Child Pornography Detection with Computer Vision

André Brandão
Department of Computer Science
Faculty of Sciences of the University of Porto
Porto, Portugal
up201908477@fc.up.pt

***Abstract*—Internet and social media has become part of most people's daily lives, fast and mass content distribution is now available to anyone. As such, tools need to be devised to prevent shared content from breaking platform's terms and conditions, or allowing for the dissemination of illegal content that may offend people's susceptibility, such as child pornography. I propose in this paper an algorithm that combines classical computer vision and deep learning algorithms to detect child pornography images.**

***Index Terms*—Computer Vision, Pattern Recognition, Deep Learning, Age Classification, Pornography Detection.**

I. Introduction

Automated child pornography detection is very important not only to internet and social media companies to block the distribution of this type of contents but also to law enforcement authorities to identify and locate possible victims.

During two weeks in 2018 police offers from 21 countries spent 8 hours per day manually analysing more than 32 millions of images of child pornography [12]. Besides expensive and slow, it's very dehumanizing for the people that had to carefully analyze the images. In this paper I propose an approach that combines face detection using Viola-Jones algorithm with deep learning for age classification and pixel manipulation using thresholding operations for pornography detection to detect child pornography.

II. Related Work

Many proposals have been made to detect pornography in images, most of them are focused on skin detection, converting the original image to different color spaces and then applying studied thresholds to create skin masks, after that the percentage of skin in the image is computed, if it is above a predefined threshold then it's considered pornography, otherwise not pornography ([6], [7], [8], [9], [10], [11]). More recently, techniques focused on deep learning such as transfer learning using pre-trained complex networks like AlexNet, GoogLeNet or VGG16 have been applied to tackle this problem, having outperformed the more classical techniques ([4], [5]). Transfer learning takes advantage of the knowledge of the best features to retrieve from an image, already learned by the pre-trained network, and uses it to train the the last part of the neural network to identify the new classes.

III. Child Detection

In order to detect if a picture contains a child or not the algorithm must find all the faces in the image and classify them according to their age, if there is at least one face classified as under 18 the picture will be classified as positive, in the child detection phase.

A. Face Detection

To find all the faces in the image I used the Viola-Jones algorithm [1]. This algorithm, when applied to human face detection, tries to find the most relevant features for a human face using rectangular filters.

1) Haar Features: The Viola-Jones algorithm classifies an object based on simple features' values. By using features instead of pixels directly it's possible to encode ad-hoc domain knowledge. This features are based in the haar wavelet functions [2].

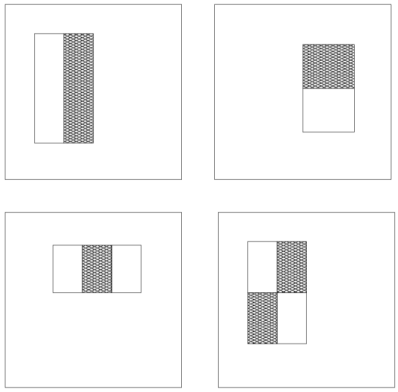

Fig. 1. Examples of Haar features. (source: [1]).

Figure 1 represents example features and there value can be computed by subtracting the sum of pixels in the white rectangles from the sum of the grey ones.

With this features we can identify many characteristics from the human face.

Fig. 2. Features' application. (Left) Eyes detection (Middle) Smile detection (Right) Nose detection.

2) Integral Images: It's possible to conclude, by looking at figure 2 that we have to compute the value of the filters for every possible position and size to find the actual features in

