [21]. ConvNets are the current state-of-the-art for many computer vision tasks such as image classification [22], object detection [13], video analysis [23–26] and image segmentation [27]. LSTMs are well suited to learn representations of sequences such as videos and texts. The conjoint use of both algorithms has been used to solve problems in video analysis [28] and scene captioning [29].

A ConvNet is a deep learning strategy that combines three ideas to ensure some degree of shifting, scale, and distortion invariance regarding the image content: local receptive fields (filters), shared weights, and spatial (or temporal) pooling [30]. The convolution operator is applied in order to replace fully-connected matrix multiplications, granting the two first mentioned ideas and considerably reducing the amount of parameters within a network. Convolutional filters are learned using the well-known backpropagation algorithm [31]. This process can be seen as *representation learning*, i.e., the network acting as a feature extractor. Learning representations from images is vital for the success of the computer vision task at hand. Eq. (1) defines a convolution, where (x, y) is a position on the jth feature map from the ith network layer; m indexes

[3] A codeword is the centroid of a given visual words cluster.

[4] A codebook is a set of codewords.

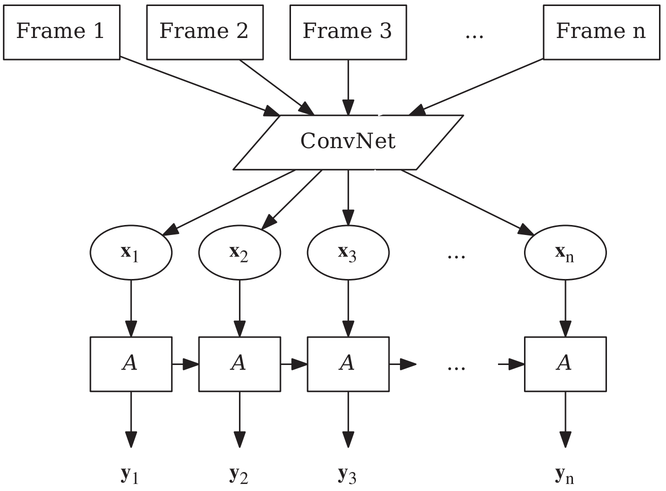


Fig. 2. *ACORDE* architecture.

the set of feature maps, b_{ij} is the corresponding bias value, w_{ijm}^{pq} is the weight's value at position (p, q), and P_i and Q_i are the height and width of the filter, respectively. The ReLU (rectifier linear unit) activation function [17] is often used as a source of non-linearity, essentially thresholding values in zero: $\text{relu}(v) = \max(0, v)$.

$$v_{ij}^{xy} = \text{relu}\left(b_{ij} + \sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)} \right) \tag{1}$$

Recurrent neural networks (RNNs) [32] are especially designed to learn sequential or time-varying patterns. LSTMs [21], for instance, are a specific type of RNN that uses a basic unit called *memory cell* to allow the constant error flow through time. A memory cell comprises gates that are designed to protect the cell from irrelevant inputs and from irrelevant content within the cell.

As in [33], the LSTM implemented within *ACORDE* is defined by the following components: block input (Eq. (2)), input gate (Eq. (3)), forget gate (Eq. (4)), cell state (Eq. (5)), output gate (Eq. (6)) and block output (Eq. (7)), where $\mathbf{x}^t$ is the input vector at time t, $\mathbf{W}$ are the weight matrices connected to the input, $\mathbf{R}$ are the recurrent weight matrices, $\mathbf{p}$ are peephole weight vectors, and $\mathbf{b}$ are bias vectors. Non-linear functions are denoted by σ, g and h. The sigmoid function is used in the gates, whereas the hyperbolic tangent is used in the block input and output. For short, the LSTM internal state is denoted by A.

