Forensic Image Inspection Assisted by Deep Learning

Felix Mayer
Fraunhofer SIT
Rheinstr. 75
Darmstadt, Germany 64295
felix.mayer@sit.fhg.de

Martin Steinebach
Fraunhofer SIT
Rheinstr. 75
Darmstadt, Germany 64295
martin.steinebach@sit.fhg.de

ABSTRACT

Investigations on the charge of possessing child pornography usually require manual forensic image inspection in order to collect evidence. When storage devices are confiscated, law enforcement authorities are hence often faced with massive image datasets which have to be screened within a limited time frame. As the ability to concentrate and time are highly limited factors of a human investigator, we believe that intelligent algorithms can effectively assist the inspection process by rearranging images based on their content. Thus, more relevant images can be discovered within a shorter time frame, which is of special importance in time-critical investigations of triage character.

While currently employed techniques are based on black- and whitelisting of known images, we propose to use deep learning algorithms trained for the detection of pornographic imagery, as they are able to identify new content. In our approach, we evaluated three state-of-the-art neural networks for the detection of pornographic images and employed them to rearrange simulated datasets of 1 million images containing a small fraction of pornographic content. The rearrangement of images according to their content allows a much earlier detection of relevant images during the actual manual inspection of the dataset, especially when the percentage of relevant images is low. With our approach, the first relevant image could be discovered between positions 8 and 9 in the rearranged list on average. Without using our approach of image rearrangement, the first relevant image was discovered at position 1,463 on average.

CCS CONCEPTS

• **Applied computing → Evidence collection, storage and analysis**; *Investigation techniques*; • **Information systems** → *Information retrieval*;

KEYWORDS

text tagging, content-based image retrieval, information retrieval, ranking, deep learning

ARES '17, Reggio Calabria, Italy
© 2017 ACM. 978-1-4503-5257-4/17/08…$15.00
DOI: 10.1145/3098954.3104051

ACM Reference format:
Felix Mayer and Martin Steinebach. 2017. Forensic Image Inspection Assisted by Deep Learning. In *Proceedings of ARES '17, Reggio Calabria, Italy, August 29-September 01, 2017,* 9 pages.
DOI: 10.1145/3098954.3104051

1 MOTIVATION

In this work, we discuss how intelligent filtering or pre-selection of images can help a forensic investigator who needs to find out if a huge set of images includes child pornography. The amount of images which can be stored on devices is steadily increasing and leads to a work load at forensic investigations which cannot be handled efficiently. As an example, the Apple iPhone 7 is available with 256 GB memory. With an assumed average image size (huge camera photos and smaller Internet downloads) of 1.5 MB, in theory more than 170,000 photos could be found on a single smartphone. A 3 TB NAS is available for € 150. Assuming similar file sizes, 2 million images could be stored here.

Interviews with German law enforcement officials revealed that the amount of image files which have to be investigated in cases of suspected child pornography possession, exceeds their available capacities by far. Therefore, and due to the fact that confiscated devices have to be handed back to the owner after a certain period of time, the vast majority of child pornography collectors is assumed to remain undetected[1].

We suggest assisting forensic investigation by providing a presorted list of images to the investigator. Images of potentially relevant nature are at the top of the list; those of lesser relevance are at the bottom. This helps to quickly decide if there is illegal content stored on the device. And the investigator can use his limited concentration on those images more likely to be of relevance.

To achieve this, we evaluate the ability of existing deep learning networks to identify nudity and erotic scenes in images. If a mixed image set can be divided into one group (A) of images that contain pornographic and erotic content and another group (B) of the remaining images, showing A first to the investigator removes the noise B would produce if the investigator would work on the mixed set of A and B. If the network is even able to provide tags that help to quantify the relevance of the individual images, for example based on assumed age or sexual activity, sorting can be even more helpful as more relevant images within A would be placed on the top of the list. An investigation on a huge image collection with only a small fraction of pornography would be assisted largely by this.

To the best of our knowledge, today's forensic image investigation is only based on white- and blacklisting powered by robust or

[1] http://sit4.me/policeinterview

cryptographic hashes as well as skin color filters. Skin color filtering is known to be error-prone and therefore of limited use, even if supported by detection of anatomic shapes. Blacklisting has the drawback that only known images can be identified, whitelisting commonly fails to be of help due to the vast amount of images produced today. Thus, our approach is the first one which applies deep learning techniques in order to support forensic triage tasks on large image datasets.

1.1 Assumptions

A strategy for improving the screening of images with potentially illegal content as suggested above is only sensible if some limitations regarding an investigation are accepted. At least in Germany, due to the federal organization of the police, this can differ from state to state. Sorting images only makes sense if the investigator is not forced to view all images on the device without an exception but white- and blacklisting based on cryptographic hashes. In some interviews, investigators state that for an actual case only a certain amount of relevant images have to be found, regardless of how many relevant images remain undetected. In other words, screening all images is not compulsory. Others argue that it is vital that all photos are found as new victims could be identified and saved.

