A Novel Pornographic Visual Content Classifier based on Sensitive Object Detection

Dinh-Duy Phan[1], Thanh-Thien Nguyen[2], Quang-Huy Nguyen[3], Hoang-Loc Tran[4],
Khac-Ngoc-Khoi Nguyen[5], Duc-Lung Vu*[6]
Faculty of Computer Engineering, University of Information Technology
Vietnam National University Ho Chi Minh City, Viet Nam

***Abstract*—With the increasing amount of pornography being uploaded on the Internet today, arises the need to detect and block such pornographic websites, especially in Eastern cultural countries. Studying pornographic images and videos, show that explicit sensitive objects are typically one of the main characteristics portraying the unique aspect of pornography content. This paper proposed a classification method on pornographic visual content, which involved detecting sensitive objects using object detection algorithms. Initially, an object detection model is used to identify sensitive objects on visual content. The detection results are then used as high-level features combined with two other high-level features including skin body and human presence information. These high-level features finally are fed into a fusion Support Vector Machine (SVM) model, thus draw the eventual decision. Based on 800 videos from the NDPI-800 dataset and the 50.000 manually collected images, the evaluation results show that our proposed approach achieved 94.06% and 94.88% in Accuracy respectively, which can be compared with the cutting-edge pornographic classification methods. In addition, a pornographic alerting and blocking extension is developed for Google Chrome to prove the proposed architecture's effectiveness and capability. Working with 200 websites, the extension achieved an outstanding result, which is 99.50% Accuracy in classification.**

***Keywords*—Computer vision; image proccessing; object detection; pornographic recognition and classification; blocking extension; machine learning; deep learning; CNN**

I. Introduction

In the digital era, information has become a powerful weapon to manipulate the development of a society. People nowadays are easy to find and upload any information they want on the Internet. On the one hand, these pieces of information sometimes are good and bring positive value to the human race. On the other hand, many harmful kinds of negative information can also be found with only some keywords by anybody, even children. Pornography content is one of them. Many women worldwide are victims of sexual cybercrimes because their private videos are spread on the social network. Furthermore, pornographic content is even restricted in many countries. From the above problems, the need for an effective pornographic visual content detector is necessary.

Many efforts have been made recently to classify pornographic images among normal ones. In the early stages, the skin-based method has been applied. These approaches check whether images have nude people or not based on the ratio of the exposed skin. Another approach is handcrafted features-based, which uses various descriptors to extract key point low-level features in an image. A visual codebook may be learned by applying the k-means algorithm on a training set. After that, the trained codebook could represent any images, and a classifier may detect pornographic ones. Applying low-level features to identify obscene images, it achieved significantly higher performance than skin-based methods. However, representing images by visual words still suffers a severe problem since it ignores the spatial relationship, which is very important to represent the image's content. The state of the art approaches for this classifying problem are based on deep learning methods. These approaches build models with neural networks that let it learns features from the image's contents itself.

Previous studies [1], [2] [3], have implemented the above approaches and achieved some particular success, especially the deep-learning-based approach has proposed a potential development for this problem. However, the context in images is very complicated, and there are many similarities between a pornographic image and a normal one. The normal image that contains a large region of exposed skin (e.g., swimming, wrestling, people wearing bikinis) or contains people with sexy poses may be misclassified as pornography. Misclassification may seriously affect the user experience while using the internet. To avoid this problem, we have to clarify what is pornography. According to Oxford Advanced Learner Dictionary, "Pornography is magazines, DVDs, websites, etc., that describe or show naked people and sexual acts to make people feel sexually excited, especially in a way that many other people find offensive."[1] From the definition, an image can be determined as pornographic if it contains naked people. In other words, pornographic images are images that consist of human's sensitive objects and organs such as breasts, anus and genitals. We called this method is the sensitive object-based approach.

Based on that insight, this paper presents a novel approach for pornographic content detection and classification, which not only leverages the advantages of previous approaches but also compensates for these methods' weaknesses. Our main approaching strategy is using the effectiveness of object detection to identify pornographic elements in visual content with steady prediction. Additionally, skin and human recognition are also integrated into our method to distinguish between images with humans from images without humans, but with human like skin colors such as sand or wood. These two modules not only capable to augment the classification's decision but also can be served as the counterweight to prevent the potential bias that comes from object detection. Ultimately, a linear classification

[1]https://www.oxfordlearnersdictionaries.com/definition/english/pornography

model SVM is adapted to make the eventual prediction using features from these three modules, eliminate the limitations during the process while keeping their advantages.

In summary, our main contributions are as follows:

- A method is proposed to detect sensitive objects on the human body. Then, these objects were combined with skin and human features feeding to SVM model to conclude if an image is pornographic or not.
- Based on the pornographic textual and visual detection, a pornographic alerting and blocking extension is developed as an initial gate warning user before accessing inappropriate websites.

The rest of the paper is organized as follows. Section II describes details of some relevant approaches included skin-based approaches, handcrafted features-based approaches, and deep learning-based approaches. Section III presents theoretical background with four main models that we used to detect sensitive objects: Mask R-CNN, YOLO, SSD, and Cascade Mask R-CNN. The detail of our proposed method will be presented in Section IV. In order to demonstrate the applicability of the proposed model, a web-based extension will be introduced in section V. After that, Section VI provides the details of our dataset, experiments, and the achieved results. Finally, we give conclusions and suggest some future works in Section VII.

