Multi-level detector for pornographic content using CNN models

1st Quang-Huy Nguyen
University of Information Technology
Vietnam National University
Ho Chi Minh City, Viet Nam
15520306@gm.uit.edu.vn

2nd Khac-Ngoc-Khoi Nguyen
University of Information Technology
Vietnam National University
Ho Chi Minh City, Viet Nam
15520386@gm.uit.edu.vn

3rd Hoang-Loc Tran
University of Information Technology
Vietnam National University
Ho Chi Minh City, Viet Nam
locth@uit.edu.vn

4th Thanh-Thien Nguyen
University of Information Technology
Vietnam National University
Ho Chi Minh City, Viet Nam
thiennt@uit.edu.vn

5th Dinh-Duy Phan
University of Information Technology
Vietnam National University
Ho Chi Minh City, Viet Nam
duypd@uit.edu.vn

6th Duc-Lung Vu
University of Information Technology
Vietnam National University
Ho Chi Minh City, Viet Nam
lungvd@uit.edu.vn

***Abstract*—This paper focuses on detecting and classifying pornographic content (images and videos) by using a multi-level CNN model with some supportive models. The main approaching method is to determine the images (keyframes extracted from videos) containing sensitives content or not by applying object detection model Mask R-CNN, which is the completely new approaching method in pornographic recognition. Moreover, the proposed model also adapts some other methods such as feature extraction and classifying based on CNN to increase the accuracy of the adaptive methods and ignore non-pornographic images and videos. Experimental results using the Pornography-800 and Pornography-2K datasets, performance of our method is reaching the accuracy of 92.13% and 90.40% respectively, show the effectiveness of the proposed method.**

***Index Terms*—image classification, object detection, convolutional neural networks, pornographic content recognition**

I. Introduction

With the advent of the internet, mobile technologies, social media, and peer-to-peer networks, obscene and pornographic contents (videos or images) has been broadly disseminated over the past decades and cause great harm to the social stability and adolescent psychology. Furthermore, it can also be the main reason for some types of crime, such as rape, sexual assault or even child abuse. Thus, the recognition of images and videos with pornographic content has become essential.

For the recognition of pornographic content on the internet, first idea to detect pornographic content is using the low-level features such as color and textures properties of human skin. These approaches tried to detect nudity content in the graphic and made decision on this result. However, this approach has a serious problem since it's hard to distinguish skin and skin-like objects and give high false positive in images which contain much exposed skin parts such as swimming, wrestling, etc. Another approach is inspired from Bag-of-Words model in text classification. These approaches, which is called Bag-of-Visual-Word (BoVW), tried to represent an image by an unsorted set of visual word. Using different descriptors such as SIFT or Hue-SIFT, BoVW creates a Codebook with the help of k-means algorithm from training dataset. Then, with the help of SVM classified, images will be classified into porn or non-porn category. These approaches still have some problems in defining task-specific visual Codebooks and mapping low-level features into a feature vector. At the same time, deep learning approaches with neural network have been becoming more popular significantly in the identification system to enhance performance and accuracy efficiently.

Therefore, this paper presents an approach to distinguish between normal content and pornographic content by improving the capacity to predict whether the images or videos are pornographic or not, even they are similar. We suppose that the pornographic image is one that contains sensitive objects such as breast, genitals, anus, etc. For that reason, the main approaching method of our work is applying convolutional neural network, one of the state-of-the-art deep learning approaches, to recognize sensitive objects on image and decide if this image is safe or not. As far as we know, this is the first time object detection is applied for pornography detection. In this paper, we proposed Mask R-CNN [2] to resolve sensitive object detection model with this implementation [3]. We also implemented some supportive methods to reduce resource cost and speed of our method, while improving our predictability.

The remainder of the paper is organized as follows. Section II presents the relevant literature. Next, in Section III, we describe our method in detail. Then, the experiments and results are reported in Section IV. Finally, in Section V, conclusion and future works are presented.

II. Related Works

A. Skin-based approaches

The first idea for detecting pornographic is finding if there are nude people in pictures or videos, which approaches tried to identify skin information of nude body. These approaches

will tried to classify each pixel as a skin pixel or non-skin pixel, which mostly based on color information and some low-level features. After that, a classifier will be trained to identify that photo or video contains nudity or not. In other words, a trained classifier will identify a graphic as pornographic or non-pornographic based on nudity-check with the help of skin information.