From our perspective, we see our approach as an optimization of used resources. The amount of images that can be screened is limited by the number of investigators. It is regularly stated by the police that there are more cases than current investigators can handle, causing a significant delay in forensic examinations of cases. If a tool is able to reduce the amount of wasted resources to enable an overall higher performance at screening illegal content at the cost of potentially missing some evidence, from a technical point this seems to be acceptable. If, by the help of our approach, an investigator could handle five times as many cases as before but may be wrong in one out of five, this would still result in three times more successful investigations than without the given strategy.

But we are aware that this may not be a technical decision but a legal one. Important for this work is the fact that at least some states would allow a sorting and filtering as described in the following paper. We also need to stress that during this work no illegal content has been used. Distinguishing between relevant and irrelevant content is simulated by identifying images containing nudity and separating them from those that do not. From our point of view, the results are sufficiently promising to trust that in real cases similar success rates can be achieved.

2 BACKGROUND

This section provides some basic understanding of deep learning which is the fundamental technology used in our approach. The term deep learning refers to machine learning algorithms based on artificial neural networks (ANNs) with many hidden layers. A layer is a collection of parallel artificial neurons, which are mathematical functions taking an arbitrary number of input values and returning one output value. This output value depends on an activation function, which means only if the input values are high enough, an output signal is produced (in analogy to mammalian brain cells this is also sometimes called firing). The hidden layers of an ANN are those between its input and output layer.

The layers are concatenated such that the outputs of neurons in layer n serve as inputs for the neurons in layer $n + 1$. This way, very complex non-linear mathematical functions can be built, which allow learning classification models for various types of data. The input layer takes original data samples and passes them on to the next layers, while the output layer finally predicts a class value for the given data sample. As the inputs of neurons are parametrized, the actual learning process of an ANN is a parameter optimization which is achieved via backpropagation.

There exist many different kinds of neurons and layers. One layer type which is commonly used in image processing tasks is the so-called convolution layer. An ANN containing this special type of layers is called convolutional neural network (CNN) and is the type of ANN we employed for our studies in this work. The main characteristic of CNNs is that, while learning to detect certain features (e. g. a car or a dog), their precise location in the image is not relevant. This is a very important characterisitc as the location of features naturally changes among different images and a detection algorithm has to deal with that.

3 STATE OF THE ART

For the identification of child pornography among a large quantity of images, which belongs to the field of forensic triage, there exist different approaches, some of which are applied in practice today. In this section we provide an overview of such techniques.

3.1 Image Blacklisting /Whitelisting

Blacklisting based on cryptographic hashes is the to our knowledge the only strategy to assist a forensic investigation centered on illegal images which is widely accepted. In Germany, a hash database with known child porn is used within the police. Included are also whitelisted images which look like child porn but are not. Blacklisting is the most efficient and reliable way to re-identify images, but also easy to counter: any change in the image file will result an a different hash. One alternative also applied in fighting child porn on the Internet are robust hashes [16] [13]. The best known example is PhotoDNA by Microsoft which is applied by various police forces.

Whitelisting by hashes is also possible by will lead to vast data collections. A forecast estimates that in 2017 1.3 trillion photos will be made. A white-list of these photos based on a 256 bit hash would have a size of 19 TB.

3.2 Skin Color Filters

In the field of skin color detection there has been a lot of research in the past. Those methods perform a pixel- or region-based image analysis operating on various color spaces like RGB, HSV and the like. The classification of skin and non-skin pixels can be done via a simple threshold or statistical and machine learning techniques. While achieving quite high true positive rates (up to 99%), most approaches exhibit rather high false positive rates as well (30% and higher) [5]. For example, [10] claim to achieve 95.6% classification accuracy with their skin pixel detection algorithm. But the image dataset ColorFERET, which they used for evaluation, contains mostly images with a homogeneous and very light background. This, of course, minimizes the overlap of skin and other pixels and allows to achieve such a good performance. Although for some application scenarios this might be realistic, it is not realistic in our scenario where images can be taken at arbitrary locations thus inducing much more color heterogeneity among different image regions.

Furthermore, skin color filters have some drawbacks for their application in our scenario: skin in grayscale images is hard to detect due to the loss of color information; varying illumination conditions and the existence of various skin tones increase the bias in the skin color distribution; non-skin image regions might get classified as skin due to an overlap with the learned skin color distribution; and the presence of many skin pixels in an image does not automatically imply the existence of pornographic content (e. g. close-up views of faces).

3.3 Machine Learning-based Forensic Triage

A technique to screen mobile devices for the presence of child pornography and other crime-related data was presented in [7]. Their approach was based on conventional machine learning algorithms and was intended to be used as a quick classification method in time-critical cases. In contrast to our approach it is fully automated, but only provides an indication of whether or not there is illegal content on the device. However, our approach allows the identification of actual means of evidence which can be used at court. Thus, the approahc presented in [7] should rather be seen as a tool for police forces on the spot, e. g. to decide if a device should be confiscated or not.

3.4 NSFW Detection

Both, nudity and pornographic content in images are often referred to as *NSFW* (*Not Safe For Work*). NSFW detection is the discipline to automatically identify corresponding imagery. Due to governmental guidelines in some countries for restricting the availability of erotic imagery on the Internet as well as company policies to deny access to such material, the research incentive in the field of NSFW detection is very high.