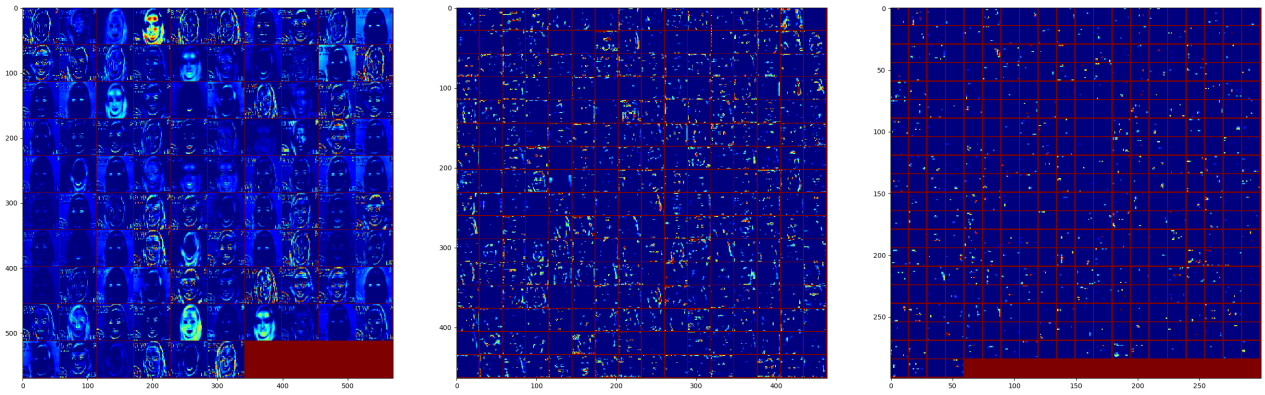

Fig. 3. Output of the convolutional layers (Left) First layer (Middle) Second layer (Right) Third layer.

the image. In order to reduce the time complexity of this task, the integral image is computed. It consists on computing, for every pixel i, the sum of all the pixels to the left and top of the pixel i.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'), \tag{1}$$

- ii - integral image.
- i - original image.

By using ii, obtained using the formula above, it's possible to compute any rectangular sum inside the image using four array references.

3) Cascading: A cascading of classifiers was constructed to achieve better performance while reducing the computation time. This concept is applied to eliminate the majority of the sub-windows of the image that are completely unrelated to faces, using simpler classifiers. As the sub-windows pass to the next levels, the classifiers become more complex and harder to pass as a false positive (see figure 4).

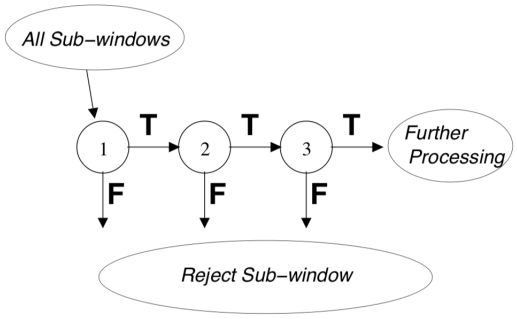

Fig. 4. Cascade strategy diagram (source: [1]).

B. Age Classification

Age classification poses a more complex problem to solve using classical computer vision strategies, mostly because it's hard to define what features better describe a person's age like hair color, wrinkles, hair loss, etc. Even knowing this features it's extremely difficult to detect them accurately. Therefore, deep learning was used to solve this part of the problem.

1) Convolutional Neural Networks: When talking about deep learning applied to image classification, CNN (Convolutional Neural Networks) are the most common and accurate type of neural networks that are used nowadays.

A convolutional neural network joins deep learning with convolutional kernels, in a normal neural network we use neurons as a way to apply general matrix multiplication, in a CNN we replace this operations with a convolutional operation.

a) Convolutional Layers: The convolutional layers will try to learn which convolutional kernels (ex.: Sobel edge detector) are the most interesting to identify the classes, however we don't need to specify anything about the filters, the network will learn all by itself. The convolutional operation (equation 2) consists in a sliding window that computes the operation between the sub-window ($\begin{bmatrix} a & b \\ e & f \end{bmatrix}$) and the filter ($\begin{bmatrix} w & x \\ y & z \end{bmatrix}$) for every sub-window in the image.

$$\begin{bmatrix} a & b & c & d \\ e & f & g & h \\ i & j & k & l \\ m & n & o & p \end{bmatrix} \circledast \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \tag{2}$$

$$= \begin{bmatrix} aw + bx + ey + fx & bw + cx + fy + gz & cw + dx + gy + hz \\ ew + fx + iy + jz & fw + gx + jy + kz & gw + hx + ky + lz \\ iw + jx + my + nz & jw + kx + ny + oz & kw + lx + oy + pz \end{bmatrix}$$

In figure 5 it's represented the filters from the first layer of the trained convolutional neural network used to classify people's age. We can observe some familiar filters, like vertical edge detection filters (in positions: $(0, 5), (1, 9), (7, 6)$) and horizontal edge detection filters (in positions: $(1, 0), (3, 8), (5, 9)$) for example. There are also many filters that we don't understand what they are detecting