$$\mathbf{z}_t = g(\mathbf{W}_z \mathbf{x}_t + \mathbf{R}_z \mathbf{y}_{t-1} + \mathbf{b}_z) \tag{2}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{R}_i \mathbf{y}_{t-1} + \mathbf{p}_i \cdot \mathbf{c}_{t-1} + \mathbf{b}_i) \tag{3}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{R}_f \mathbf{y}_{t-1} + \mathbf{p}_f \cdot \mathbf{c}_{t-1} + \mathbf{b}_f) \tag{4}$$

$$\mathbf{c}_t = \mathbf{i}_t \cdot \mathbf{z}_t + \mathbf{f}_t \cdot \mathbf{c}_{t-1} \tag{5}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{R}_o \mathbf{y}_{t-1} + \mathbf{p}_o \cdot \mathbf{c}_t + \mathbf{b}_o) \tag{6}$$

$$\mathbf{y}_t = \mathbf{o}_t \cdot h(\mathbf{c}_t) \tag{7}$$

Fig. 2 depicts *ACORDE*'s architecture. Let $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n\} \in \mathcal{F}$ be a keyframe sequence, $\mathbf{x}_t$ be the set of features extracted from the tth frame by the ConvNet, and $\mathbf{y}_t$ the predicted probability of existing adult content at the tth temporal iteration. The final prediction $y_{|\mathcal{F}|}$ given by *ACORDE* uses as features the LSTM's internal state A from its final iteration. State A has recurrent changes given new inputs and the outputs of the internal gates. Note that in Fig. 2 the recurrence is unrolled. Thus, when $t = 3$, there is no $\{\mathbf{x}_t, \mathbf{y}_t\} \ \forall \ (t > 3)$.

Table 1
Amount of parameters in well-known ConvNets.

Network	#Parameters
AlexNet [17]	≈ 66M
GoogleNet [18]	≈ 6M
ResNet-50 [22]	≈ 25M
ResNet-101 [22]	≈ 44M
ResNet-152 [22]	≈ 60M

Since *ACORDE* makes use of a RNN to analyze each frame, we can obtain probability y_i iteratively. This is useful for detecting and blocking videos in real-time applications. The work of [16] has difficulties in classifying video streams because of its majority voting strategy, which harms predictions performed in real-time. For instance, a video containing only a few pornographic scenes will most probably be misclassified. On the other hand, methods based on histogram representations (e.g., [10]) depend on the training of the final classifier (often a non-linear SVM) to perform accurate intermediate predictions. Training a classifier based on the entire videos' histograms introduces a bias, in which the induced model may expect a large amount of explicit content at test time. Such an assumption is probably false, especially for the initial stages of the adult video.

3.1. Image feature extraction

Training deep models in large datasets such as ImageNet [19] generates discriminative models capable of encoding semantic information from the images. *ACORDE* makes use of pre-trained ConvNets to extract representative feature vectors $\mathbf{x}_i$ from images/frames. We have experimented with both GoogleNet [18] and ResNet [22] architectures. GoogleNet is a compact ConvNet with the same prediction power of a VGG-19 [34], whereas ResNet holds the actual state-of-the-art for the ImageNet challenge. Common ResNet incarnations are composed by 52, 101, and 152 layers.

ACORDE extracts 1024 features from the last convolutional layer from GoogleNet and 2048 features from ResNets 52, 101, and 152. It normalizes the images by subtracting the RGB mean from the available pornographic dataset (in our experiments, NPDI). For generating more robust features, *ACORDE* employs 10 crops from each original frame: top-left, top-right, bottom-left, bottom-right, and center (and then the same crops but horizontally mirrored). The final features are defined by the average of the vectors extracted from the 10 crops.

Table 1 shows the amount of parameters in well-known ConvNet architectures. The model proposed by [16] is composed by both an AlexNet and a GoogleNet, totalizing ≈72M of parameters, much more than the networks explored in *ACORDE*. In addition, *ACORDE* requires only the forward pass of the architectures, since it does not train a ConvNet model over the available pornographic data.