Ahmadia et al. [4] proposed the skin detection method used color based to discriminate skin and non-skin areas. Since color base feature are easy to obtain and also robust to the orientation and scaling. This method transform RGB color to color space YCbCr, eliminating Y axis and using just chrominance axes Cb and Cr. Then they calculate the ratio of skin region and others to obtain features and use as 1 criteria beside textual content to identify a graphic as pornographic or not. Zaidan et al. [5] has discussed models of skin detection and their advantages as well as disadvantages in real life. Through a review of 28 papers about skin detection method for classifying pornographic and non-pornographic images, they have pointed out that many encouraging results have been achieved in the area of pornographic detection. However, there are still many problems remain such as the quality and resolution of images may reduce the accuracy. Furthermore, skin-based approaches have a serious problem when input images contain many skin-like objects or some non-pornographic images contain much exposed skin parts like swimming, wrestling, etc. the false positive value will be a problem we have to deal with.

B. BoVW-based approaches

Inspired from Bag-of-Words model which turns text into vector and uses those vectors for calculations such as information retrieval or text classification, Bag-of-Visual-Word or BoVW approaches have been used recently to deal with image classification problems which pornography recognition is one of it. Generally, BoVW approaches extract key points with features in images by using different feature descriptor. From the training set, a visual Codebook may be learned by k-means algorithm. With a visual Codebook, the features of key points from any images may be converted into an intermediate representation using uniform feature vector. After all, an SVM classifier can be trained to classify images.

Many different descriptors can be used for extracting images' features. Lopes et al. [6] proposed an extension to the well-known SIFT descriptor - called Hue-SIFT descriptor which aimed at adding color information to the original SIFT. A comparison between SIFT and Hue-SIFT descriptor has been made on pornography recognition problems. The results showed that the combinations of color information and local feature information performed better. In [7], Avila et al. proposed BossaNova, which was an extension for BoVW approach, for representing content-based concept in images and videos and used this for pornographic video detection task. Key frames were extracted from segmented shots for representing videos. They used Hue-SIFT descriptor to extract features from key frames. Then, BossaNova - a mid-level representation based on a histogram of distances between the descriptors found in the image and those in the codebook - was used to encode local features. Finally, an SVM classifier for each key frame was employed and a major voting scheme for identifying video as pornographic or not. Moreira et al. [8] introduced a space-temporal interest point detector and descriptor called Temporal Robust Features (TRoF). TRoF was used to extract local features, then these features were aggregated into mid-level representation using Fisher Vector, the state-of-the-art model of BoVW.

The BoVW approaches improved the performance of pornography recognizing than skin-based approaches. However, these approaches still have some problems in defining task-specific visual Codebooks and mapping low-level features into a feature vector.

C. CNN-based approaches

In recent years, CNN-based approaches have shown empirically superior performance on computer vision problems, including pornography detection. These approaches did not use the hand-crafted visual features but let the machine learn the feature itself. Instead of training model from the beginning, most of approaches choose transfer learning to fine-tune a pre-trained CNN model.

Fudong Nian et al. [9] proposed a model utilizing a deep neural network (CNN) to detect pornographic images in a single model. The training data was obtained followed by an improved sliding window method and some data augmentations approaches. Two strategies for training algorithms were proposed. The first one is the pre-trained mid-level representations non-fixed fine-tuning strategy. Another one is adjusting the training data at the appropriate time on the

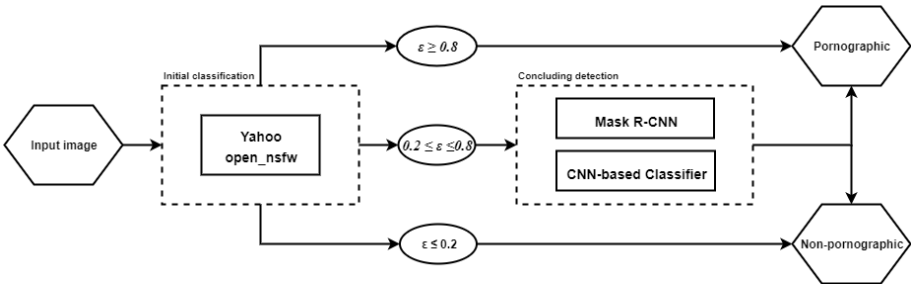

Fig. 1. Diagram of the proposed model

basis of the performance of the proposed network on the validation set. Mahadeokar et al. [10] from Yahoo proposed a deep learning solution for detecting not-safe-for-work (NSFW) images by released a train ResNet-50 model. This model also employed transfer learning to a fine-tune pre-trained model. The deep models were first pre-trained on the ImageNet 1000 class dataset. For each network, the last layer (FC1000) was replaced with a two-node fully-connected layer. Then the weights were fine-tuned on the NSFW dataset. After running, the results are NSFW marks for the input images. If the mark is high, it means that image is not safe (porn, violence, etc.), in contrast, the image is safe.