II. Related Work

Pornographic recognition and classification have been a long-lasting problem with many existing studies. In this paper, previous studies are grouped into four primary categories according to their main approach, which are: skin-based approach, handcrafted features-based approach, deep learning-based approach and object-based approach.

The skin-based approach is considered the earliest and most basic method of recognizing visual pornography, as it focuses on estimating the exposed skin area within an image. Several low-level or high-level features such as shape, color, facial, or belly can be utilized in order to achieved higher estimation. The final decision of this method mostly depended on mathematic or statistic thresholds based on the predicted skin ratio and the image's ground-truth. Previous studies with skin-based approach can be determinating seperate skin ratios on different body areas [4], combining facial recognition and skin recognition [1], or utilizing a pre-train discrimination neural network following a skin extractor [5]. Although the effectiveness of recognizing most cases of pornographic visual content, as [6] pointed out, the skin-based approach comes with vital weakness as its performance can be affected highly by the quality and resolution of the input image. Additionally, the strong resemblance between athletic or sporting images with a vast amount of exposed skin and pornographic images under the skin-ratio threshold can be a serious problem of recognizing the right one, which certainly affects the performance of skin-based methods.

The handcrafted feature-based approach, mostly applying the Bag-of-Visual-Word technique, extracts key-point features inside visual data using feature descriptors and maps it into vectors. To obtain features, various feature descriptors can be adapted such as: SIFT and Hue-SIFT [2]; BossaNova [7], which is an image representation based on histogram of distance; or Temporal Robust Features [8], a spatial-temporal interest point descriptor that adapts Fisher Vector intermediate representation. When these representative vectors are formed, a visual codebook can be constructed and concatenated with discrimination algorithm likes SVM to determine the pornographic of input visual content. However, the diversity and discrepancy of pornographic visual content along with the omission of spatial relationship make it difficult to determine the appropriate features to describe visual pornography comprehensively.

The robustness of deep learning-based methods in recent years has brought a significant result in visual pornography identification. Rather than adapting descriptors to obtain representative features manually, the advantage of neural network architecture helps model extract features and refine learning parameters themselves, thus improve the performance of predicting pornographic visual content. Previous studies with pornography detection [9], [10] utilize pre-trained neural network models and fine-tune on a custom pornography dataset instead of training the architecture from scratch. However, training strategies and network customization must be made to optimize architectural's parameters and prevent the model suffer from over-fitting.

Lastest studies about visual pornography [11], [12], [13] identify that sensitive objects, organs or body parts, likes vulva, dildo, female breast or anus, have been known to carry rich information of describing a major part of pornographic content. By recognizing these organs or objects with object detection models, it is possible to draw the conclusion about the pornographic of input visual content with high accuracy. Normally, studies under the object-based method often adapted exist object detector and fine-tuned it on a custom annotated dataset. However, the selection of appropriate pornographic organs or objects as well as the method of annotating depended heavily on the study scale and perspective. Noticeably, Tabone et al. [14] proposed seven sexual organs and objects for pornography classification included buttocks, female breast, female genital (which are divided into two sub-classes: female genital posing and female genital active), male genital, sex toy and benign object. Eventually, they annotated those classes with a five-set labeled point: one center point and four perpendicularly offset for each. Additionally, some common sexual poses or intercourses can be learned in pattern to improve the performance of prediction. Although methods under object-based approach can ensure the high performance of prediction on most pornographic visual data, in some special cases, the strong resemblance of some everyday object with sexual organs under certain conditional (light, viewpoint, or shape) such as sausage and male genital make it difficult to make the right decision.

In our previous studies with sensitive objects and organs on object-based approach in [15] [16], not only we developed a pornographic object detector with Mask R-CNN to identify the most four common sensitive objects includes women breast, male/female genitals and anus, but also we utilized the training strategy of object detection model with two-step learning to overcome the false positive prediction of pornographic objects, thus improve the overall performance of prediction.

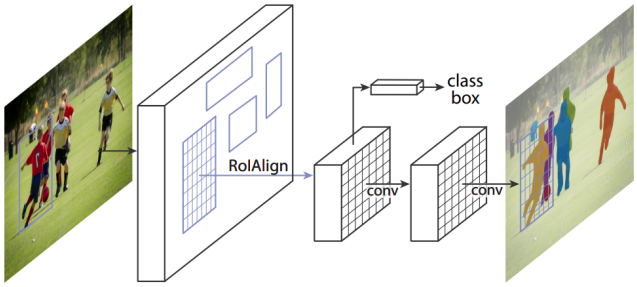


Fig. 1. Mask R-CNN Main Architecture [17].