To evaluate the discriminative power of the extracted features, we trained a linear SVM ($C = 1$) at image level for the NPDI data. We have labeled each keyframe based on the respective video class. Note, however, that not every frame in an adult video has adult content. Therefore, the results presented in Table 2 should be interpreted as the capability of each feature extractor to recognize frames from adult movies (and not of recognizing adult content per se). The results presented in Table 2 are the average accuracy obtained per fold. Note that the multiple crops improve the resulting accuracy in all cases. Results show that the high-level features extracted by the deep networks have similar discriminative power, but all of them outperform the well-known SIFT (BoW) descriptor by ≈20%. SIFT (BOF) was generated by using 100,000 randomly sampled points clustered in 128 keywords.

Table 2
Average accuracy (%) ± standard deviation from different feature extractors.

Feature extractor	1 crop	10 crops
GoogleNet	86.89 ± 1	89.34 ± 1
ResNet-50	89.07 ± 1	90.01 ± 1
ResNet-101	88.68 ± 1	90.27 ± 1
ResNet-152	87.91 ± 0	89.28 ± 1
SIFT (BOF)	67.68 ± 3	–

Table 3
Hyper-parameters setup.

Hyper-parameter	GoogleNet	ResNets	LSTM
Optimizer	SGD	SGD	Adam
Learning rate (α)	1×10^{-2}	1×10^{-1}	1×10^{-3}
Decay of α (γ)	4% every 8 epochs	10× when error plateaus	–
Momentum	0.9	0.9	–
Weight decay	2×10^{-4}	1×10^{-4}	1×10^{-4}
Weight initialization	[36]	[36]	[36]

3.2. Video-based learning

ACORDE's architecture is trainable in an end-to-end fashion and it accepts variable-length inputs. Whereas one does not need to retrain the ConvNet over the pornographic dataset, the LSTM must be trained in the available data (e.g., NPDI). The parameter updating process of the LSTM is performed by using the full gradient through time [35], which means the model can learn complex long-term relationships among frames.

Let $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n\}$ be a video composed by n RGB frames, and $\mathbf{x_t} = \phi(\mathcal{F}_\mathbf{t})$ the features extracted from frame t after the last convolutional layer's activation. The process performed by ϕ is time-independent and can be massively parallelized. The final prediction step is then represented by $y_n = \varphi(\mathbf{x_n}, \varphi(\mathbf{x_{n-1}}))$, where φ indicates the LSTM recurrence. Note that the forward pass of the nth frame is equivalent to a very deep neural network with n layers performing several nonlinearities.

The use of an LSTM network as a sequence-learner causes values of y_n to be generated by using information from all previous frames $\mathcal{F}_t$ $\forall t \in \{1, 2, \ldots, n-1\}$. The assumption here is that the LSTM hidden state A that has been used to compute y_n contains information collected from the video as a whole.

To summarize, the advantages of using an LSTM within *ACORDE* are twofold: (i) allows for the learning of fixed-sized representations of variable-length inputs in an end-to-end fashion; and (ii) allows for the mapping of long-term dependencies within frames.

4. Experimental methodology

In this section, we present the experimental methodology that we employed for evaluating *ACORDE*'s performance. We describe the baseline algorithms that are compared with *ACORDE* in Section 5.2, the evaluation measures that assess the quality of the classifiers in Section 4.2, and the hyper-parameters required for running *ACORDE* in Section 4.3.

4.1. Baseline algorithms

We compare *ACORDE* with five classification strategies that have been previously applied for adult video classification: (i) BossaNova-HueSIFT [8]; (ii) BossaNova-BRISK [9], (iii) BNVD [9]; (iv) BoW-VD [10]; and (v) AGbNet [16].

4.2. Evaluation measures

As discussed in Section 3, the outputs of *ACORDE* for each class are probability values, and the same is true for the baseline algorithms. Hence, the final predictions are often generated after thresholding these probability values in 50%. However, this choice is arbitrary and defining an optimal threshold is difficult and subjective. Hence, we avoid choosing thresholds by employing the *Receiver Operating Characteristic* (ROC) curve as the evaluation criterion for comparing the different approaches. For generating a ROC curve for a given classification method, one must select a predefined number of different thresholds within [0,1] to be applied over the outputs of each method. Finally, the true positive rate is plotted in function of the false positive rate for different cut-off points. The interpolation of these points generates a ROC-curve, and then we use the area under such a curve ($AU(ROC)$) as the indication of quantitative performance obtained by each method. In addition, we also show the values of accuracy (thresholding the probabilities in 50%) since the classes in NPDI are properly balanced.