Former approaches were based on skin detection and human body part detection using classifiers with hand-crafted features, such as [12], [9] and [1], for which free JavaScript (nude.js[2]) and Python (nude.py[3]) implementations exist. [15] propose using color visual words for the detection of child pornography. Similar to our approach they aim to provide support for forensic investigators in the process of image inspection by filtering large image sets. Their algorithm comprises a statistical classification scheme based on the texture and color distribution of images, which are used to discriminate between child pornography and other images.

However, modern state-of-the-art approaches employ deep learning, specifically, deep convolutional neural networks (CNNs) which do not require any feature engineering beforehand. One big challenge in NSFW detection is the ability of algorithms to distinguish between harmless images, such as beach photos showing half-naked people in bikinis and trunks, and actual pornographic content. Considering solely the amount of skin color pixels in an image would not be sufficient for a reliable NSFW detection algorithm. Deep learning can overcome this challenge much better than other approaches, by training on massive image sets which allow the algorithm to learn specific features of pornographic content. Today, there exist several reliable deep-learning-based nudity algorithms, some of which are offered as commercial solutions (usually in the form of a web service). Those algorithms will be described in more detail in section 4.

4 NUDITY DETECTION BY DEEP LEARNING

In this section we describe the principles of deep-learning-based NSFW detection and identify the currently available solutions. Additionally we explain how investigations in child pornography cases can be supported by Deep Learning.

4.1 Deep-Learning-Based NSFW Detection

The concepts of artificial neural networks (ANNs) have been well-known for decades. However, the applicability of deep learning had to wait until the early 2010s. This is due to the fact that just recently the computational power of CPUs and especially GPUs reached a level where such algorithms can run within reasonable times. In particular, the training phase affords an intense amount of calculations performed on large datasets. In 2012 Krizhevsky et al. were the first to win the annual image classification competition ILSVRC[4] with a convolutional neural network (CNN) [6]. Just three years later, in 2015, He et al. claimed that their CNN even outperformed human-level classification performance [4]. Since then, CNNs quickly became the new state-of-the-art in image classification and other machine learning tasks, which applies not only to the research community but also to many commercial and non-commercial products.

Rather than asserting definite labels to an image, such as *NSFW* or *SFW* (this refers to the opposite of NSWF, i. e. "Safe For Work"), the classifiers rather return confidence scores for both labels. Those scores are numbers between 0 an 1 which can be interpreted as probabilities. For example, a NSFW score of 0.9 might be a strong indicator for actual NSFW content in a given image. For a given NSFW score p the corresponding SFW score is just its opposite, i. e. 1 - p. As nudity is a culture- and policy-dependent term, API users can adjust the sensitivity of the algorithms to their individual needs by setting an individual threshold value for the NSFW or SFW score.

4.2 Available NSFW Detection CNNs

Due to the high demand for the detection of NSFW content in digital images by public and private organizations, there already exist several commercial as well as free solutions based on deep learning. Some of them claim that their algorithms are up to 99% accurate. Commercial solutions are offered by Sightengine[5], Clarifai[6] and Nude Detect[7]. They are typically provided as a web service and

[2] https://www.patrick-wied.at/static/nudejs

[3] https://github.com/hhatto/nude.py

[4] http://www.image-net.org/challenges/LSVRC

[5] https://sightengine.com/

[6] https://clarifai.com/

[7] http://nudedetect.com

fees are based on the number of requests per month. For small amounts up to a few thousand requests per month their use is free of charge. Yahoo released an open-source NSFW detection API[8] in 2016. Another free NSFW detection CNN, called Illustration2Vec (I2V)[9], was originally trained on Anime images [11]. In contrast to other solutions, I2V additionally allows the detection of various explicit pornographic actions and specific body parts. However, those algorithms that focus on pure NSFW detection achieve much higher accuracy scores.

5 OUR APPROACH

To enable law enforcement authorities to efficiently asses huge quantities of images found on a computer, we propose to utilize trained deep learning networks as a filter or selection tool. In cases of child pornography possession, hundreds of thousands or millions of images might be found on confiscated storage devices, while only a little fraction contains relevant content. Without any pre-processing of the image set, this makes it necessary to screen the complete set which is a cumbersome and time-consuming task, requiring the human investigator to stay concentrated during the whole screening process. Consider a hard drive disk containing 1 million image files, where 99.9% of the images are harmless, like holiday or kitten pictures, and only 0.1% are child pornography, which makes one out of a thousand images. If those relevant images are stored at some unfavorable place on the disk and images are inspected in the order in which they are retrieved from the storage device, this means that in the worst case the investigator has to view 999,000 irrelevant images until the first case of child pornography is found.

Studies showed that it takes the human brain about 150 milliseconds [14] to recognize high-level content in an image, e. g. whether or not an animal is present in the picture. Since decision-making in the context of child pornography affords many more classification steps (detection of a human body, nudity, a certain context and age estimation) and additional latency for the investigator to finally hit a button, a realistic classification time of a human investigator during image screening can be assumed to be much higher.

For the following example we therefore assume an average reaction time of a human investigator of 0.5 seconds per image. Thus, using the 1 million image dataset, in the worst case when all relevant images appear at the end of the list, it would take 138.75 hours (or 5.7 days) of uninterrupted image screening until the first relevant image is detected. Assuming a 40-hour week, the investigator would need more than 17 work days for this task and would be blocked from all other duties.