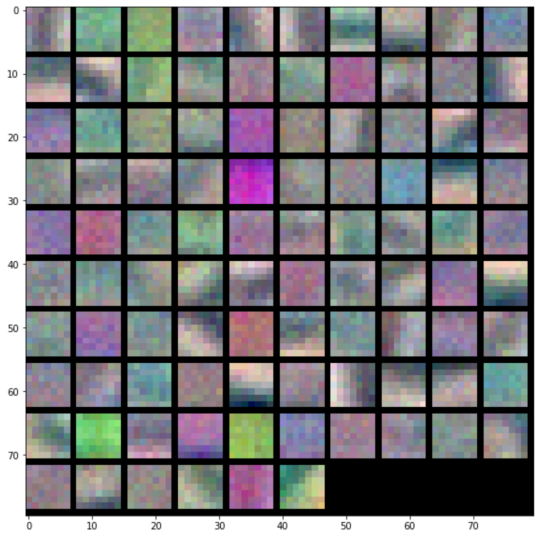

Fig. 5. First convolutional layer filters.

and that is the power off deep learning, the ability to find patterns that humans cannot see.

The filters shown in figure 5 are applied to the input image that will produce the output presented in figure 3. In the first layer is possible to identify many similarities to the original image (the same used in figure 2), in the second layer this similarities start to fade away and in the third layer the images are already to abstract for humans to understand.

b) Fully Connected Layers: Fully connected layers follow the same principle as the traditional multi-layer perceptron. The goal of this layers is to take the interesting features found by the convolutional layers and use them to learn the best mapping between input features and labels using backpropagation. After the network is trained it will use the features extracted from the image by the convolutional layers to classify the image into the correct label.

c) Architecture: The architecture used to classify the age of a face was the one proposed in [3] and represented in figure 6. The three color channels from an RGB image are processed by the network, the image is first rescaled to 256 x 256 pixels and a cropped version of 227 x 227 pixels is fed to the network.

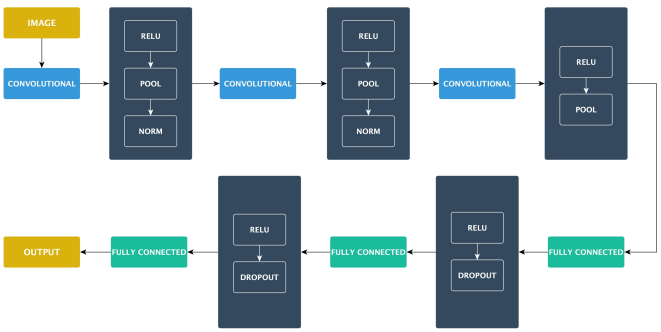

Fig. 6. Network architecture.

The RELU, POOL, NORM and DROPOUT layers between convolutional and fully connected layers are used to prevent overfitting. The ReLU layer keeps the positive part of its argument, *i.e.* $max(0, x)$, the pooling layer (POOL) is used to reduce the spatial dimensions, but not depth, this improves performance, prevents overfitting and it allows some translation invariance. The NORM layer normalizes the data, in this network the Local Response Normalization (LRN) is used. The DROPOUT layer consists in ignoring some of the neurons during the training phase, not taking them in account during a particular forward or backward pass, this will help the network to prevent overfitting.

C. Combining the Algorithms

Finally, to identify if there is at least one child in the picture the face detection algorithm is used to identify the faces in the picture, then this information is used to cut the faces from the image and build new images with just the faces of the people. For every image the deep neural network is used to determine if the face belongs to a child or not, if there is at lest one, then the algorithm returns true, otherwise false.

1) Results: The deep neural network is implemented in Caffe framework and it was trained by G. Levi and T. Hassner in a GPU machine with 1.526 CUDA cores and 4GB of video memory. The dataset used to train and test the CNN was the Adience benchmark (https://talhassner.github.io/home/projects/Adience/Adience-data.html#agegender) that consists of images uploaded to Flickr from smartphone devices. The entire dataset contains 26.580 images of 2.284 different subjects, in figure 7 we can see the distribution of the ages.