CNN-based approaches present the state-of-the-art performance on pornography detection problem. In this paper, we proposed a multi-level approach for pornography detection. This method combines two different CNN models, each model used in one level. The minor level uses Open-NSFW [10] to determine the safety of an image. Then in major level, we proposed the Mask R-CNN model and another CNN model to give the final result.

III. Proposed Model

A. Overview

Having ideas from this article [1], the pornographic images detection method presented in this paper is divided into two main phases: the initial classification and the concluding detection. The initial classification is adapted from Yahoo's Open-NSFW to extract features and exclude normal images and high-probability sensitive images. The rest images will be recognized in the concluding detection by applying two different methods, a pre-trained Mask R-CNN model and a classification CNN-based model. The approach of proposed model presented in this paper is shown in Figure 1.

B. Initial classification

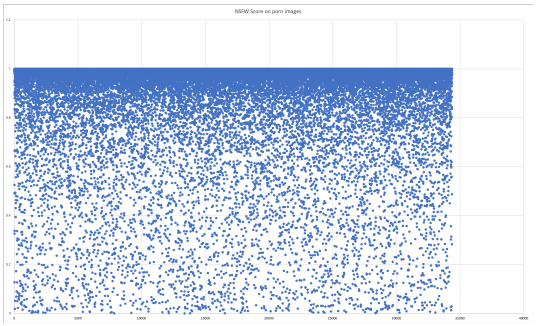

Fig. 2. Distribution of Open-NSFW score for pornographic images

In this phase, we use Yahoo's Open-NSFW model [10] to extract features and exclude normal or obvious pornographic images. It can not only improve integral recognition efficiency but also contribute to the improvement of recognition speed. The model yield a number ε between 0 and 1 corresponding to the pornography probability of the image. From the experiments (also follow the recommended usage of the model), we conclude that if the ε is lower than 0.2, the image is considered to be a normal image, and pornographic when ε is higher than 0.8. If the ε between 0.2 and 0.8, then we will take the image to the conclude detection for further identification. Figure 2 and Figure 3 show our result from the experiment to determine the threshold ε. The pornographic image set and the non-pornographic image set that we used for the experiment have more than 35000 and 8000 images respectively, both of them were extracted from Pornography-800 and Pornography-2k datasets.

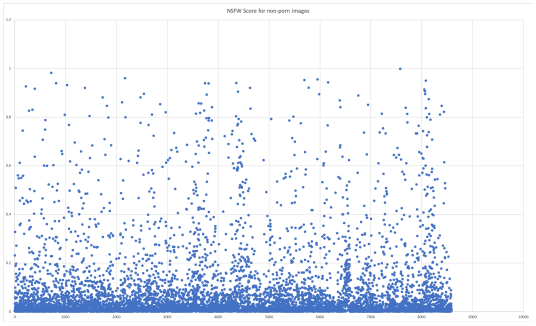

Fig. 3. Distribution of Open-NSFW score for non-pornographic images

C. Concluding detection

In the concluding detection, we adapted two different models, Mask R-CNN with ResNet101 backbone and a CNN-based classifier with ResNet50 backbone.

1) Mask R-CNN: In the first model, we modified a Mask R-CNN model [3] from Keras and TensorFlow to detect each instance object on the images. The Mask R-CNN model generates bounding boxes and segmentation masks to identify sensitive objects on each image and determined whether an image contains sensitive content or not.

This model adapts Faster R-CNN network based on ResNet101 + FPN (Feature Pyramid Network) backbone for feature extraction. Subsequently, a lightweight neural network termed Region Proposal Network with ROI classifier is applied to generate bounding boxes to encapsulate sensitive objects. Finally, we use a fully-connected CNN layer to draw a segmentation mask for each object.

The sensitive objects that we identify on each image for object detection including male genital, female genital, breast, and anus with some detected images that can be seen in Figure 4. We created a manual dataset with approximately 11000 images for training this model, gained the results of 70.28$\pm$1% mAP.

2) CNN-based classifier: In the second model, we modified a classification model[1] implemented from TensorFlow to extract features and divide images into binary classes: pornography and normal. The model was an adjustment from the original five-class detection to detecting just only pornography and normal images. The backbone we implemented on this model is ResNet50, using a single block for ResNetv2, without

[1]https://github.com/MaybeShewill-CV/nsfw-classification-tensorflow

Fig. 4. Sensitive objects detected with Mask R-CNN

a bottleneck, batch normalization then ReLu then convolution as described by this article [11].