Rearranging the image files based on their likelihood of containing child pornography would facilitate the inspection process by far, as those images which are presumably relevant appear at the beginning of this rearranged list and can be inspected first. To this end, we use the estimations of NSFW detection algorithms to tag each image with a label or score relevant for the investigation. An investigation regarding the possession of child pornography would benefit from automatically identifying images with erotic or pornographic content. Images that are tagged as porn would be presented to the investigator with higher priority than others. Thereby, the chance for the investigator to discover relevant images earlier in the dataset increases. This allows screening only a small fraction of the complete dataset, thus saving time and human resources. Even if legal regulations require the investigator to screen the complete dataset, positive cases can be identified much earlier and thus be prioritized.

Theoretically, any classification algorithm suited for nudity detection can be used to achieve such rearrangement of images by simply moving those images which are classified as *NSFW* to the top of the list. However, deep learning algorithms for NSFW detection perform remarkably well, and are more suitable than simple skin detection algorithms, due to the reasons mentioned in section 3. We therefore employ several deep learning NSFW detection solutions in our approach, in order to separate relevant from non-relevant images by rearrangement. Of course, this pre-processing step causes some time overhead in advance of the manual inspection process, but it is still low compared to a cumbersome manual inspection of hundreds of thousands or even millions of images. Parallelization and the use of powerful GPUs even allow diminishing this time overhead. Furthermore, algorithms executed on neuromorphic hardware as a cutting-edge technology, already achieve classification results near those on traditional hardware architectures, however, processing 1,200 to 2,600 images per second [3].

5.1 Usage of Deep Learning for Image Rearrangement

In order to build a rearrangement system for large image sets with only small fractions of child pornographic imagery we used different CNN classifiers trained for NSFW detection, which will be detailed in section 6. Their NSFW scores per image were used to create relevance rankings of images. First we tested each classifier on a dataset comprising 1,000 NSFW images and 1,000 SFW images. Then we sorted the images according to the obtained NSFW scores in descending order, such that those images with high NSFW scores are located at the top of the list. We then used the respective ROC curves to determine the optimal threshold for each classifier to discriminate between NSFW ans SFW images. Finally, we applied three strategies for image rearrangement:

1. A **random** shuffling of the dataset: this served as a baseline arrangement which can be considered the average arrangement of images in a real-world image screening scenario, without any pre-processing.
2. A **binary** ordering of the dataset: the dataset was split into two sets. Data points with scores greater or equal to the measured threshold for a certain classifier were put into set A (the set of predicted NSFW images). Data points with scores below the threshold were put into set B (the set of predicted SFW images).
3. A **ranked** list of the dataset by sorting the data points in descending order based on their assigned scores.

The evaluation of the three strategies in section 6 shows the advantage that our binary and ranked order strategies provide over a random arrangement of images.

[8] https://github.com/yahoo/open_nsfw
[9] http://illustration2vec.net/

(a) NSFW images

(b) SFW images

Figure 1: Representative example images of the NSFW and SFW classes of our collected dataset. Inappropriate body regions and associated eye areas are blacked here for modesty. The images were collected from Reddit.com.

5.2 Dataset

Because there are no suitable image datasets available that could be used in the context of NSFW detection, we generated our own dataset. To this end, we collected 2,000 images from the online platform Reddit.com. The dataset splits into two classes: NSFW and SFW images, each of them containing 1,000 samples. Those images were manually sleected from a larger amount of collected images. As every positive training image for NSFW classifiers must inevitably contain human bodies or body parts, we assume that the well-trained CNNs employed in our approach do not have a considerable false positive rate on negative test images that do not display humans. So, in order to provide a challenging test set, the SFW class of our test set comprised only images of one or more persons. Figure **??** illustrates some sample images from our collected dataset.

Using real child pornography in our dataset might allow more reliable results, however, due to legal and moral reasons the collection and usage of such datasets is not feasible. On our collected dataset we performed the classifier evaluation. For the evaluation of our rearrangement strategies, a dataset of feasible size was not available and could not easily be collected. Therefore, we simulated our approach on randomly generated datasets by extrapolating the measured NSFW scores from our 2,000 samples dataset to a size of 1 million images.

6 EVALUATION

The evaluation of our approach was performed in a two-stage manner: first, we evaluated the chosen NSFW classifiers on a test set in order to measure their performances and to determine optimal classification thresholds for the binary rearrangement strategy; second, we evaluated our rearrangement strategies using those classifiers on a simulated dataset to show the potential of our approach.

6.1 Classifier Evaluation on Test Set

For the evaluation of our approach we chose three NSFW classifiers based on CNN architectures: Yahoo, Clarifai and I2V. In order to identify the optimal classification threshold for each of the CNN classifiers, we applied them on our dataset of 2,000 sample images. Additionally, we examined two baseline algorithms: a random coin toss classifier and the nude.py classifier which was mentioned earlier in section 3. Figure 2 shows the ROC curves for all five classifiers. Both Yahoo and Clarifai exhibit an area under the curve (AUC) close to 1, which means that they perform nearly optimal. I2V performs still much better than the random coin toss classifier and nude.py. The ROC curve of nude.py reveals that, although it is actually a nudity detection algorithm, it has a very poor performance, even close to the random coin toss classifier. This observation strengthens our assumption that deep learning algorithms are much stronger than traditional nudity detection mechanisms. The concrete AUC values are provided in Table 1.

