Technology and Innovation, Vol. 19, pp. 593-599, 2018
Printed in the USA. All rights reserved.
Copyright © 2018 National Academy of Inventors.

ISSN 1949-8241 • E-ISSN 1949-825X
http://dx.doi.org/10.21300/19.3.2018.593
www.technologyandinnovation.org

REINING IN ONLINE ABUSES

Hany Farid

Dartmouth College, Hanover, NH, USA

Online platforms today are being used in deplorably diverse ways: recruiting and radicalizing terrorists; exploiting children; buying and selling illegal weapons and underage prostitutes; bullying, stalking, and trolling on social media; distributing revenge porn; stealing personal and financial data; propagating fake and hateful news; and more. Technology companies have been and continue to be frustratingly slow in responding to these real threats with real consequences. I advocate for the development and deployment of new technologies that allow for the free flow of ideas while reining in abuses. As a case study, I will describe the development and deployment of two such technologies—photoDNA and eGlyph—that are currently being used in the global fight against child exploitation and extremism.

Key words: Social media; Child protection; Counter-extremism

INTRODUCTION

Here are some sobering statistics: In 2016, the National Center for Missing and Exploited Children (NCMEC) received 8,000,000 reports of child pornography (CP), 460,000 reports of missing children, and 220,000 reports of sexual exploitation. Moreover, NCMEC reports a 1000% increase in sex trafficking over the past five years, and 12 is the average age of a child involved in sex trafficking and CP. These are deeply troubling numbers particularly when you consider that these are primarily U.S.-based statistics and the U.S. accounts for only 5% of the world's population. While all of these numbers are troubling, I would like to focus on the 8,000,000 reports of CP that NCMEC received last year.

It is helpful to look at the historical record to understand how we arrived at such a staggering number of CP reports. In the early 1980s, it was illegal in New York State for an individual to "promote any performance which includes sexual conduct by a child less than sixteen years of age." In 1982, Paul Ferber was charged under this law with selling material that depicted underage children involved in sexual acts. After having been found guilty under the New York State obscenity laws, Ferber appealed and the New York Court of Appeals overturned the conviction, finding that the obscenity law violated the First Amendment of the U.S. Constitution. The U.S. Supreme Court, however, reversed the appeal, finding that the New York State obscenity law was constitutional (1). Among several reasons for their ruling, the Supreme Court found that the government has a compelling interest in preventing the sexual exploitation of children and that this interest outweighs any speech protections.

Accepted: October 15, 2017.

Address correspondence to Hany Farid, 6211 Sudikoff Lab, Dartmouth College, Hanover, NH 03755, USA. Tel: +1 (603) 646-2761. E-mail: farid@dartmouth.edu

The landmark case of New York v. Ferber made it illegal to create, distribute, or possess CP. The result of this ruling, along with significant law enforcement efforts, was effective, and by the mid-1990s, child pornography was, according to NCMEC, largely a "solved problem." By the early 2000s, the rise of the internet brought with it an explosion in the global distribution of CP. Alarmed by this growth, in 2003, Attorney General Ashcroft convened executives from the top technology firms to ask them to propose a solution to eliminate this harmful content from their networks. Between 2003 and 2008, despite continued pressure from the attorney general's office, these technology companies did nothing to address the ever-growing problem of their online platforms being used to distribute a staggering amount of CP with increasingly violent acts on increasingly younger children (as young, in some cases, as a only a few months old).

In 2008, Microsoft and NCMEC invited me to attend a yearly meeting of a dozen or so technology companies to provide insight into why, after five years, there was no solution to the growing and troubling spread of CP online. This meeting led me on a nearly decade-long journey to develop and deploy technology to curb harmful online speech. Along the way, I learned many lessons about how to develop and deploy technology at internet scale, as well as learning about public and media relations, corporate indifference, and the horrific things that are being done online and offline to some of the most vulnerable in our society. I will share some of these insights along with some technical details of the technology that we developed.