4.3. Hyper-parameters settings

Table 3 shows the setup that has been used in the ImageNet-based ConvNet training for the GoogleNet and ResNet architectures, as well as the parameters for the training of the LSTM over the NPDI dataset. Note that these parameters are the default of the original papers and that no effort has been made in order to tune them.

Regarding the LSTM training, we use separated folds for validation and testing. *ACORDE* learns the model by using three out of the five folds. The training stops when the loss function plateaus for ten consecutive epochs. The best model is chosen by using the validation set as a proxy of the test set.

5. Experiments and discussion

In this section we present the experimental analysis that was performed in order to evaluate *ACORDE*. In Section 5.1, we evaluate the importance of coupling the LSTM within *ACORDE*'s architecture, whereas in Section 5.2 we compare *ACORDE* with the current state-of-the-art in the NPDI dataset. In all experiments, we show the performance of four distinct versions of *ACORDE: ACORDE-GN* (based on the GoogleNet architecture), *ACORDE-50* (based on the ResNet-50 architecture), *ACORDE-101* (based on the ResNet-101 architecture), and *ACORDE-152* (based on the ResNet-152 architecture).

5.1. Evaluating ACORDE's architecture

In order to show the gains provided by *ACORDE* when coupling the LSTM in the output of the convolutional neural network, we compare the performance of *ACORDE* within the NPDI dataset with convolutional-only approaches that employ either average pooling or max pooling over the features of the videos' keyframes. In addition, we also compare *ACORDE* with the use of only the convolutional network with per-frame predictions, in which the final prediction is simply given by majority voting of the per-frame predictions. Table 4 shows the results of such an analysis. Note that all *ACORDE*'s versions outperform their corresponding convolutional-only counterparts, clearly indicating that *ACORDE*'s architecture with the coupled LSTM is indeed beneficial for the video classification problem.

Table 4
Comparison between convolutional-only networks and *ACORDE* for video classification in NPDI.

CNN	Aggregation strategy	Accuracy (%)
GoogleNet CNN	Average pooling	88.4 ± 3
GoogleNet CNN	Max pooling	90.8 ± 1
ResNet-50 CNN	Average pooling	92.1 ± 3
ResNet-50 CNN	Max pooling	93.8 ± 3
ResNet-101 CNN	Average pooling	92.1 ± 3
ResNet-101 CNN	Max pooling	93.9 ± 1
ResNet-152 CNN	Average pooling	92.9 ± 2
ResNet-152 CNN	Max pooling	93.9 ± 2
GoogleNet CNN	Majority voting	90.5 ± 3
ResNet-50 CNN	Majority voting	92.8 ± 3
ResNet-101 CNN	Majority voting	92.8 ± 2
ResNet-152 CNN	Majority voting	91.6 ± 2
ACORDE-GN		92.8 ± 1
ACORDE-50		94.1 ± 1
ACORDE-101		94.0 ± 1
ACORDE-152		94.5 ± 1

Table 5
Results for video classification. (*) denotes the use of 10 crops during feature extraction. In bold the results that outperform the current state-of-the-art, and the best result achieved in the NPDI dataset are underlined.

Method	Accuracy (%)	*AU*(*ROC*)
BossaNova-HueSIFT [8]	89.5 ± 1	0.954
BossaNova-BRISK [9]	88.6 ± 2	0.960
BNVD [9]	92.0 ± 1	0.973
BoW-VD [10]	92.4 ± 2	0.976
AGbNet [16]	94.1 ± 2	–
ACORDE-GN	92.8 ± 1	**0.978**
ACORDE-50	**94.1 ± 1**	**0.986**
ACORDE-101	94.0 ± 1	**0.985**
ACORDE-152	**94.5 ± 1**	**0.986**
*ACORDE-GN**	93.3 ± 2	**0.981**
*ACORDE-50**	**94.8 ± 2**	**0.988**
*ACORDE-101**	**95.6 ± 1**	**0.990**
*ACORDE-152**	**95.3 ± 1**	**0.990**

5.2. State-of-the-art comparison

In this section, we show the results obtained by comparing the predictive performance of the following algorithms: BossaNova-HueSIFT, BossaNova-BRISK, BNVD, BoW-VD, AGbNet, and all versions of *ACORDE*.