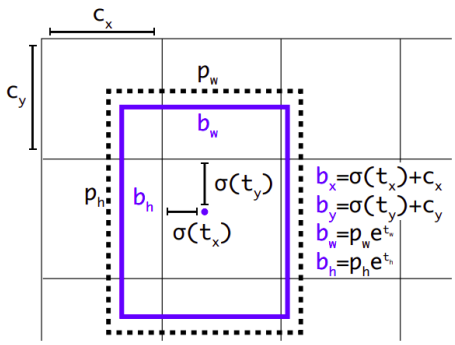


Fig. 2. YOLO Bounding Box with Location Prediction [21].

III. Theoretical Background

A. Mask R-CNN

Mask R-CNN [17] is a robust object detection and instance segmentation framework that draws bounding boxes for object detection and generates a high detail segmentation mask for instance segmentation. In this paper, the Matterport's version of Mask R-CNN [18] is adapted, which uses ResNet101 [19] and Feature Pyramid Network (FPN) [20] as main backbone.

The Mask R-CNN model included multiple stages and branches from the original Faster R-CNN framework, whith the main architecture can be seen in Fig. 1. Tts network hierarchy included the Region Proposal Network branch for object bounding boxes identification and Box Regression Network branch for bounding box regression. Furthermore, a simple FCN is added to predict the segmentation masks on each Region of Interest (RoI) in a pixel-to-pixel manner. Furthermore, for the core operation of feature extraction, Mask R-CNN applies pooling algorithm RoIAlign to extract small feature maps called Region of Interest (RoI) features to be well aligned and preserved with the corresponded object in the pixel level. Hence, RoIAlign helps Mask R-CNN achieves pixel-to-pixel alignment and high accuracy in prediction and mask segmentation. On each RoI, the new multi-task loss function of Mask R-CNN has been introduced, which combines the loss of classification, localization, and segmentation mask by Eq. 1.

$$\mathcal{L}_{\text{Mask R-CNN}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} \tag{1}$$

The loss function $\mathcal{L}_{\text{class}}$ and $\mathcal{L}_{\text{box}}$, which is adapted in Mask R-CNN, are defined by Eq. 2 and Eq. 3 respectively, with p_i, p_i^*, t_i, t_i^* are the predict class, ground truth label, predicted bounding box coordinates and ground truth bounding box coordiniates.

$$\mathcal{L}_{\text{class}} = \frac{1}{N_{\text{cls}}} \sum_i -p_i^* \log p_i - (1 - p_i^*) \log(1 - p_i) \tag{2}$$

$$\mathcal{L}_{\text{box}} = \frac{\lambda}{N_{\text{box}}} \sum_i p_i^* \cdot L_1^{\text{smooth}}(t_i - t_i^*) \tag{3}$$

On the other hand, the loss function of instance segmentation mask $\mathcal{L}_{\text{mask}}$ is defined as the average binary cross-entropy loss between the ground-truth mask y_{ij} and the predicted mask y_{ij}^k, only in terms of class k-th in Eq. 4:

$$\mathcal{L}_{\text{mask}} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} [y_{ij} \log y_{ij}^k + (1 - y_{ij}) log(1 - y_{ij}^k)] \tag{4}$$

B. Yolo

You Only Look Once (YOLO) is the state-of-the-art, real-time object detection system proposed by Joseph Redmon. YOLO [22] uses a single neural network to predict bounding boxes and class probabilities directly from full images in one evaluation. In YOLOv2 [21], they remove the fully connected layers from YOLO and use anchor boxes to predict bounding boxes. This research applied YOLOv3 [23], the latest version of YOLO using Darknet53 network architecture that helps the model achieves the outstanding benchmark comparing to others. Fig. 2 illustrates how YOLO model predicts the bounding box location on the image. Network model identifies 4 coordinates of the bounding boxes t_x, t_y, t_w, t_h, then the bounding box predictions is calculated by Eq. 5:

$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \\ b_w = p_w e^{t_w} \\ b_h = p_h e^{t_h} \end{cases} \tag{5}$$

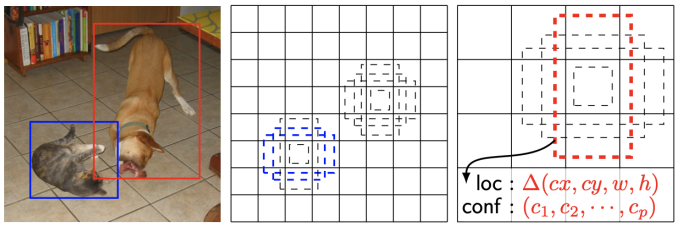
with the the top left corner point of the predicted box(c_x, c_y) as well as the width and height p_w, p_w, respectively. To optimized training parameters, YOLO adapted the loss function as shown in Eq. 6

$$\mathcal{L}_{\text{YOLO}} = \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{loc}} \tag{6}$$

in terms of

$$\mathcal{L}_{\text{loc}} = \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} \Big[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \Big] \tag{7}$$

and



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Fig. 3. SSD Framework. [24] (a) SSD only needs an Input Image and Ground Truth Boxes for Each Object During Training. In a Convolutional Fashion, SSD Evaluates a Small Set (e.g. 4) of Default Boxes of Different Aspect Ratios at Each Location in Several Feature Maps with Different Scales (e.g. 8 x 8 and 4 x 4 in (b) and (c)).