COUNTERING CHILD EXPLOITATION

At the first of what would be many meetings on this topic, I listened to several hours of discussion on the scope and scale of the problem of online child exploitation. I heard why various technological solutions did not or would not work, and I heard many lawyers talk about liability, profits, and user privacy. Around midday, I was asked to share my thoughts. I started with a simple question: Just out of curiosity, how many of you are engineers, mathematicians, or computer scientists? One or two hands shot up, out of a room of 60 or so people. I then asked how many were lawyers. More than half of the remaining hands shot up. I don't recall if I said this out loud or not, but I certainly thought, "Well, there is at least part of your problem. It is difficult to get things done when the lawyers outnumber the scientists and engineers."

Throughout the day of that first meeting, I repeatedly heard that it is incredibly difficult to automatically and efficiently scrub CP from online platforms without interfering with the business interests of the titans of tech represented in the room. Among several challenges, managing the massive volume of data uploaded every day to social media platforms was of particular concern. My second question to the group was, therefore, "Specifically, how hard is the problem?" Here are the numbers that all the attendees agreed upon. Any technology must satisfy the following requirements:

1. Analyze an image in under two milliseconds (500 images/second)
2. Misclassify an image as CP at a rate of no more than one in 50 billion
3. Correctly classify an image as CP at a rate of no less than 99%
4. Do not extract or share any identifiable image content (because of the sensitive nature of CP)

Developing a fully automatic algorithm to distinguish CP from other content with these engineering demands was, in my opinion, not feasible. It was not feasible in 2008 when we started to work on this problem, and I would argue that it is not feasible today despite all of the advances in machine learning and computer vision in the intervening years.

I was ready to concede that a solution was not possible until I heard NCMEC's then-CEO Ernie Allen mention two interesting facts: 1) NCMEC is home to millions of known CP images that have been manually reviewed and determined to contain explicit sexual contact with a minor (in many cases, under the age of 12) and 2) These same images are continually distributed for years and even decades after they are first reported to NCMEC. I thought that even if we did not have the technological innovation to fully distinguish CP from other content, we could perhaps stop the redistribution of known CP content instead. While this would not address the problem in its entirety, surely it would, given what we know,

Figure 1. The MD5 hash of this image is 78ba217bccd6e6b4d032e54213006928.

be a first step to disrupting the global distribution of CP.

In collaboration with NCMEC and researchers at Microsoft, we set out to develop technology that could quickly and reliably identify images from the NCMEC database of known CP images. At first glance, this may seem like an easy problem to solve. Hard-hashing algorithms such as MD5 or SHA-1 (2,3) can be used to extract from an image a unique compact alphanumeric signature (Figure 1). This signature can then be compared against all uploads to an online service like Facebook or Twitter. In practice, however, this type of hard hash would not work because most online services automatically modify all uploaded images. Facebook, for example, resizes, recompresses, and strips metadata from every image. The result of these and similar modifications is that, although the original and modified images are perceptually similar, the signature (hash) is completely different. The reason is that hard hashing is designed to yield distinct signatures in light of any modification to the underlying image. Hard hashing, therefore, is ineffective at matching images that are modified in any way at the time of upload.

At a conceptual level, however, hashing has many desirable properties: A signature is computationally efficient to extract; the signature is unique and compact; and hashing completely sidesteps the difficult task of content-based image analysis that would be needed to recognize the presence of a person, determine the person's age, and recognize the difficult-to-define concept of sexually explicit. Building on the basic framework of hard hashing, we sought to develop a robust hashing algorithm that generates a compact and distinct signature that is stable to simple modifications to an image, such as re-compression, resizing, color changes, and annotated text.

Although I will not go into too much detail on the algorithmic specifics, I will provide a broad overview of the robust hashing algorithm—named PhotoDNA—that we developed (see also (4,5)). Shown in Figure 2 is an overview of the basic steps involved in extracting a robust hash. First, a full-resolution color image is converted to grayscale and downsized to a lower and fixed resolution of 400 × 400 pixels. This step reduces the processing complexity in subsequent steps, makes the robust hash invariant to image resolution, and eliminates high-frequency differences that