In the experiment, we applied the Pornography dataset with 5-fold cross-validation method by [6], [7] for training and testing, finally gained the best prediction result of 92.00% accuracy.

IV. Experimental Results

A. Datasets

In experiment, we use two separate datasets for training and testing the adaptive model. We create a manual dataset to train our proposed model (Mask R-CNN model), and testing on two publicly datasets, the Pornography-800 [7] and the Pornography-2k [8] dataset.

1) Manual Dataset: The manual dataset has approximately 11.000 images that we collected from the internet and subdivided into three main categories, including:

- 5500 pornographic images with sensitive object content (a brief look of those images can be seen in Figure 4)
- 2750 sexy images (with high-rate skin color, bikini, erotic or sexual content,...)
- 2750 normal images (landscapes, people, animals,...)

Those images are divided into three different sets: the train set with 8000 images, the validation set with 2000 images and the test set with 1000 images and each set includes 50%, 25% and 25% of those categories above, respectively. The datasets total number of images and the quantity of images in the three major categories are shown in Table I.

TABLE I
Images in the manual dataset

Categories	Train set	Val set	Test set	**Total**
Pornographic	4000	1000	500	5500
Sexy	2000	500	250	2750
Normal	2000	500	250	2750
Total	**8000**	**2000**	**1000**	**11000**

In the 5600 pornographic images, we annotated 4 main sensitive objects with mask polygon for Mask R-CNN detection, including:

- Breast
- Male genitals
- Female genitals
- Anus

2) Publicly Dataset: We experiment with our method on the Pornography-800 dataset and the Pornography-2K dataset. The Pornography-800 dataset includes 800 videos and the Pornography-2k includes 2000 videos, divided into two parts, 50% non-pornographic videos and 50% pornographic videos with various lengths from several seconds to about half an hour relatively.

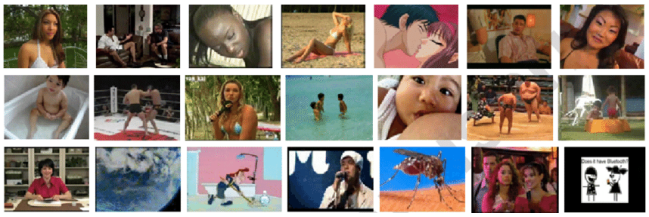

Fig. 5. Pornography-2k dataset. Top Row: pornographic videos. Middle Row: difficult cases of non-pornographic videos. Bottom Row: easy cases of non-pornographic videos

B. Results

On the result, we work with two publicly datasets and, Pornography-800 and Pornography-2k dataset. On each video, we extract one key-frame every 5 seconds and tests our model on those frames.

In the initial classification, we extract images which ε are greater than 0.8 will obviously be a pornographic image, and the images have ε lesser than 0.2 are classified as obviously non-pornographic image.

Consequently, we divided into two different experimental directions which are suitable for two approaching models for the concluding detection of the rest key-frames. On one hand with Mask R-CNN, we detect sensitive objects on each key-frames to identify it. Then, we calculate with the previous results from the initial classification to determine whether the video is pornographic or not. We have experimented with various evaluation methods between the number of pornographic frames and non-pornographic frames, and we concluded that the best accuracy result would be achieved when the pornographic video is considered to have 15 or more pornographic key-frames and in contrast, the video is normal. The prediction of Open-NSFW combine with Mask R-CNN on the Pornography-800 and the Pornography-2k are 92.13% and 90.40%, respectively.

On the other hands, we tested the CNN-based classifier by using the same 5-fold cross validation protocol applied in [7], [8]. The final video label is obtained by majority voting over the frames, and the final results is the average of results that we calculated through the 5-fold cross testing. We achieved an accuracy of 92.00% on the Pornography-800 and 84.82% on the Pornography-2k dataset.

Table II presents a comparison of the video classification accuracy between our approaches and the methods of Avila et al. [7], Vitorino et al. [13], Xinyu Ou et al. [14] and Caetano et al. [12] work result on the Pornography datasets. Overall, our results have a significantly better than other methods with the

TABLE II
RESULTS ON THE PORNOGRAPHY-800 AND PORNOGRAPHY-2K DATASETS

Method	Pornography-800	Pornography-2k
Avila et al. [7]	89.50%	N/A
Vitorino et al. [13]	N/A	88.00%
Xinyu Ou et al. [14]	85.30%	N/A
Caetano et al. [12]	**92.40%**	N/A
NSFW + Mask R-CNN	92.13%	**90.40%**
NSFW + CNN Classifier	92.00%	84.82%

Pornography-2k dataset, and slightly lower than the accuracy of [12] with Pornography-800 dataset, gaining the second-best result on the comparison. These results demonstrate the good performance of our method compared with other studies, and we aim to increase the efficiency of our model in future works.