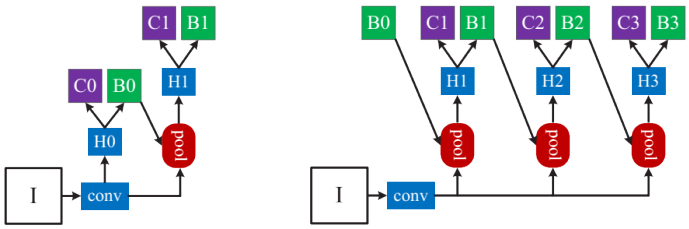


Fig. 4. Architecture of Faster-RCNN (left) and Cascade-RCNN (right) [25].

$$\mathcal{L}_{\text{cls}} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \left[1_{ij}^{\text{obj}} + \lambda_{\text{noobj}} (1 - 1_{ij}^{\text{obj}}) \right] (C_{ij} - \hat{C}_{ij})^2 + \sum_{i=0}^{S^2} \sum_{c \in \mathcal{C}} 1_i^{\text{obj}} \left[p_i(c) - \hat{p}_i(c) \right]^2 \tag{8}$$

C. SSD

Single Shot Multibox Detector (SSD) is proposed by Wei Liu et al. [24] which is known as the object detection model with high accuracy and speed. Unlike the other models, which need to hypothesize bounding boxes, followed by some complicated steps to handle them, SSD uses a fixed set of default boxes, then predicts scores and box offsets by using small convolutional filters to feature maps as illustrate in Fig. 3. One of the advantages of SSD is that it can detect objects of mixtures of scales with high accuracy. To achieve that, SSD produces predictions of different scales from feature maps, which are also in various scales and explicitly separates predictions by aspect ratio. For the balance between swiftness and precision, SSD runs a convolution network on the input image only once and then computes the feature map. SSD framework can be summarized as follows:

- The training phase inputs images with ground truth boxes for each object.
- With different scales of images, SSD evaluates a small set of default boxes in different aspect ratios at each location in several feature maps.
- For each default box, the framework predicts both offsets and confidences of shape for all object categories. At the training phase, SSD tries to match these default boxes to the ground truth boxes.

During the training process, SSD adapts the multi-task loss function described in Eq. 9:

$$\mathcal{L}_{\text{SSD}} = \frac{1}{N} (\mathcal{L}_{\text{conf}} + \alpha \mathcal{L}_{\text{loc}}) \tag{9}$$

in terms of:

$$\mathcal{L}_{\text{loc}} = \sum_{i,j} \sum_{m \in \{x,y,w,h\}} 1_{ij}^{\text{match}} L_1^{\text{smooth}} (d_m^i - t_m^j)^2 \tag{10}$$

and

$$\mathcal{L}_{\text{cls}} = -\sum_{i \in \text{pos}} 1_{ij}^k \log(\hat{c}_i^k) - \sum_{i \in \text{neg}} \log(\hat{c}_i^0), \text{ where } \hat{c}_i^k = \text{softmax}(c_i^k) \tag{11}$$

D. Cascade Mask R-CNN

Cascade R-CNN [25] is a multi-stage object detection based on Faster R-CNN. It adds two extra stages to the standard two-stage Faster R-CNN architecture, as can be seen in Fig. 4. Besides, the data of each stage is trained with increasing IoU thresholds, which aims to reduce the overfitting problem between the training and testing processes.

The Cascade R-CNN architecture can be integrated with other models to improve the performance. For experiments, we use Cascade Mask R-CNN [26], a combination of Cascade R-CNN and Mask R-CNN, on Detectron2 [27]. As being the extended version of Mask R-CNN, Cascade Mask R-CNN shares a similar loss function with Mask R-CNN, as shown in Eq. 12.

$$\mathcal{L}_{\text{Cascade Mask R-CNN}} = \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} \tag{12}$$

IV. Methodology

In this article, instead of using low-level image features, a novel model for extracting high-level features, which can be used for classification afterward is proposed. The proposed method consists of two stages as shown in Fig. 5. The first stage is the high-level feature extraction which combines 10 high-level features from 3 blocks: sensitive objects detection (8 features), Human presence (1 feature) and Skin color extraction (1 feature). The second stage of the model is an SVM classifier for making the final decision of discriminating pornographic or non-pornographic content. The details of each block are described below.

A. First Stage - High-level Feature Extraction

1) Sensitive object detection module: When examining pornographic visual contents, including images and videos, we realize that the sensitive objects in the human body appear frequently, so that these objects can be used for detecting pornographic content effectively. In other words, we can apply object detection models to detect these objects and use it as high-level features. Throughout the image, four sensitive