Table 5 shows the results regarding both accuracy and *AU*(*ROC*). For simplicity, the table is split into three sections: (i) baselines results; (ii) *ACORDE*'s results with no cropping; and (iii) *ACORDE*'s results with cropping.

Regarding the baselines, the maximum accuracy that is reached is 94.1% ± 2 for AGbNet, which employs two ConvNets for classifying images and a majority voting scheme for classifying the videos. Both second and third placed methods are from [10], namely BNVD and BoW-VD, reaching ≈92% of accuracy. Note that even *ACORDE*'s weakest approach, namely *ACORDE-GN*, outperforms all non-network baselines. It is outperformed by AGbNet, but recall that *ACORDE-GN* contains ≈12× less parameters than AGbNet, as well as a more stable behavior. Overall, five out of the eight *ACORDE* variations outperform all baselines for all evaluation measures.

ACORDE -101 and *ACORDE-152* achieve the largest accuracy values, ≈95.5%. These results surpass the best baseline approach by ≈1.5% and the second-best by ≈3.2%. The *ACORDE* variations that employ residual connections achieve an *AU*(*ROC*) of around 0.99. Given the insufficient description of AGbNet's parameters, we could not reproduce it in order to generate its respective *AU*(*ROC*).

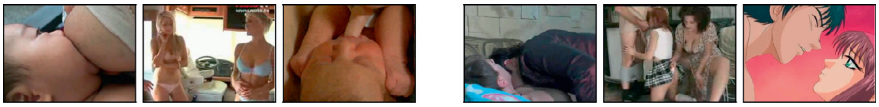
The use of the multiple-crop strategy (denoted by *) improves the predictive performance at video level for all *ACORDE*'s variations. The improvement ranges between 0.5% and 1.6%, with a greater impact for the deeper architectures. Shallower architectures such as *ACORDE*-GN and *ACORDE*-50 presented higher variability when using multiple crops.

Fig. 3 presents randomly selected keyframes from the videos misclassified by *ACORDE*-101. The first three keyframes (Fig. 3(a)) are examples of the non-adult class and the last three (Fig. 3(b)) of the adult class. Videos of breastfeeding babies usually present the exposure of women's breasts, which can be misleading to the classifier. *ACORDE* generates a probability of $y = 54\%$ of the first video being pornographic, leading to the incorrect prediction. The second video presents a game in which women in underwear have to perform certain moves and positions, though it is labeled as non-adult. *ACORDE* classifies that video as adult with $y = 83\%$ of probability. The last false positive is regarding a video with a single tricky keyframe. We highlight the fact that *ACORDE-101* perfectly classified all beach and sumo videos, which are often misclassified by the baseline approaches. We also noticed that misclassified videos with large probability scores are very rare.

The lack of exposure of intimate body parts is the most frequent characteristic in the false negatives. For instance, both fourth and fifth videos are 1-keyframe videos in which the intimate body parts are mostly hidden. We do believe that training a model with all movie frames could minimize those classification errors. The last keyframe shown in Fig. 3 is from a long *anime*-style video with very few explicit scenes. *Anime* content is much less frequent in the training data and greatly differs from the rest of the dataset, hence affecting the generalization performance of *ACORDE*.