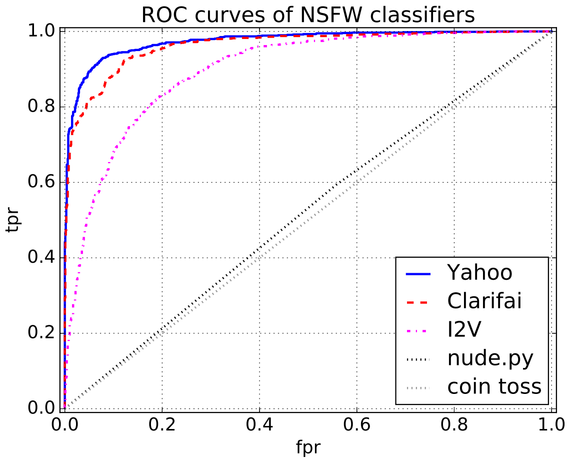


Figure 2: ROC curves of Yahoo, Clarifai and I2V compared to a random coin toss classifier on our 2,000 samples test set. The x-axis represents the false positive rate (fpr) and the y-axis the true positive rate (tpr).

Based on the ROC curves, we calculated Youden’s J statistic [17] which is defined by

$$\begin{aligned} J &= sensitivity + specificity - 1 \\ &= tpr - fpr, \end{aligned}$$

in order to get optimal thresholds for NSFW classification. The maximum of the J value determines the optimal threshold for each classifier. In addition to the AUC values, Table 1 shows the optimal thresholds for Clarifai, Yahoo and I2V together with the corresponding true positive rate (tpr) and false positive rate (fpr).

While the optimal thresholds for Yahoo and Clarifai are roughly in the middle between 0 and 1, it is interesting to note that I2V’s optimal threshold is close to 0. Interpreting I2V as a probabilistic classifier, this low threshold does not reflect a well calibrated prediction model. As nude.py uses a fixed threshold and only returns true/false labels, it does not allow threshold optimization.

Table 1: AUC, optimal thresholds and corresponding tpr and fpr for the evaluated NSFW classifiers.

Classifier	AUC	Threshold	tpr	fpr
Yahoo	0.975	0.384	0.928	0.076
Clarifai	0.963	0.682	0.922	0.121
I2V	0.896	0.090	0.826	0.190
nude.py	0.518	-	0.594	0.558
coin toss	0.500	-	0.500	0.500

6.2 Simulation of a Large Dataset

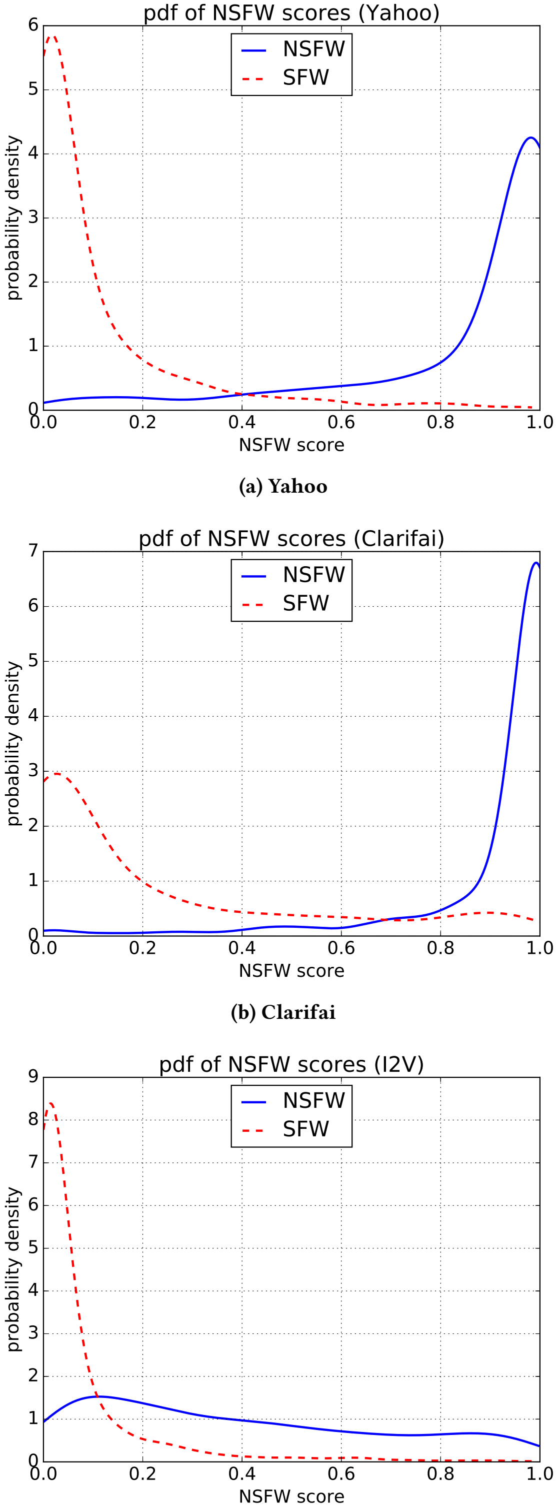
Due to limited resources we cannot test our approach on datasets of sizes, where manual inspection gets really cumbersome, such as hundreds of thousands or even millions of images. Nonetheless, we can evaluate the strength of our approach under realistic circumstances by extrapolating the measured results from section 6.1 on a simulated dataset of sufficient size. A preliminary sorting of images is especially helpful when the dataset is very large while the number of relevant images is relatively small. To simulate this, we generated a series of 100 sets, each comprising 1 million data points, while each data point represented an image. In each of the 100 datasets 1,000 of those data points were labelled as NSFW and the remaining 999,000 were labelled as SFW, i. e. 0.1% of the data points represented NSFW images.