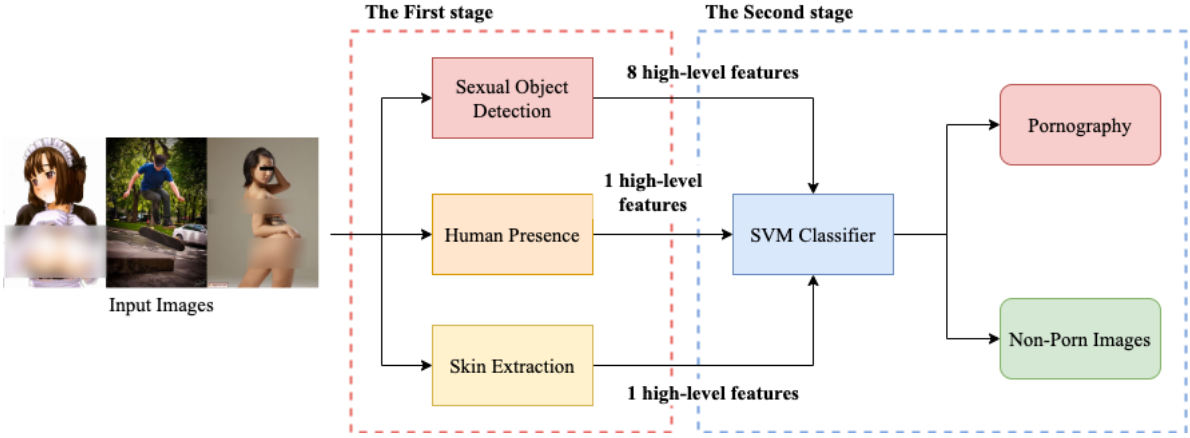


Fig. 5. Block Diagram of the Proposed Pornography Classifier.

objects and organs included female breasts, male/female genitals, and anus, are identified along with bounding boxes and corresponding confidence scores. Two high-level features are extracted for each type of sensitive object, including the number of detected objects and each type's highest confidence score, thus making eight high-level features for feeding to the main fusion model.

2) Human presence module: Images or videos are considered sexual if and only if it contains people or parts of people, therefore the second information can be used to solve the task in this paper is human presence. In other words, human detection is a sub-task of the object detection problem, thus the object detection model can be adapted to identify humans' presence on visual content. The existing human in images detected is another high-level feature we proposed to improve our model.

In this module, a human detector on Detectron 2, which used the pre-trained model on COCO dataset, is adapted to identify humans' existence from the input image. The segment mask of detected people was then used to recognize these people's presence on the image and adapted as the input for the skin extraction module. Thus the feature of human presence, in terms of a binary value, is adapted to the main fusion model.

3) Skin extraction module: For the skin extraction, we adapted two color spaces, HSV and YCbCr, to recognize skin areas from the image. While HSV is being used for its advantage in describing a high-quality color as well as reducing the problem of illuminating color identification, YCbCr is adapted for its advantage to describe skin on various races with significant results. Moreover, YCbCr is one of the most popular color spaces applied to skin extraction methods.

The color spaces thresholds that we applied for skin extraction in our proposed approach is described in Eq. 13. Based on these boundaries, we decided whether the pixel is describing skin or not by combining two extracted results from HSV and YCbCr range.The ratio of the total body skin areas on the image, which value varies between 0 and 1, as another high-level feature along nine upper describing features.

$$\begin{cases} 0 \leq H \leq 17 \\ 15 \leq S \leq 170 \\ 0 \leq V \leq 255 \end{cases} \text{ and } \begin{cases} 0 \leq Y \leq 255 \\ 85 \leq Cb \leq 135 \\ 13 \leq Cr \leq 180 \end{cases} \quad (13)$$

The skin extraction task is performed on the human segmentation areas, which we get from the Human Presence module described above. We calculated the skin ratio only on human segments, not on the entire original image. This gives a better view of the human body's skin ratio, where the higher exposed skin, the higher pornographic probabilities.

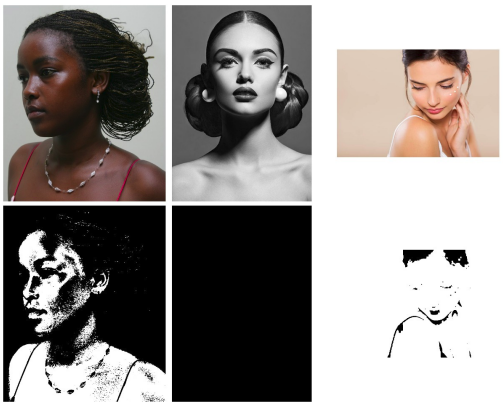


Fig. 6. Skin Extraction Algorithms Results.

However, we must admit that the skin extraction algorithm has a certain weakness when it deals with gray-scale or black/white images as well as images with skin-liked color, such as sand, wood, etc. Under these cases, the skin extraction neither working nor achieving great results, as described in Fig. 6. Thus, that will be one of the problems we have to overcome in the future.

B. The Second Stage - SVM Classifier

Ten high-level features from the three modules above then feed into a discriminative SVM model, which works as a fusion mechanism, to become the proposed approaching method. Generally, the overview of the fusion model is illustrated in

Fig. 5 where the sensitive object detection module can be replaced by Mask R-CNN, YOLO, SSD, and Cascade Mask R-CNN model, respectively. For the SVM model, kernel Radial Basis Function (RBF) is adapted as it comes with our proposed method's highest performance.