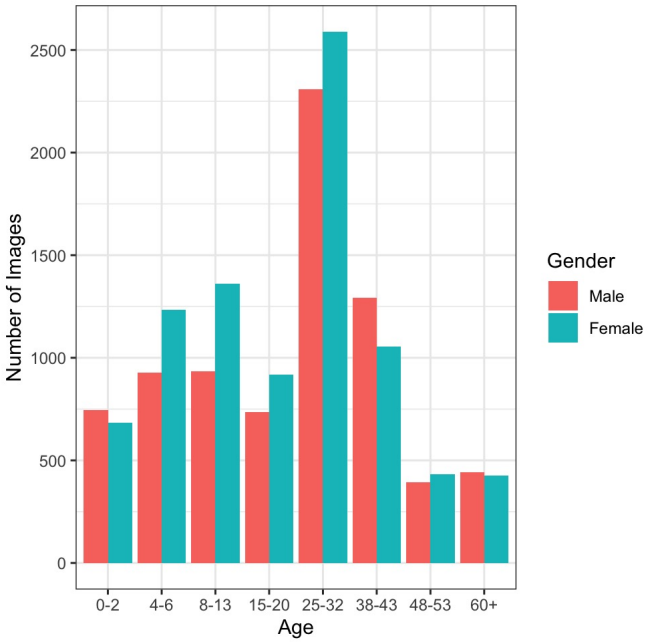

Fig. 7. Number of images per age.

To understand how the model was performing a confusion matrix was made (table I). The model performs well for ages $0-2$, $4-6$ and $8-13$, however in age $15-20$ the model isn't very good since it classifies more than half of the observations

TABLE I
CONFUSION MATRIX FOR AGE DETECTION MODEL.

		True Age							
		0-2	4-6	8-13	15-20	25-32	38-43	48-53	60+
Predicted Age	0-2	**0.699**	0.147	0.028	0.006	0.005	0.008	0.007	0.009
	4-6	0.256	**0.573**	0.166	0.023	0.010	0.011	0.010	0.005
	8-13	0.027	0.223	**0.552**	0.150	0.091	0.068	0.055	0.061
	15-20	0.003	0.019	0.081	**0.239**	0.106	0.055	0.049	0.028
	25-32	0.006	0.029	0.138	0.510	**0.613**	0.461	0.260	0.108
	38-43	0.004	0.007	0.023	0.058	0.149	**0.293**	0.339	0.268
	48-53	0.002	0.001	0.004	0.007	0.017	0.055	**0.146**	0.165
	60+	0.001	0.001	0.008	0.007	0.009	0.050	0.134	**0.357**

as $25-32$, what is understandable given that most people in this age group want to look older than they actually are. The opposite also happens for ages $38-42$ and $48-53$ where the model classifies almost half of the observations as $25-32$ probably because most people in this age groups want to look younger.

IV. PORNOGRAPHY DETECTION

Pornography is any material that triggers sexual thoughts in an explicit way. In order to detect this type of images the algorithm will try to identify the pixels that contain skin, this way we can compute the percentage of skin in a image. Then, given a predefined threshold it's possible to classify the image as pornographic or not.

A. Skin Detection

To detect skin in a picture the algorithm will focus on detecting the human skin color pixels by converting the image to three different color spaces (HSV, RGB and YCrCb) and then applying thresholds to the converted images, the final result will be the average of the three different masks. The best thresholds used for each color space were found by [10].

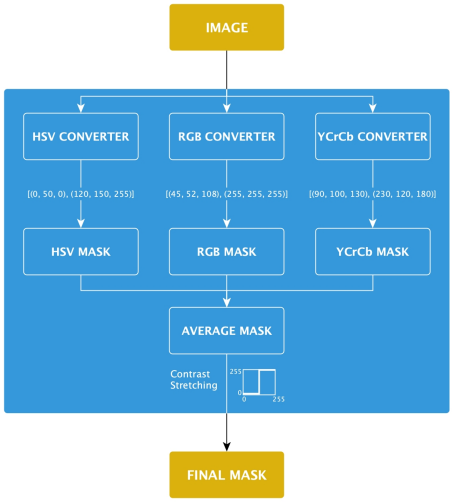

Fig. 8. Skin detection pipeline.

In figure 8 we can see the diagram of the algorithm for skin detection as well the thresholds used for each converter, after we obtain the masks for each converter we average the values for each pixel and apply a contrast stretching to get the final mask where every value is 0 or 255.