First, for each classifier we calculated the probability density function (pdf) of predicted NSFW scores for both classes (NSFW and SFW) in our collected dataset containing 2,000 images, using a Gaussian kernel density estimation (kde). Using the pdf we generated NSFW scores for our simulated datasets. Figure **??** illustrates the pdfs of predicted scores by Yahoo, Clarifai and I2V. The scores predicted by I2V are distributed much more homogeneously than those by Yaho and Clarifai, which reflects I2V's lower performance on the discrimination of NSFW and SFW images.

The simulated datasets with scores assigned to every data point according to the pdfs served as a statistical model to predict the performance of our approach on massive datasets. We employed two rearrangement strategies for each of the three NSFW classifiers Yahoo, Clarifai and I2V to generate two image lists respectively: a binary ordering based on the previously measured NSFW classification threshold, and a ranked list according to pure NSFW scores. From each of the six resulting lists we successively withdrew data points, thus simulating a manual screening of the corresponding image dataset. Performing this step for each of the 100 simulated datasets and counting the number of retrieved NSFW data points at each step, we finally obtained an average performance graph for every combination of a classifier and rearrangement strategy.

This represents a classical information retrieval problem, for which many performance measures exist. However, in our scenario, depending on legal constraints for the prosecution of child pornography possession or production, demands on recall and precision may differ. We evaluated our approach using the following performance measures, which were also proposed by [8], due to the following reasons:

- **r@k**, recall after k images are retrieved: due to the fact that time and personnel are limited resources in criminal



(a) Yahoo

(b) Clarifai

(c) I2V

Figure 3: Probability density functions of predicted scores for the two classes NSFW and SFW on the 2,000 samples dataset.

investigations, and that the concentration of investigators decreases over time, only a limited amount of images can be screened at once.

- **Rank$_1$**, rank of the first relevant image: to determine whether or not there is child pornographic imagery on a given storage device is crucial, independent of the actual number of such images.

The evaluation results of the two different rearrangement strategies producing the binary and ranked list, on our simulated dataset of 1 million images can be found in Table 2 for each of the three NSFW classifiers separately. Additionally, a random sorting of the list is provided as a baseline, which corresponds to the average case of investigations without any pre-processing rearrangement. As one observes, ranking the images according to their predicted NSWF score yields a great advantage over a random or binary ordering of the dataset.

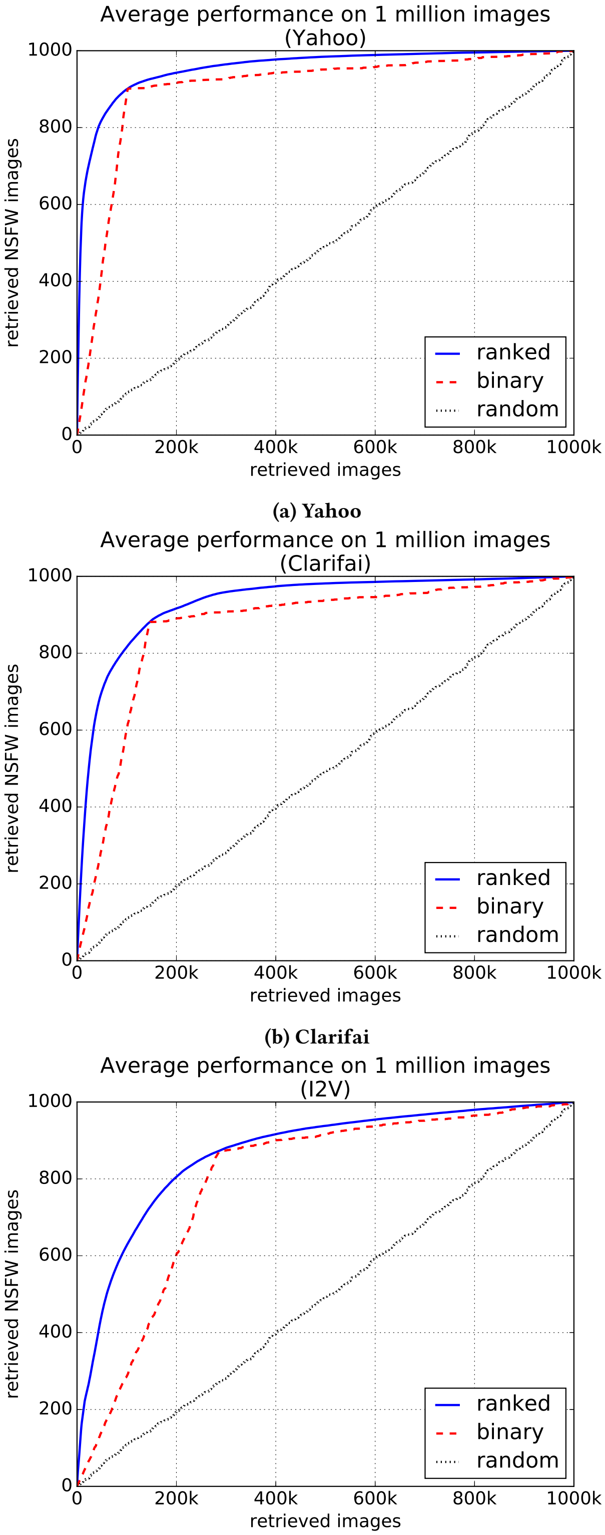
Using the ranked lists, the first NSFW image could be observed on average at ranks 8.74, 32.83 and 41.79 for Yahoo, Clarifai and I2V respectively. Using the binary sorting on the other hand, the first NSFW image could be observed much later, while in the randomly sorted lists, more than 1,400 images had to be inspected until the first NSFW image could be found. Also in terms of recall, the ranked list rearrangement strategy outperforms the binary and random sorting. After 5,000 inspected images from the randomly sorted list, the recall is just 0.5%. For the binary sorted lists it is already much higher (Yahoo: 3.97%, Clarifai: 2.83%, I2V: 1.42%). Using the ranked lists however, the recall increases by multiples (Yahoo: 38.51%, Clarifai: 13.86%, I2V: 9.18%).