In the proposed model, sensitive object information plays the most crucial part in determining the sexuality of input visual content as it accounts for 80% of input features feeding to the fusion module. Moreover, the presence of humans from input images and the ratio of explicit body skin on these human segments play an additional role in ensuring the effectiveness of the method's performance. Ten high-level features feeding to the SVM model can leverage their effectiveness in prediction as well as eliminate the limitations from their own modules.

V. Pornographic Blocking Extension for Website

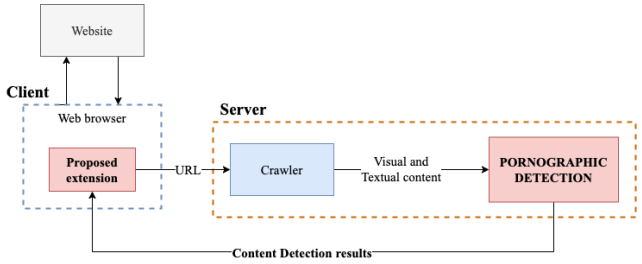


Fig. 7. Procedure of Pornographic Blocking Extension.

From our previous work in [28], [29] and [15], we developed an extension on Google Chrome for pornographic website recognition works as an initial gate alerting and blocking people before accessing the unappropriated website. The extension involves pornographic textual content identification using the NaiveBayes classifier and visual content identification based on YOLO sexual object detection to determine the pornography of the input website.

A. Procedure Flow

The main procedure of this extension can be observed in Fig. 7. Initially, visual and textual contents are crawled from the input website on the crawler module. Then, pornographic visual and textual identification modules are adapted to identify how pornography the website is.

In the visual identification module, a sensitive object detection was adapted to determined if website images are pornography or not. Due to the fast execution and great performance, YOLOv3 was chosen for the extension. On the other hand, the Vietnamese/English pornographic textual classifier was adapted from [29] for the textual identification module. This classifier uses the NaiveBayes algorithm to discriminate whether the sensitiveness of textual contents from the website.

Based on the predicted results of these contents, the extension alerting three safety levels to the user, which are: (1) Safety: textual and visual contents are both recognized as safety; (2) Porn-Type-1: one of the textual or visual content is recognized as pornography; (3) Porn-Type-2: textual and visual contents are both recognized as pornography.

TABLE I. The Website Dataset

	Safety	Porn-Type-1	Porn-Type-2	**Total**
Vienamese	65	5	11	81
English	35	15	69	119
Total	100	20	80	200

TABLE II. Result on the Manual Dataset

Object detection model	**Object-based method (object detection)**	**Proposed method (Skin + Human + object detection)**
Mask R-CNN	81.64%	**85.63%**
YOLO	93.87%	**94.06%**
SSD	88.93%	**89.32%**
Cascade Mask R-CNN	**93.43%**	93.39%

B. Pornographic Website Dataset

To evaluate the performance of the pornographic alerting and blocking extension, we collected a website dataset including 100 pornographic websites and 100 safety websites with Vietnamese and English languages. The websites are divided into three categories: Safety, Porn-Type-1, and Porn-Type-2 corresponded with the definitions of extension outcome notifications. The distribution of the pornographic website dataset by category is described in Table I.

VI. Experimental Results

A. Results and Evaluation

On the experiment, we evaluated our method with a manual image dataset along with the public dataset NPDI-800 [7]. The manual image dataset includes 25.000 normal and 25.000 pornographic images, while the NPDI-800 dataset contains 400 pornographic videos and 400 normal (non-porn) videos.

For video evaluation, one frame per second is extracted to determine the pornography of input video. The evaluating strategy of pornographic video is modelized as shown in Fig. 8. As can be seen, extracted frames are classified into two groups, namely pornographic frames, and normal frames, by a classifier. Based on the ratio of porn frames on total frames, the input video is classified as pornographic or normal.

Fig. 9 and Fig. 10 show our experiments on NPDI-800 dataset with a various "porn rate" threshold from 1% to 15%. With that strategy, we achieved the best result when we considered pornographic video contain 1% or more frames detected as pornographic. To assess the effectiveness of our proposed approach, we have done two experiments on the same testing set: (i) using object detection's outputs directly to distinguish pornography from normal images, (ii) feeding the SVM with object detection's output information combining with extracted skin and human segmentation features.

TABLE III. Result on NPDI-800 Dataset

Object detection model	**Object-based method (object detection)**	**Proposed method (Skin + Human + object detection)**
Mask R-CNN	54.38%	**74.13%**
YOLO	94.88%	**94.88%**
SSD	89.00%	**89.13%**
Cascade Mask R-CNN	76.63%	**82.75%**

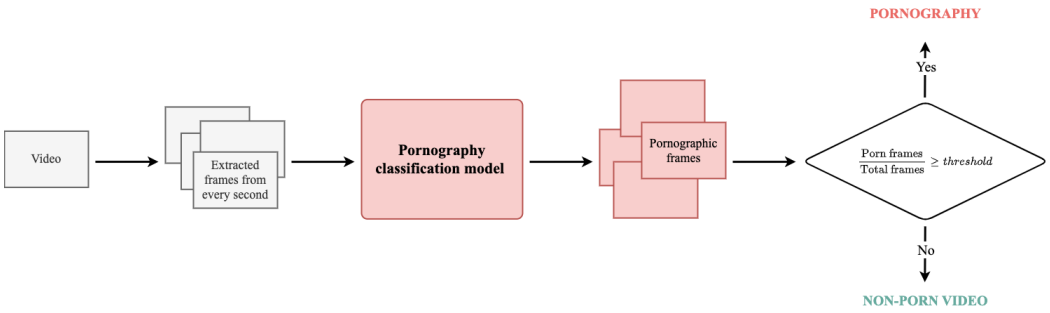


Fig. 8. Evaluating Strategy on NPDI-800 Dataset.

