Journal of Visual Communication and Image Representation 50 (2018) 303–313

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

# Leveraging deep neural networks to fight child pornography in the age of social media☆

Paulo Vitorino[a,b], Sandra Avila[c,*], Mauricio Perez[d], Anderson Rocha[c,*]

[a] Department of Electrical Engineering, University of Brasilia, Brazil
[b] Brazilian Federal Police, Brazil
[c] Institute of Computing, University of Campinas, Brazil
[d] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

*Keywords:*
Child pornography
SEIC content
Deep learning
Transfer learning
Fine tuning

ABSTRACT

Over the past two decades, the nature of child pornography in terms of generation, distribution and possession of images drastically changed, evolving from basically covert and offline exchanges of content to a massive network of contacts and data sharing. Nowadays, the internet has become not only a transmission channel but, probably, a child pornography enabling factor by itself. As a consequence, most countries worldwide consider a crime to take, or permit to be taken, to store or to distribute images or videos depicting any child pornography grammar. But before action can even be taken, we must detect the very existence or presence of sexually exploitative imagery of children when gleaning over vast troves of data. With this backdrop, veering away from virtually all off-the-shelf solutions and existing methods in the literature, in this work, we leverage cutting-edge data-driven concepts and deep convolutional neural networks (CNNs) to harness enough characterization aspects from a wide range of images and point out the presence of child pornography content in an image. We explore different transfer-learning strategies for CNN modeling. CNNs are first trained with problems for which we can gather more training examples and upon which there are no serious concerns regarding collection and storage and then fine-tuned with data from the target problem of interest. The learned networks outperform different existing solutions and seem to represent an important step forward when dealing with child pornography content detection. The proposed solutions are encapsulated in a sandbox virtual machine ready for deployment by experts and practitioners. Experimental results with tens of thousands of real cases show the effectiveness of the proposed methods.

## 1. Introduction

It is no wonder that virtually all modern societies condemn the sexual abuse of children. In the past two decades, awareness and recognition of such problems have grown systematically and now permeate the forefront of discussions in several governments, media outlets and society circles [1]. Unfortunately, some aspects of sexual abuse of children still lag behind in terms of public policies and immediate actions of eradication. It was only recently that child pornography started to be cast as a significant element in the lineup of activities related to sexual abuse. According to Taylor and Quayle [1], although child pornography has been a recognized problem for decades, it was until recently deemed as a "rather small and essentially specialist correlate of a much broader and more significant problem". Nonetheless, some recent studies report some staggering projections pointing out that every forth girl and sixth boy in the U.S. alone will experience some form of sexual abuse before turning 18 [2]. Equally alarming, we must also be aware that it is very likely that a significant fraction of these cases will be video taped for further distribution and sharing in online platforms and social networks.

In recent years, new communication and computing advancements along with the rise of social networks have prompted unremarkable societal advances in our world. However, at the same time, these advancements have also brought dishonest elements of our society closer to us. In this vein, as Taylor and Quayle [1] properly put it, since the mid-1990s, we have seen a significant change in the nature of child pornography in terms of generation, distribution and possession of images. Until some years ago, child pornography was mostly done offline and, thus, had less impact and it was more easily traceable. However, in the last few years, it evolved to a much more difficult

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding authors.
*E-mail addresses:* sandra@ic.unicamp.br (S. Avila), anderson.rocha@ic.unicamp.br (A. Rocha).

https://doi.org/10.1016/j.jvcir.2017.12.005
Received 16 May 2017; Received in revised form 22 September 2017; Accepted 16 December 2017
Available online 19 December 2017

problem with the advent of social networks, transforming the internet not only in a transmission channel but, probably, in a child pornography enabling factor by itself [1,3].

The problem turns even more complicated when we analyze it under a behavioral optics [4], with which we seek to learn as much as possible about the perpetrators, victims and the dynamics of an offense. Under this vantage point, pedophiles[1] now not only share contents online but also organize themselves in their own social networks sharing interests and experiences, blurring the edge between the virtual and real worlds [6]. Moreover, child pornography offenses seem not to have any specific boundary of class, income or profession [1]. Although controversial, some of these studies also suggest that child pornography offending is an indicator of pedophilia [7,8]. Finally, with more children also having access to uncontrolled materials and uncensored "friendships" online, they are also more easily approachable and encouraged to engage in dubious relationships with offenders. The Tech Innovation to Fight Child Sexual Exploitation (THORNE) foundation reports that up to 42% of "sextortion" victims met perpetrators online [9].