It is interesting to note that the recall for the randomly sorted lists appears to be lower than one would actually expect, as pure probability would imply a r@100 of 0.0001 and a r@1000 of 0.001; r@5000 meets the expectancy. This can be explained by the nature of our experiment which involved just a limited number of simulations and is thus subject to deviations from pure probability. Alltogether, the ranking of images according to their NSFW score predicted by Yahoo and Clarifai reveals a huge advantage over all other sorting strategies.

If a larger loss of relevant images is tolerable, the performance of the binary rearrangement strategy can be increased by choosing a stricter threshold. If the threshold is picked high enough such that the scores of SFW images which led to false positive classifications before, are now below the threshold, they are classified correctly as SFW. This lets both actual NSFW and SFW images drop out of the set A of predicted NSFW images. However, assuming that NSFW images generally have a higher NSFW score, more SFW images move to set B, which increases the chance to discover actual NSFW images in set A.

In order to provide an overview of the performance of the different sorting strategies and classifiers, figure 4 shows the average retrieval graphs for the different lists and classifiers.

We also want to note that we did not take speed measurements at this point. This can become relevant when huge data sets are processed and the time between automated sorting of the images and human-based screening is of importance. Processing speed depends on the classifiers and their integration in the overall system.



(a) Yahoo

(b) Clarifai

(c) I2V

Figure 4: Number of retrieved NSFW images plotted against the total number of retrieved images, averaged over 100 simulated datasets for Yahoo, Clarifai and I2V.

Table 2: Average performance measured on 100 randomly generated datasets with 1 million images containing 1,000 NSFW images.

Rearrangement strategy	Rank$_1$	r@100	r@1000	r@5000
random	1463.00	0.0000	0.0000	0.0050
binary (Yahoo)	144.39	0.0009	0.0071	0.0397
ranked (Yahoo)	8.74	0.0109	0.0977	0.3851
binary (Clarifai)	169.41	0.0006	0.0294	0.0283
ranked (Clarifai)	32.83	0.0031	0.0608	0.1386
binary (I2V)	324.65	0.0003	0.0030	0.0142
ranked (I2V)	41.79	0.0024	0.0234	0.0918

Speed optimization is easily done as distributed parallel processing, e.g. by a Hadoop cluster is not a challenge.

7 FUTURE WORK

In the near future the collection of large image datasets for NSFW detection would enable us to confirm our observations made on the simulated datasets in this paper. Also, as we actually intend to provide a means for easier detection of child pornography, our approach should be applied on imagery collected from real cases. This can only be done by law enforcement authorities or authorized insitutions who might have access to such imagery at large scale, as the collection of child pornographic imagery itself is generally illegal.

Our approach can also be extended to distinguish between multiple classes of image content in order to achieve a fine-grained separation of pornographic content depending on severity. Instead of ranking images solely according to their NSFW scores it is possible to assign multiple scores, e. g. for primary and secondary sexual characteristics or explicit sexual activities. Especially I2V provides a broad classification scheme which could be used to this end.

A combination of our approach with face and scene detection algorithms would additionally ease the identification of suspects and locations of child pornography production. For a better separation of child pornography and legal adult pornography, a combination with age classification algorithms is also conceivable. In future work, it should be investigated whether the combination with age classification algorithms allows achieving image rankings where child pornographic images are assigned higher scores than adult pornography.

As deep learning can be applied on various data classification tasks, our approach is transferable to other scenarios where large amounts of data have to be inspected manually and time and resources are limited, such as the discovery and deletion of fake news and hate speech [2] in social media networks.