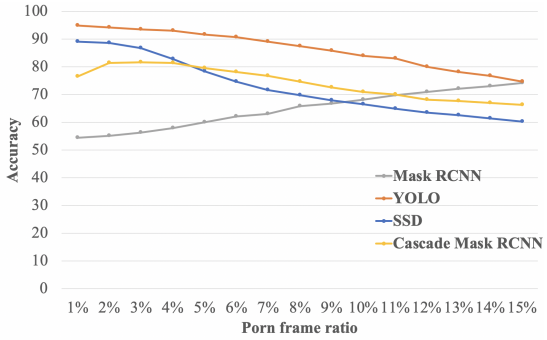


Fig. 9. Using Object Detection Approach to Evaluate on NPDI-800 Dataset with Different Porn Frame Ratios.

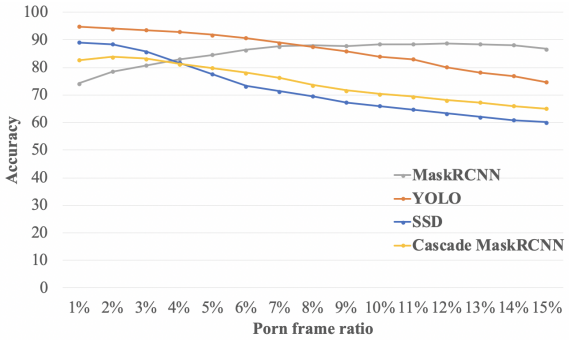


Fig. 10. Using Proposed Method to Evaluate on NPDI-800 Dataset with Different Porn Frame Ratios.

TABLE IV. COMPARED RESULTS ON THE NPDI DATASET

Methods	NPDI-800
SIFT & Hue-SIFT Decriptor [2]	84.60%
BossaNova + SVM [7]	89.50%
DCNN-based Learning [11]	97.50%
BossaNova Video Descriptor [30]	92.40%
Deep Multicontext Network [3]	85.30%
Our model	**94.88%**

TABLE V. PERFORMANCE OF EXTENSION IN ALERTING WEBSITES

Website	Ground-Truth	Correct-Predicted	Accuracy
Safety	100	99	99.00%
Porn-Type-1	20	20	100.0%
Porn-Type-2	80	62	77.50%
Total	200	181	**90.05%**

TABLE VI. PERFORMANCE OF EXTENSION IN CLASSIFYING WEBSITES

Website	Ground-Truth	Correct-Predicted	Accuracy
Non-Porn	100	99	99.00%
Porn	100	100	100.0%
Total	200	199	**99.50%**

The first experiment was done by using the results from sensitive object detection directly to judge an image is pornographic or not. If the model detects at least one of four sensitive objects in an image, this image will be concluded as pornography and vice versa. From the results on the manual dataset, which showed in Table II, the YOLO model achieved the best Accuracy score with 93.87%. While evaluating videos, this approach gives us the highest accuracy of 94.88% when the YOLO model is used to detect sensitive objects on the NPDI-800 dataset. Detail result at 1% of porn rate ratio has been shown in Table III. As shown in Fig. 9, both YOLO and SSD models had the best result at 1% porn frame ratio while Mask R-CNN achieved its best at 15% and this number was 3% of porn frame ratio on Cascade Mask R-CNN.

The second experiment was conducted by adapting our proposed approach, which used an SVM with ten high-level features inputs from sensitive object detection, human presence, and skin extraction modules. Generally, we achieved better results in comparison with the method using sexual object detection algorithms only, which can be observed on Table II and Table III. On the NPDI-800 dataset, Mask R-CNN and Cascade Mask R-CNN's results improved significantly when adapting the proposed approach, from 54.38% to 74.13% and 76.63% to 82.75%, respectively. However, the highest results still are achieved by the YOLO model with 94.06% accuracy on our manual dataset and 94.88% on the NPDI-800 dataset. The comparison with other methods on the NPDI-800 dataset can be observed in Table IV. As can be seen, the performance of our method in the NPDI-800 dataset is quite impressive.

B. Website Extension Experiment

In the experiment with pornography blocking extension, our manual website dataset is adapted to evaluate the effectiveness of recognizing not-safe-for-work websites. The results in prediction on three types of website can be observed on Table V and Table VI.