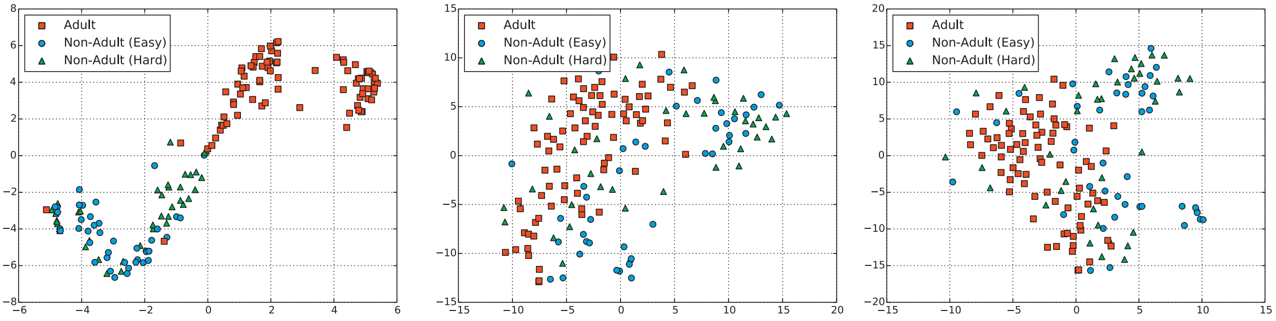
Fig. 4 shows a 2-dimensional *t-distributed stochastic neighbor* plotting (*t*-SNE) [37] for further analysis of *ACORDE* and the baseline methods. *t*-SNE is a dimensionality reduction technique for high-dimensional data visualization. Fig. 4(a) presents the plotting of the video-based features extracted from *ACORDE*'s LSTM. Fig. 4(b) and (c) presents, respectively, BNVD's and BoW-VD's generated video features. Note that *ACORDE* generates discriminative features that practically allows a linear separation of porn and non-porn videos. An interesting fact is also the possibility of separating the easy and hard non-adult subclasses, which are automatically recognized by *ACORDE* without explicitly training with these categories.

Figs. 5–7 show the confusion matrix for the best *ACORDE* version (*ACORDE*-101) and the baseline non-network methods. Note that *ACORDE*-101 decreases by half the number of false positives of BNVD and BOW-VD, clearly indicating that the high-level features from the ConvNet and the sequence learning of the LSTM are more robust in identifying adult content in scenarios with large skin-exposure. In addition, *ACORDE*-101 also decreases by a third



(a) Keyframes from non-adult videos. (b) Keyframes from adult videos.

Fig. 3. *ACORDE*-101 misclassified videos.



(a) *ACORDE*-101. (b) BNVD. (c) BOW-VD.

Fig. 4. *t*-SNE plots.

		Predicted Class	
		Adult	Non-Adult
Actual Class	Adult	381	20
	Non-Adult	16	385

Fig. 5. *ACORDE*-101 confusion matrix.

		Predicted Class	
		Adult	Non-Adult
Actual Class	Adult	370	31
	Non-Adult	33	368

Fig. 6. BNVD confusion matrix.

		Predicted Class	
		Adult	Non-Adult
Actual Class	Adult	373	28
	Non-Adult	33	368

Fig. 7. BoW-VD confusion matrix.

the number of false negatives of the baseline approaches (20 versus 33), proving to be safer to use in a parental control device than the previous state-of-the-art approaches.

6. Conclusions and future work

In this paper we proposed *ACORDE*, a novel deep neural network architecture that comprises both a ConvNet and an LSTM for adult video classification. To the best of our knowledge, *ACORDE* is the first method in the literature that makes use of such a type of architecture for detecting adult content in videos. We performed several experiments in the NPDI pornography dataset for verifying the best design choices for *ACORDE*. After a thorough empirical analysis, most of the *ACORDE*'s variations were capable of outperforming the current state-of-the-art methods for adult content detection.

As future work, we intend to make available a novel image-based adult dataset, which will be the largest dataset publicly-released to date. Moreover, we would like to verify which is the best transfer learning strategy for using in the pornography context. Finally, we want to develop a novel method for real-time segmentation of body parts in adult videos.

Acknowledgments

We would like to thank Motorola Mobility and the Brazilian research agencies CAPES and CNPq for funding this work. In addition, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPUs used for this research.