8 SUMMARY AND CONCLUSION

In this paper we investigated the benefit of deep learning algorithms for the manual inspection of large image sets in child pornography investigations by a binary and relevance-based rearrangement of images. We showed that neural networks for NSFW classification can be employed for sorting millions of images by relevance in order to aid human inspectors in discovering pornographic content. Although a manual screening of images in cases of child pornography possession is indispensable, our ranking-based approach allows saving large amounts of time when the dataset is large and only a small fraction is relevant. The first relevant image within a dataset of 1 million samples could be discovered at rank 8 or 9 on average, using our ranking-based strategy, regardless of the choice of NSFW classifier. Approximately 10% of the pornographic images could be found within the top 1,000 images (0.1% of the whole dataset) when using the ranking-based strategy with the Clarifai NSFW classifier.

However, we point out that due to legal reasons we could not evaluate our approach on real child pornographic imagery. Therefore, we evaluated it on the more general case of NSFW detection. Thus, the human inspector still has to discriminate between legal pornography and actual child pornography. Nevertheless, our approach provides an auxilliary tool to aid investigators

We are aware of the fact that in criminal investigations related to child pornography the occurrence of a false negative (i. e. when there is actual child pornography present on a storage device but is not detected by the classifier/investigator) might have serious effects and should be avoided. To avoid the possibility of missing relevant content, our approach does not truncate the dataset in any way. Instead, a prioritization of images takes place. So, if legal constraints require human inspectors to view the complete dataset, this is still possible with our approach. But even in this case our approach is still advantageous since relevant content can be found earlier, allowing to spur further actions and to avoid overlooking relevant images due to a loss of concentration.

At the end, we see our approach as an alternative or compromise between blacklist-based automated screening with its inability to deal with unknown content and unassisted human-based screening.

REFERENCES

[1] Rigan Ap-Apid. 2005. An Algorithm for Nudity Detection. In *PCSC'05*. Computing Society of the Philippines, Manila, PHL.

[2] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *WWW '17 Companion*. International World Wide Web Conferences Steering Committee, Geneva, CH, 759–760.

[3] Steven K. Esser, Paul A. Merolla, John V. Arthur, Andrew S. Cassidy, Rathinakumar Appuswamy, Alexander Andreopoulos, David J. Berg, Jeffrey L. McKinstry, Timothy Melano, Davis R. Barch, Carmelo di Nolfo, Pallab Datta, Arnon Amir, Brian Taba, Myron D. Flickner, and Dharmendra S. Modha. Convolutional Networks for Fast, Energy-Efficient Neuromorphic Computing. In *PNAS*, Vol. 113. Issue 41.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Washington, DC, USA, 1026–1034.

[5] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. 2007. A survey of skin-color modeling and detection methods. In *Pattern Recognition*, Vol. 40. 1106–1122.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc., New York, 1097–1105.

[7] Fabio Marturana and Simone Tacconi. 2013. A Machine Learning-based Triage methodology for automated categorization of digital media. In *Digital Investigation*, Vol. 10. Elsevier, New York, NY, USA, 193–204.

[8] Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet, and Thierry Pun. 2001. Performance Evaluation in Content-based Image Retrieval: Overview and Proposals. In *Pattern Recognition Letters*, Vol. 22. Elsevier, New York, NY, USA, 593–601.

[9] Christian Platzer, Martin Stuetz, and Martina Lindorfer. 2014. Skin Sheriff: A Machine Learning Solution for Detecting Explicit Images. In *SFCS '14*. 45–56.

[10] Siddharth Roheda and Hari Kalva. 2017. A Multi-Scale Approach to Skin Pixel Detection. In *Journal of Electronic Imaging*. Society for Imaging Science and Technology, Springfield, VA, USA, 18–23.

[11] Masaki Saito and Yusuke Matsui. 2015. Illustration2Vec: A Semantic Vector Representation of Illustrations. In *SIGGRAPH Asia 2015 Technical Briefs*. ACM, New York, NY, USA, 5:1–5:4.

[12] Clayton Santos, Eulanda M. dos Santos, and Eduardo Souto. 2012. Nudity Detection Based on Image Zoning. In *ISSPA’12*. 1098–1103.

[13] Jin S Seo, Jaap Haitsma, Ton Kalker, and Chang D Yoo. 2004. A robust image fingerprinting system using the Radon transform. *Signal Processing: Image Communication* 19, 4 (2004), 325–339.

[14] Simon Thorpe, Denis Fize, and Catherine Marlot. 1996. Speed of processing in the human visual system. In *Nature*, Vol. 381. Macmillan Publishers, London, UK, 520–522.

[15] Adrian Ulges and Armin Stahl. 2011. Automatic Detection of Child Pornography Using Color Visual Words. In *ICME’11*. ACM, Washington, DC, USA, 1–6.

[16] Ramarathnam Venkatesan, S-M Koon, Mariusz H Jakubowski, and Pierre Moulin. 2000. Robust image hashing. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, Vol. 3. IEEE, New York, NY, USA, 664–666.

[17] William John Youden. 1950. Index for Rating Diagnostic Tests. In *Cancer*, Vol. 3. Wiley, New Jersey, USA, 32–35.