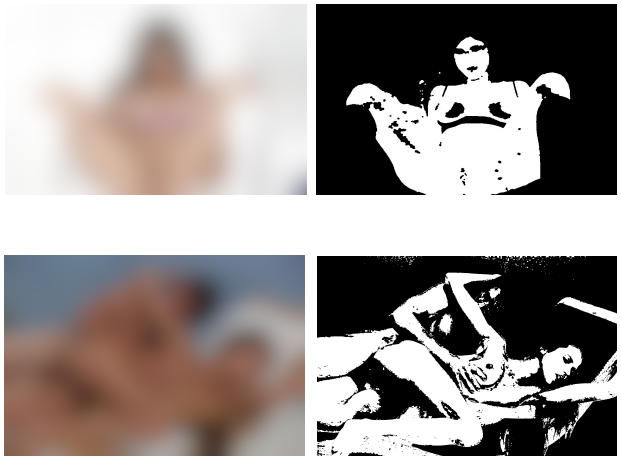

Fig. 9. Skin detection algorithm (Left) Original images blurred (Right) Final masks.

In figure 9 we can observe the resulting masks (right column images) from the algorithm described before applied to the images on the left column. The results are good, detecting most of the skin pixels although it also detects some pixels that aren't skin, *e.g.* the borders of the bench in the second image which in this case the consequences aren't critical since we can allow false positives (Non child pornography images classified as child pornography images) but we want as few false negatives as possible.

B. Face Close Up Problem

One problem that arises when using the strategy above is that when a picture of a close up face appears, most of the pixels are going to be skin, therefore the algorithm is going to classify this image as pornographic. To solve this problem the resulting bounding boxes for the faces found in the Viola-Jones algorithm were used to compute a ratio $\frac{\#SKIN\_PIXELS}{\#FACE\_PIXELS}$.

Using the figure 10 and the strategy described above we get the following values:

- #TOTAL_PIXELS: 209160.
- #FACE_PIXELS: 97344.
- #SKIN_PIXELS: 74602.
- $\frac{\#SKIN\_PIXELS}{\#FACE\_PIXELS}$: 0.766.

Fig. 10. Example of the face close up problem.

Since the ration $\frac{\#SKIN\_PIXELS}{\#FACE\_PIXELS}$ is less than one, *i.e.* there are more pixels for the face than for the skin, we can conclude that this image is not pornographic, even having more than $\frac{1}{3}$ of its pixels representing skin.

C. Thresholds

After retrieving the skin/face ratio and skin/total ratio we now need to define the thresholds to apply to this features. To better define these thresholds a decision tree model was used to automatically select the best values according to a dataset of 60 images, 30 pornographic images and 30 non pornographic.

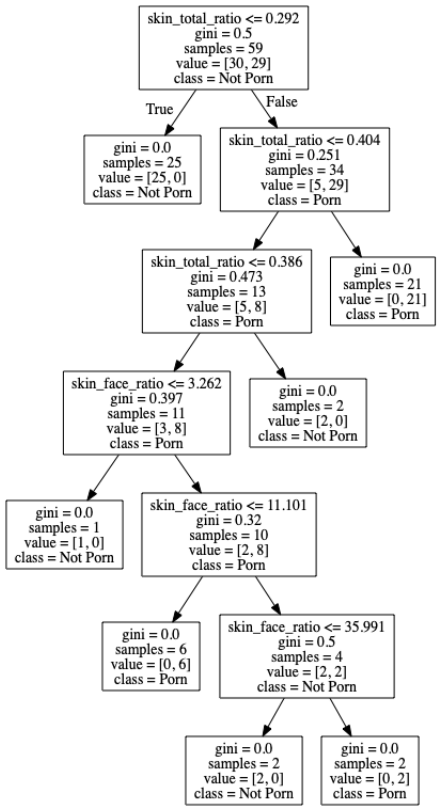

Fig. 11. Decision tree visualization.

Figure 11 represents the final tree created by the model, the skin/total ratio is used first and then the skin/face ratio is used to deal with the face close up problem.

D. Results

Using 70% of the images to train de decision tree model and 30% to test. The confusion matrix is presented in table II.

The accuracy of the model was 94.44%, however the dataset is to small to draw any viable conclusions from this result.

TABLE II
CONFUSION MATRIX PORNOGRAPHY DETECTION.

		True	
		Porn	Not Porn
Pred	Porn	**10**	1
	Not Porn	0	**7**

V. CHILD PORNOGRAPHY DETECTION

Finally, in order to detect if a image contains child pornography we need to join all the algorithms described in the previous sections.