V. CONCLUSIONS AND FUTURE WORK

In conclusion, it would appear that we proposed an pornographic content detector method using multi-level CNN models. The proposed model is divided into two main phases: the initial classification and the concluding detection. The initial classification recognizes obviously normal images and pornographic images quickly. Then, we use the concluding detection to identify the rest of unclassified images. We approach the final decision by two different models: CNN-based classifier for feature extraction and Mask R-CNN for object detection.

During the experiment, we create a manual dataset for training Mask R-CNN, and apply the training protocol in [7], [8] for CNN-based classifier. The adaptive model can be extended to videos using the sampling approach to extract key-frames. We identify pornographic videos primarily by making major voting between pornographic frames and non-pornographic frames. At the result, we obtained the accuracy of 92.13% on the Pornography-800 dataset and 90.40% on the Pornography-2k dataset by using Mask R-CNN model for sensitive objects detection. Our model achieved a great performance when comparing with other results on the same pornographic datasets.

For future works, we will seek to promote the performance and accuracy of our model for better results. Furthermore, we also want to approach some new methods to help us increase the quality of the Mask R-CNN prediction to achieve higher recognition precision, reaching the state-of-the-art.

ACKNOWLEDGMENTS

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2019-26-02.

REFERENCES

[1] Kailong Zhou, Li Zhuo, Zhen Geng, Jing Zhang, Xiao Guang Li, *Convolutional Neural Networks Based Pornographic Image Classification*, 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), pp.198-205, 2016.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollr and Ross B. Girshick, *Mask R-CNN* 2017 IEEE International Conference on Computer Vision (ICCV), pp.2980-2988, 2017.

[3] Waleed Abdulla, *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*, GitHub repository, GitHub, https://github.com/matterport/Mask_RCNN, 2017.

[4] Ali Ahmadia, Mehran Fotouhib, Mahmoud Khaleghic, *Intelligent classification of web pages using contextual and visual features*,Applied Soft Computing, 2010.

[5] A. A. Zaidan, H. Abdul Karim, N. N. Ahmad, B. B. Zaidan, A. Sali, *An automated anti-pornography system using a skin detector based on artificial intelligence: A Review*, International Journal of Pattern Recognition and Artificial Intelligence Vol. 27 No. 4, 2013.

[6] Ana P. B. Lopes, Sandra E. F. de Avila, Anderson N. A. Peixoto Rodrigo S. Oliveira and Arnaldo de A. Arajo, *A Bag-of-features approach base on HUE-SIFT descriptor for nude detection*, Signal Processing Conference, 2009 17th European, IEEE, pp. 15521556, 2009.

[7] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, Arnaldo de A. Arajo, *Pooling in Image Representation: the Visual Codeword Point of View*, Computer Vision and Image Understanding (CVIU), volume 117, issue 5, pp.453-465, 2013.

[8] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E.Valle, S. Goldenstein, A. Rocha, *Pornography classification: the hidden clues in video space-time*, Forensic Science International (Forensic Sci. Int.), volume 268, pp.4661, 2016.

[9] Fudong Nian, Teng Li, Yan Wang, Mingliang Xu, Jun Wu, *Pornographic Image Detection Utilizing Deep Convolutional Neural Networks*, Neurocomputing, 2015.

[10] Jay Mahadeokar, Gerry Pesavento, *Open Sourcing a Deep Learning Solution for Detecting NSFW Images*, https://yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for, 2016.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2016), *Identity Mappings in Deep Residual Networks*, Computer Vision ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 1114, 2016, Proceedings, Part IV, volume 9908, pp.630-645, 2016.

[12] Carlos Caetano, Sandra Avila, William Robson Schwartz, Silvio Jamil F. Guimares, A. Arnaldo de A. Arajo, *A mid-level video representation based on binary descriptors: A case study for pornography detection*, Neurocomputing, volume 213, pp. 102114, 2016.

[13] Paulo Vitorino, Sandra Avila, Mauricio Perez, Anderson Rocha, *Leveraging Deep Neural Networks to Fight Child Pornography in the Age of Social Media*, Journal of Visual Communication and Image Representation, Volume 50, pp.303-313, 2018.

[14] Xinyu Ou, Hefei Ling, Han Yu, Ping Li, Fuhao Zou, Si Liu, *Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-coarse Strategy*, ACM Transactions on Intelligent Systems and Technology, volume 8, no. 5, 2017.