According to the International Criminal Police Organization (Interpol) [10,11], child pornography is defined as "…the consequence of the exploitation or sexual abuse perpetrated against a child. It can be defined as any means of depicting or promoting sexual abuse of a child, including print and/or audio, centered on sex acts or the genital organs of children." As a consequence, most countries worldwide consider a crime to take, or permit to be taken, to store or to distribute images or videos depicting any child pornography grammar [1,3]. According to Taylor et al. [12], in spite of the actual scene depicted in an image or video, whenever "an image of a child is accessed for a sexual purpose, it victimizes the individual concerned". With this backdrop, it is paramount that we devise and deploy proper mechanisms (technical and legal) to combat child pornography online. In this vein, in this paper, we aim at the automatic detection of child pornography from images. For a proper nomenclature, whenever we refer to images depicting child pornography content, we adopt the term *sexually exploitative imagery of children* (SEIC) [4].

In recent years, some researchers took aim at this problem by proposing a diverse set of solutions in the literature. As we shall discuss in Section 2, solutions range from nudity detection [13] and facial analytics [14,15] as proxies for child pornography classification to bags of visual words [16,17] and behavioral analytics [4] to network profiling [4,18–21] and sensitive hashing techniques [14,22,23]. Departing from virtually all existing methods, in this work, we leverage data-driven techniques and deep convolutional neural networks (CNNs) to harness enough characterization aspects from a wide range of images and point out the presence of child pornography content in an image. We propose a two-tiered CNN modeling, first trained with the source-related problem of general pornography detection – for which we can gather more training examples and upon which there are no serious concerns regarding collection and storage – and then fine-tuned, in a second refinement stage, with child pornography concepts properly controlled by a government agent in a secured setup. The learned network outperforms different existing solutions and seems to represent a major leap forward when it comes to dealing with this complex problem.

In a nutshell, our contributions in this paper are threefold:

- We introduce data-driven solutions able to distinguish sexually exploitative imagery of children from adult (related to normal porn) and seemingly innocuous (related to everyday imagery) content. The methods can pinpoint images depicting assault, gross assault and sadistic/bestiality involving children, which are considered to be of high importance in an international scale for combating child pornography (see Section 2 for more details on the scale and Section 3 on the method).
- We also introduce an adult content detector able to detect adult content not necessarily involving children, including the ones depicting secretive photographs showing underwear/nakedness with sexual intent, intentional posing suggesting sexual content and erotic posing (intentional sexual or provocative poses), which are ranked as of medium importance in the same international scale for combating child pornography (see Section 2 for more details on the scale and Section 3 on the method).
- Finally, we produce a self-contained virtual machine with our solutions, free of cost, and ready for initial deployments by different law-enforcement agents and practitioners for combating SEIC content nowadays.

We organize the remaining of this paper into four sections. Section 2 presents works related to child pornography. Section 3 introduces our solution to detecting SEIC contents. Section 4 describes the used datasets and experimental setup while Section 5 presents the experiments and results comparing the proposed method with different counterparts in the literature and with some off-the-shelf solutions. Finally, Section 6 concludes the paper and sheds some light on future research directions.

## 2. Related work

The legal definition of child pornography does not capture the entire nuance of the problem [1,12]. In a study of online content at the Combating Paedophile Information Networks in Europe Center (COPINE), Taylor and Quayle [1,12] identified 10 categories of pictures that may be sexualized by an adult and created the so-called COPINE scale as Table 1 shows.

Establishing the COPINE scale is pivotal to define to which extent someone caught red-handed is involved with sexually exploitative imagery of children. It also sheds some light on the research directions we should take in order to deal with this challenging problem [2]. Clearly, focus should be given to detecting activities involving levels L8–L10. However, many techniques in the literature, as we describe next, focus on simple models that are only able to detect nudity, or levels L2 and L3 above. Moreover, studies such as the COPINE one can also help driving some appropriate legislation on the theme [3] and defining a typology of online child pornography offending [24], especially because the law, itself, might not be well defined [25].

As the public awareness about child pornography has increased over the past years, so has the interest of researchers in presenting effective methods to block or, at least, cope with the problem. On one hand, there have been efforts toward working on the server side and on the network itself by developing appropriate filtering and blocking tools [4,20,21]. On the other, researchers have been developing detection solutions exploiting the intrinsic distributed nature of the internet, thus focusing on the users. While the former group of methods might be effective by actively detecting suspicious content and users trying to have access to it, it raises serious questions regarding censorship and high false positive rates, oftentimes blocking otherwise legitimate content [26,27]. The second group, in turn, works as a passive solution in which law-enforcement agents can sift through large amounts of data looking for inappropriate content in apprehended materials or as an active filtering solution in which families can protect their children and beloved ones from accessing inappropriate content by installing such solutions in their computing devices.

Spearheading the first group, we can see the works of [18,19], which focus on developing network profiling techniques to pinpoint abnormal behavior linked to child pornography. In the same vein, some researchers focused on specific active analyses of peer-to-peer networks [4,20,21], by and large, one of the most important channels for exchange of SEIC content online.

Veering away from the network profiling models, we have the

[1] According to Dorland [5], pedophilia refers to an abnormal fondness for children; sexual activity of adults with children.