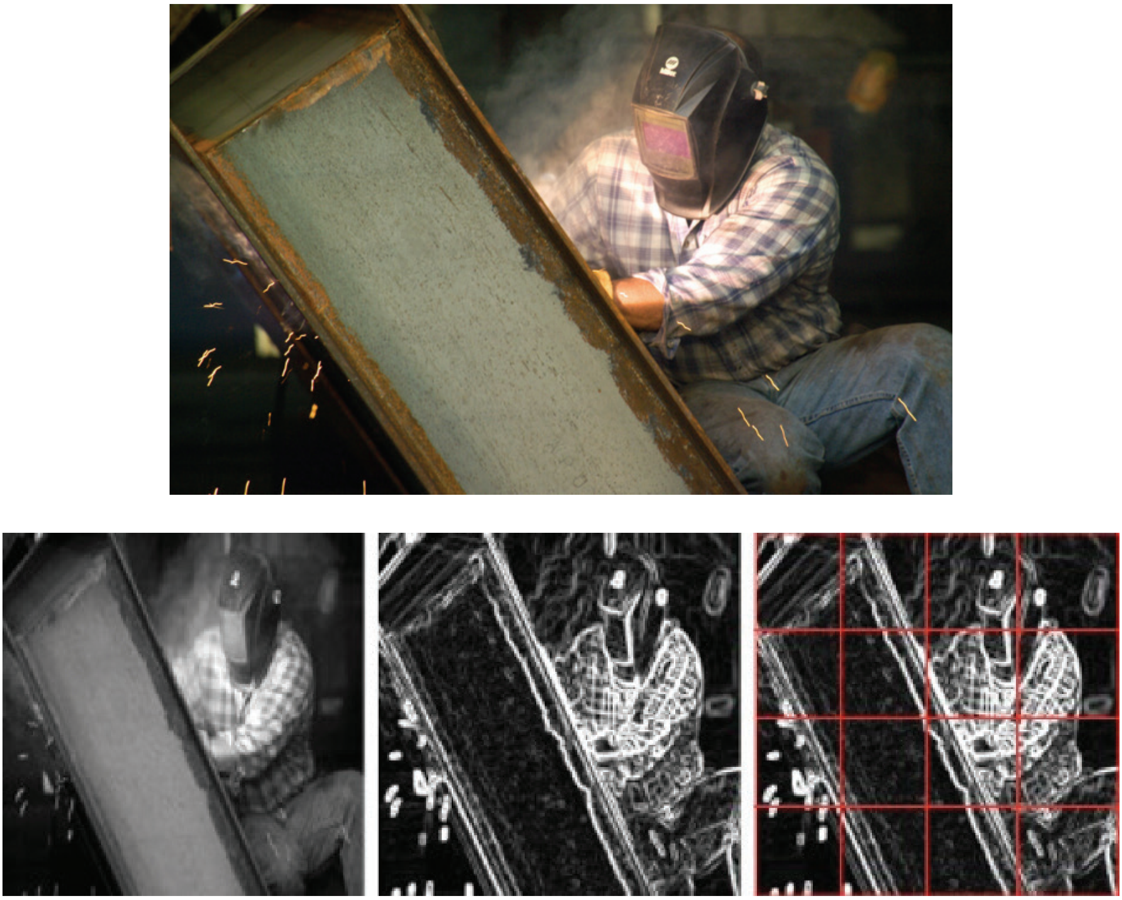

Figure 2. The three basic processing steps of photoDNA: 1) convert a full-resolution color image (top) to grayscale and lower resolution (bottom left); 2) use a high-pass filter to highlight salient image features (bottom center); and 3) partition the high-pass image into quadrants from which basic statistical measurements are extracted to form the photoDNA hash (bottom right).

may result from compression artifacts. Next, a high-pass filter is applied to the reduced resolution image to highlight the most informative parts of the image. Then, the image is partitioned into non-overlapping quadrants from which basic statistical measurements of the underlying content are extracted and packed into a feature vector. Finally, we compute the similarity of two hashes as the Euclidean distance between two feature vectors, with distances below a specified threshold qualifying as a match. Despite its simplicity, this robust-hashing algorithm has proved to be highly accurate and computationally efficient to calculate.

After a year and a half of development and testing, photoDNA was launched in 2009 on Microsoft's SkyDrive and search engine Bing. In 2010, Facebook deployed photoDNA on their entire network. In 2011, Twitter followed suit, while Google waited until 2016 to deploy. In addition to these titans of technology, photoDNA is now in worldwide deployment. In 2016, with an NCMEC-supplied database of approximately 80,000 images, photoDNA was responsible for removing over 10,000,000 CP images, without any disputed take-downs. This database could just as easily be three orders of magnitude bigger, giving you a sense of the massive scale of the global production and distribution of CP.