References

[1] M.M. Fleck, D.A. Forsyth, C. Bregler, Finding naked people, in: 4th European Conference on Computer Vision, 1996, pp. 593–602.

[2] M.J. Jones, J.M. Rehg, Statistical color models with application to skin detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1, 1999, p. 280.

[3] J.-S. Lee, Y.-M. Kuo, P.-C. Chung, E.-L. Chen, Naked image detection based on adaptive and extensible skin color model, Pattern Recognit. 40 (8) (2007) 2261–2270. Part Special Issue on Visual Information Processing.

[4] H. Zuo, W. Hu, O. Wu, Patch-based skin color detection and its application to pornography image filtering, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 1227–1228.

[5] T. Deselaers, L. Pimenidis, H. Ney, Bag-of-visual-words models for adult image classification and filtering, in: International Conference on Pattern Recognition, 2008, pp. 1–4.

[6] A.P.B. Lopes, S.E.F. de Avila, A.N.A. Peixoto, R.S. Oliveira, A.A. Araújo, A bag-of-features approach based on hue-sift descriptor for nude detection, in: European Signal Processing Conference, 2009.

[7] A.P.B. Lopes, S.E.F. de Avila, A.N.A. Peixoto, R.S. Oliveira, M.D.M. Coelho, A.D.A. Araújo, Nude detection in video using bag-of-visual-features, in: XXII Brazilian Symposium on Computer Graphics and Image Processing, 2009b, pp. 224–231.

[8] S. Avila, N. Thome, M. Cord, E. Valle, A.D.A. AraúJo, Pooling in image representation: the visual codeword point of view, Comput. Vision Image Understanding 117 (5) (2013) 453–465.

[9] C. Caetano, S. Avila, S. Guimaraes, A.D.A. Araújo, Pornography detection using BossaNova video descriptor, in: 2014 22nd European Signal Processing Conference (EUSIPCO), IEEE, 2014, pp. 1681–1685.

[10] C. Caetano, S. Avila, W.R. Schwartz, S.J.F. Guimarães, A.D.A. Araújo, A mid-level video representation based on binary descriptors: a case study for pornography detection, Neurocomputing (2016).

[11] T. Trzcinski, M. Christoudias, P. Fua, V. Lepetit, Boosting binary keypoint descriptors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2874–2881.

[12] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vision 111 (1) (2015) 98–136.

[13] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1.

[14] J. Yang, Y.-G. Jiang, A.G. Hauptmann, C.-W. Ngo, Evaluating bag-of-visual-words representations in scene classification, in: Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, ACM, 2007, pp. 197–206.

[15] S. Avila, N. Thome, M. Cord, E. Valle, A.D.A. Araújo, Bossa: extended bow formalism for image classification, in: 2011 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 2909–2912.

[16] M. Moustafa, Applying deep learning to classify pornographic images and videos, in: 7th Pacific-Rim Symposium on Image and Video Technology (PSIVT), 2015.

[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.

[20] Y. Le Cun, B. Boser, J.S. Denker, R.E. Howard, W. Habbard, L.D. Jackel, D. Henderson, Handwritten digitrecognition with a back-propagation network, Adv. Neural Inf. Process. Syst. 2 (1990) 396–404.

[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[23] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.

[24] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Adv. Neural Inf. Process. Syst. (2014) 568–576.

[25] J. Wehrmann, G. Simões, B. Rodrigo, T. Paula, D. Ruiz, (Deep) learning from frames, in: Proceedings of the Brazilian Conference on Intelligent System, 2016.

[26] G.S. Simões, J. Wehrmann, R.C. Barros, D.D. Ruiz, Movie genre classification with convolutional neural networks, in: International Joint Conference on Neural Networks (IJCNN 2016), 2016.

[27] G. Lin, C. Shen, I.D. Reid, A. van den Hengel, Efficient piecewise training of deep structured models for semantic segmentation, in: IEEE Computer Vision and Pattern Recognition (CVPR) 2016, 2016.