As can be observed, the extension achieved the outstanding results with 99.00% Accuracy in recognizing Safety websites, 100.0% and 77.50% Accuracy on Porn-Type-1 and Porn-Type-2 websites, respectively. However, the extension recognizes Porn-Type-2 websites with only 77.50% Accuracy as some certain websites are identified into Porn-Type-1. Thus, all the pornographic websites during the experiment are identified by the extension with the absolute 100% Accuracy in recognition, leading to the total result of 99.50% Accuracy in porn/non-porn website classification. Still, we have to notice that there might be some confusing cases, which are websites containing sensitive contents. These kinds of websites might be civil intent, such as sex education lecture/advice or nude photographs for medical purposes. This makes the model quite difficult to determine the suitable decision, which might reduce our extension's performance in the real practical implementation.

VII. Conclusion

This paper proposed a new approach to identify pornographic visual content based on sensitive object detection and skin color. The proposed approach detects four sensitive objects including female breast, male or female genital and anus. Then these sensitive object information are used as high-level features along with human precense and skin extraction information as input of the fusion SVM model. Eventually, the SVM model decides whether the input visual content is pornography or not. Applying the four most notable object detection algorithms, which are Mask R-CNN, YOLO, SSD, and Cascade Mask R-CNN, sensitive object features are identified. We achieved the best results on our custom dataset and the public NPDI-800 dataset, which are 94.06% and 94.88% Accuracy respectively, when the YOLOv3 is adapted as the sensitive object detector and RBF is used as the SVM kernel.

In addition, to prove the effectiveness of our proposed method in practical application, an extension for alerting and blocking pornographic website is built. Measuring the accuracy by surfing 200 websites, the extension has shown impressive results: 99.00% of normal websites and 100% of pornographic websites was identified correctly.

Acknowledgment

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

References

[1] R. Balamurali and A. Chandrasekar, "Multiple parameter algorithm approach for adult image identification," *Cluster Computing*, vol. 22, no. 5, pp. 11 909–11 917, 2019.

[2] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, and A. de A. Araújo, "A bag-of-features approach based on hue-sift descriptor for nude detection," in *2009 17th European Signal Processing Conference*, 2009, pp. 1552–1556.

[3] X. Ou, H. Ling, H. Yu, P. Li, F. Zou, and S. Liu, "Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 5, pp. 1–25, 2017.

[4] D. C. Moreira and J. M. Fechine, "A machine learning-based forensic discriminator of pornographic and bikini images," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

[5] K. Zhou, L. Zhuo, Z. Geng, J. Zhang, and X. G. Li, "Convolutional neural networks based pornographic image classification," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, 2016, pp. 206–209.

[6] A. Zaidan, H. A. Karim, N. Ahmad, B. Zaidan, and 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. 04, p. 1350012, 2013.

[7] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. A. AraúJo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[8] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Pornography classification: The hidden clues in video space–time," *Forensic science international*, vol. 268, pp. 46–61, 2016.

[9] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," *Neurocomputing*, vol. 210, pp. 283–293, 2016.

[10] J. Mahadeokar and G. Pesavento, "Open sourcing a deep learning solution for detecting nsfw images," *Retrieved August*, vol. 24, p. 2018, 2016. [Online]. Available: yahooeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for

[11] Y. Wang, X. Jin, and X. Tan, "Pornographic image recognition by strongly-supervised deep multiple instance learning," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 4418–4422.

[12] C. Tian, X. Zhang, W. Wei, and X. Gao, "Color pornographic image detection based on color-saliency preserved mixture deformable part model," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 6629–6645, 2018.

[13] H. A. Nugroho, D. Hardiyanto, and T. B. Adji, "Nipple detection to identify negative content on digital images," in *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, 2016, pp. 43–48.

[14] A. Tabone, A. Bonnici, S. Cristina, R. A. Farrugia, and K. P. Camilleri, "Private body part detection using deep learning," in *ICPRAM*, 2020, pp. 205–211.

[15] Q. H. Nguyen, K. N. K. Nguyen, H. L. Tran, T. T. Nguyen, D. D. Phan, and D. L. Vu, "Multi-level detector for pornographic content using cnn models," in *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020, pp. 1–5.

[16] H. L. Tran, Q. H. Nguyen, D. D. Phan, T. T. Nguyen, D. L. Vu *et al.*, "Additional learning on object detection: A novel approach in pornography classification," in *International Conference on Future Data and Security Engineering*. Springer, 2020, pp. 311–324.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[18] W. Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow https://github.com/matterport," *GitHub*, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[21] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.

[22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[25] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[26] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "Hybrid task cascade for instance segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4969–4978.

[27] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2 https://github.com/facebookresearch/detectron2," *GitHub*, 2019.

[28] T. A. Dinh, T. B. Ngo, and D. L. Vu, "A model for automatically detecting and blocking pornographic websites," pp. 244–249, 2015.

[29] D. D. Phan, V. T. Nguyen, and D. L. Vu, "Nhan dang trang web co noi dung khieu dam dua tren text va website," in *Fundamental and Applied Information Technology (FAIR) Domestic Conference in Vietnam*, 2019.

[30] C. Caetano, S. Avila, W. R. Schwartz, S. J. F. Guimarães, and A. d. A. Araújo, "A mid-level video representation based on binary descriptors: A case study for pornography detection," *Neurocomputing*, vol. 213, pp. 102–114, 2016.