Child exploitation is, of course, not the only harmful content online. The internet has been a boon for extremist groups, cybercriminals, and trolls. Since 2015, I have been thinking about how technology like photoDNA can be deployed to mitigate some of the damage caused by these individuals and groups.

COUNTERING ONLINE EXTREMISM

Over the past few years, our world leaders have expressed grave concern about how extremist groups have harnessed the power of the internet to spread hate and violence. In 2015, President Obama said, "The high-quality videos, the online magazines, the use of social media, terrorist Twitter accounts—it's all designed to target today's young people online, in cyberspace."

And in 2017, Prime Minister May's office said, "The fight against terrorism and hate speech has to be a joint one. The government and security services are doing everything they can and it is clear that social media companies can and must do more."

Since 2015, I have been working with the Counter Extremism Project (a non-governmental organization) to develop the next generation of robust hashing technologies with the goal of eliminating the worst-of-the-worst extremism-related content, including content with explicit violence, explicit calls to violence, and glorification of violence (each of which are violations of most terms of service—more on this in the next section).

Conceptually, eliminating extremism-related material can follow a similar model as eliminating CP: Build a database of known harmful content, extract a hash from each piece of material, and automatically screen all uploaded material against a database of hashes. Because extremism-related material tends to come in the form of audio and video recordings, we had to generalize the image-based robust hashing to be applicable to videos and audios. Although I will not go into too much detail on the algorithmic specifics, I will provide a broad overview of the multimedia robust hashing algorithm—named eGlyph—that we have developed.

The largest challenge with analyzing video is the massive amount of data in even a short video: At 24 frames per second, a three-minute video contains 4,320 still images. At even a modest resolution of 640 × 480 pixels per frame, a three-minute video contains over 1.3 billion pixels. The complexity of hashing a video, as compared to analyzing a single image, is at least three orders of magnitude larger. There are, however, typically only small changes between successive frames of a video leading to a large amount of information redundancy in a video. We can, therefore, reduce the complexity of analyzing a video by first reducing this redundancy.

We, conveniently, just described a mechanism for measuring the similarity between two images—photoDNA. In addition to finding nearly identical images, robust hashing can be used to find similar images by controlling the threshold on the Euclidean metric for image similarity (as described in the previous section). We start a video analysis by using robust image hashing to eliminate redundant video frames (this variant is a modified version of photoDNA that is slightly more computationally efficient and yields slightly more compact hashes). This elimination of redundant frames typically reduces the length of a video by approximately 75%. The image hash is then extracted from each of the remaining frames and concatenated to yield a final video hash. Unlike the image-based hashing that yields a fixed length hash, a video hash can be of arbitrary length. This presents both a challenge and an opportunity for comparing two hashes.

A Euclidean distance cannot, of course, be used to compare two hashes of arbitrary length. Instead, we utilize the longest common substring (LCS) (not to be confused with the longest common sub-sequence algorithm) (6). By way of intuition, the LCS of the two strings "ABABACABBC" and "ABACABACBBCA" is six because the longest common string shared by these strings is "ABACAB." Note that these strings also have the substring "BBC" in common, but this is shorter than the substring of length six. Given two strings of length *m* and *n*, the LCS can be found efficiently using dynamic programming with a run-time complexity of $O(mn)$. The advantage of using LCS to compare two hashes is that it allows us to find not just matching videos but also video segments that are extracted or video segments that are embedded within a larger video (e.g., a video compilation). Running on a standard Linux machine, a Java-based implementation of this robust video hashing requires approximately 10 ms to process a single video frame and approximately 2.5 ms to compare two hashes. To improve the efficiency, we have implemented a multi-core version of this algorithm that allows for a video to be partitioned into an arbitrary number of short segments, each of which can be analyzed on a separate computer core. The individual results from each segment are then combined to create a single hash. With this approach, the rate-limiting step to analyze any video is simply the number of computing cores that are available.

DISCUSSION

The First Amendment of the U.S. Constitution reads as follows: "Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances."

Facebook's terms of service, which all users agree to, reads in part: "You will not post content that: is hate speech, threatening, or pornographic; incites violence; or contains nudity or graphic or gratuitous violence. We can remove any content or information you post on Facebook if we believe that it violates this Statement or our policies."