[28] A. Rohrbach, M. Rohrbach, B. Schiele, The long-short story of movie description, in: German Conference on Pattern Recognition, Springer, 2015, pp. 209–221.

[29] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[30] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016. In preparation

[31] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Technical Report, DTIC Document, 1985.

[32] L. Fausett (Ed.), Fundamentals of Neural Networks: Architectures, Algorithms, and Applications, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1994.

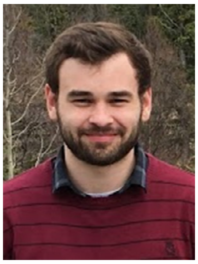
[33] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: a search space odyssey, IEEE Trans. Neural Netw. Learn. Syst. (99) (2016) 1–11. http://ieeexplore.ieee.org/document/7508408/

[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[35] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (5) (2005) 602–610.

[36] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034.

[37] L.V.D. Maaten, G. Hinton, Visualizingdata using t-SNE, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.



Jônatas Wehrmann is a Ph.D. student in computer science at Pontifícia Univeridade Católica do Rio Grande do Sul (PUCRS). He received a bachelor's degree in information systems from Sociedade Educacional Três de Maio (2015) and a master's degree with honors in computer science from PUCRS (2016). Recently he received the Google Research Awards for Latin America (2016) for his PhD project. He has published several papers in leading peer-reviewed conferences and journals, and received a best paper award at the Brazilian Conference on Artificial Intelligence (BRACIS 2016), the main national venue in AI. His current research interests are video and text understanding for multimodal retrieval, content summarization, and captioning via efficient Convolutional Neural Networks.

Gabriel Simões obtained his B.Sc. in computer science from Catholic University of Pelotas (UCPel, Pelotas, Brazil, 2004) and his M.Sc. in computer science from Federal University of Rio Grande do Sul (UFRGS, Porto Alegre, Brazil, 2007). Currently, he is a Ph.D. student in computer science at Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), working with machine learning, computer vision, and object detection.

Rodrigo Coelho Barros obtained his B.Sc. in computer science from Federal University of Pelotas (UFPel, Pelotas, Brazil, 2007), his M.Sc. in computer science from Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS, Porto Alegre, Brazil, 2009), and his Ph.D. in computer science and computational mathematics from University of São Paulo (ICMC-USP, São Carlos, Brazil, 2013). He was a visiting research assistant at the University of Kent, UK, from August 2012 to July 2013, where he worked with machine learning and data mining under the supervision of Dr. Alex A. Freitas. Dr. Barros has published several papers in peer-reviewed journals and conferences in the areas of machine learning, data mining, knowledge discovery in databases, and evolutionary computation, which are his main research interests. He received six best-paper awards from leading conferences, as well as an award for outstanding M.Sc. dissertation from PUCRS in 2009. He also received the Best Doctoral Thesis in Computer Science award from the Brazilian Computer Society in July 2014, and the Best Doctoral Thesis in Computer Science award from the Brazilian research agency CAPES, the latter being the most prestigious academic award for outstanding doctoral theses in Brazil. Dr. Barros joined the Faculty of Informatics at PUCRS in February 2014 as an associate professor to lead the machine learning research. Currently, he leads the Business Intelligence and Machine Learning Research Group (GPIN), working in machine learning, computer vision, and data mining topics.

Victor Cavalcante is a senior researcher in Analytics. Currently, he is the research manager responsible by the initiatives of Research and Innovation being conducted inside the Software Group at Motorola Mobility R&D in Brazil. Before joining Motorola Mobility, he was a research scientist at IBM Research, working within the Cognitive Computing strategy and member of the Social Data Analytics group and the IBM Technology Leadership Council in Brazil. He has previous professional experience as researcher, professor, optimization consulting and software engineering. Victor holds a Ph.D. and a Master in Computer Science from the State University of Campinas (UNICAMP), Brazil, with emphasis, respectively, on Mathematical Programming and Optimization heuristics. His main background is in Combinatorial Optimization and Operations Research, but his research interests also include other disciplines strongly related to Analytics like Machine Learning, Data/Graph Mining and Visual Aspects of data representation.