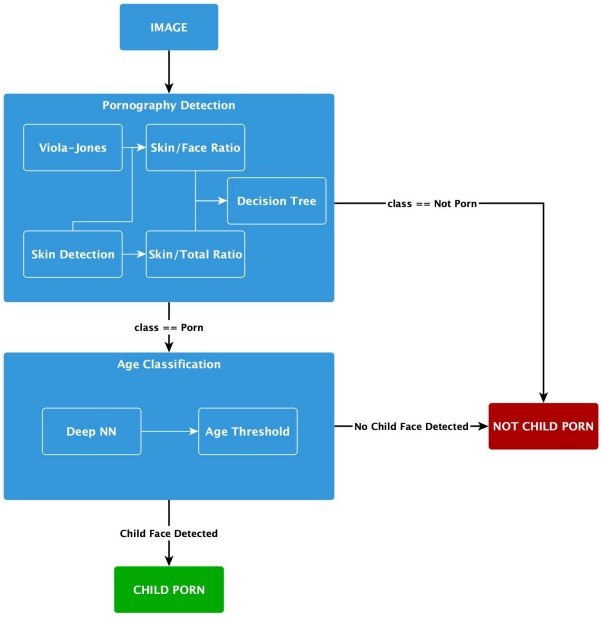

Fig. 12. Diagram of the entire pipeline.

The image is first fed into the pornography detection module where all the faces in the image will be identified, the skin/face ratio and the skin/total ratio are computed and the output of the decision tree will decide if the image is pornographic or not. If it's not pornography, then the image is classified as "not child pornography", if it's pornography, the image of the faces found in Viola-Jones will be used in the age classification model. If there is at least on child face, then the image is classified as "child pornography", otherwise as "not child pornography".

VI. LIMITATIONS

Despite the good results obtained in both age detection and pornography detection algorithms this strategies have flaws, for example if there is occlusion of the head, *e.g.* the head is covered by some object or it doesn't appear in the image, the algorithm will fail. The pornography detection algorithm is also susceptible to shadows and objects with color similar to human skin color, in figure 13 the background color is very similar to the skin color and lena's body has many shadows, this results in a bad mask.

Fig. 13. Example of a bad mask.

A. Conclusion and Future Work

It's no possible to test the model in real child pornography images given the legal issues so it's impossible to actual understand how good or bad the complete model would perform. However the two strategies work well alone so I consider that it probably has some potential when working together.

With this work I can conclude that automated process to detect child pornography can be viable in the real world, however more complex algorithms should be used with real child pornography images to train the models, because these images have different characteristics from normal pornography like the lighting, camera quality and others.

REFERENCES

[1] P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features*, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 2001, pp. I-I.

[2] Haar, Alfréd, *Zur Theorie der orthogonalen Funktionensysteme*, 1910 Mathematische Annalen.

[3] G. Levi and T. Hassner, *Age and gender classification using convolutional neural networks*, 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015, pp. 34-42.

[4] Moustafa, M. (2015). *Applying deep learning to classify pornographic images and videos*. ArXiv, abs/1511.08899.

[5] A. Gangwar, E. Fidalgo, E. Alegre and V. González-Castro, *Pornography and child sexual abuse detection in image and video: A comparative evaluation*, 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017), Madrid, 2017, pp. 37-42.

[6] Saxen, F., Al-Hamadi, A. (2014). *Color-based skin segmentation: An evaluation of the state of the art.* 2014 IEEE International Conference on Image Processing (ICIP).

[7] Sengamedu, S. H., Sanyal, S., Satish, S. (2011). *Detection of pornographic content in internet images.* Proceedings of the 19th ACM International Conference on Multimedia - MM '11.

[8] Lumini, Alessandra, and Loris Nanni. *Fair comparison of skin detection approaches on publicly available datasets.* arXiv preprint arXiv:1802.02531 (2018).

[9] Kolkur, S., et al. *Human skin detection using RGB, HSV and YCbCr color models.* arXiv preprint arXiv:1708.02694 (2017).

[10] Gasparini, Francesca and Schettini, Raimondo. (2006). *Skin segmentation using multiple thresholding - art. no. 60610F*. Proceedings of SPIE - The International Society for Optical Engineering.

[11] Conaire, C. O., O'Connor, N. E., Smeaton, A. F. (2007). *Detector adaptation by maximising agreement between independent data sources.* 2007 IEEE Conference on Computer Vision and Pattern Recognition.

[12] Carolina Branco, *Polícias analisam 32 milhões de imagens de pornografia infantil. PJ descobre foto tirada em Portugal*, Observador, Nov. 16, 2018. [Online], Available: https://observador.pt/especiais/policias-analisam-32-milhoes-de-imagens-de-pornografia-infantil-pj-descobre-foto-tirada-em-portugal [Accesed Dec. 8, 2019].