Because the First Amendment states that *Congress* "shall make no law…," the restrictions imposed by Facebook (and virtually all online services) are not at odds with our Constitution. This does not mean, however, that we should not discuss the implications of technology like photoDNA and eGlyph in the context of encouraging and supporting an open and free internet.

The most common question that we have received during the development and deployment of photoDNA and eGlyph is, "Who will decide what is and what is not CP and extremism-related material?," while the most common criticism has been, "This technology will eventually be misused to restrict protected speech, political dissent, or unpopular ideas." These are legitimate questions and concerns worthy of debate.

Although it may seem that the definition of CP should be straightforward, the federal statute is anything but (7):

> Images of child pornography are not protected under First Amendment rights, and are illegal contraband under federal law. Section 2256 of Title 18, United States Code, defines child pornography as any visual depiction of sexually explicit conduct involving a minor (someone under 18 years of age).

While a person's age is generally straightforward to determine, this determination is less straightforward if that determination is based on a single image. The definition of "sexually explicit" is also open to interpretation. When deploying photoDNA, we avoid the complexity of classifying content whose legality might be disputed by only adding content to the database that contains images of children under the age of 12 involved in an explicit sexual act. Because children under the age of 12 are typically prepubescent, there is no disagreement that the child is under the age of 18. And, because the images contain an explicit sexual act, there is no disagreement of the legal statute of "sexually explicit." This content—termed the worst-of-the-worst by former NCMEC CEO Ernie Allen—eliminates any ambiguity in the interpretation of the federal statute and ensures that photoDNA eliminates only clearly illegal content.

In the counter-extremism space, eGlyph faces similar challenges in classifying material. In building a database of extremism-related content, we want to avoid any content that might be considered political dissent or commentary or otherwise protected under a company's terms of service. Fortunately, Facebook, and most other internet-based terms of service, clearly specify that explicit violence or explicit calls to violence are forbidden. Following the model of eliminating the worst-of-the-worst, we populate the extremism-related database only with content that clearly and unambiguously falls into the categories of explicit violence or explicit calls to violence.

I take solace from the fact that some have argued that this conservative approach to defining CP and extremism-related content is not aggressive enough, while others have argued that it is too aggressive. We should work to ensure an open and free internet that allows for an open exchange of ideas and for vigorous debate. At the same time, we must acknowledge the real harm that is resulting from certain types of content and do everything we can to eliminate this type of content from our online platforms.

It is important to understand that any technology such as that which we have developed and deployed can be misused. The underlying technology is agnostic as to what it searches for and removes. When deploying photoDNA and eGlyph, we have been exceedingly cautious to control its distribution through strict licensing arrangements. It is my hope and expectation that this technology will not be used to impinge on an open and free internet but to eliminate some of the worst and most heinous content online.

ACKNOWLEDGMENTS

Hany Farid is the Albert Bradley 1915 Third Century Professor and chair of computer science at Dartmouth College. His research focuses on digital forensics, image analysis, and human perception. He is the chief technology officer and co-founder of Fourandsix Technologies and a senior adviser to the Counter Extremism Project. This work was supported by funding from Microsoft Corporation and the Counter Extremism Project.

REFERENCES

1. New York v. Ferber. (U.S. Supreme Ct. 458 U.S. 747, 1982).
2. Rivest R. The MD5 message-digest algorithm. Internet RFC 1321; 1992. http://people.csail.mit.edu/rivest/Rivest-MD5.txt
3. Eastlake DE, Jones PE. US Secure Hash Algorithm 1 (SHA1). Internet RFC 3174; 2001.
4. Venkatesan R, Koon SM, Jakubowski MH, Moulin P. Robust image hashing. In: Proceedings 2000 International Conference on Image Processing. International Conference on Image Processing 2000; 2000 Sep 10-13; Vancouver. IEEE; 2000. 3:664–666.
5. Swaminathan A, Mao Y, Wu, M. Robust and secure image hashing. IEEE Trans Inf Forensic Secur. 2006;1(2):215–230.
6. Gusfield D. Algorithms on strings, trees and sequences: computer science and computational biology. New York (NY): Cambridge University Press; 1997.
7. Sexual Exploitation and Other Abuse Of Children Definitions for Chapter, Section 2256 of Title 18, United States Code.