Literature Review - Word count: 1922

James Hines

How can technology be used to raise awareness of child sexual abuse material between content platform providers.

Introduction

There are many positives born out of the ongoing development and improvement of technology and the internet, but there is also a dark side which is also utilizing the advances in online sharing. The dark side is child sexual abuse material (CSAM), which has seen an increase from 1 million reports of CSAM in 2010, to 17 million reports in 2019 (Negreiro, 2020). An analysis carried out in 2016 had discovered that 63.4% of CSAM content involved children under 8 years of age (Cybertip, 2016), with Thorn (2023) now stating that “there are approximately 563,461 reports of CSAM images per week” being reported.

Previously, there appeared to be several organizations who had been collecting perceptual hashes to be matched for detecting potential CSAM, but it had become obvious that a centralised system would be required. This is where Thorn developed and introduced an AI/ML perceptual hashing and matching database technology called Safer, containing more than 32 Million Known CSAM hashes which could be used for detecting known and unknown CSAM. According to the Internet Watch Foundation Report (2022), IWF themselves have successfully hashed 1,663,106 individual images since 2016.

The focus of this review is to raise awareness of perceptual hashing and its capability to identify CSAM on content platforms.

Understanding the hashing technology that is successfully detecting CSAM images online with the hope that offenders could also eventually be brought to justice, is of great significance for the children involved. The National Crime Agency has estimated that the United Kingdom has between 680,000 and 830,000 adult offenders (NCA, 2023), with the NCA (2023) stating that “1 in 10 children experience child sexual abuse before their 16th birthday”. If this review leads to alternative ideas that could potentially increase the usage of perceptual hashing to identify other areas related to CSAM activity, then that could be substantially significant in aiding and supporting investigations.

Perceptual Hashing

Perceptual hashing is a hash function that had been designed specifically for visual and audio data (Yuenan et al, 2018) and described as a technique of assigning a short binary string to an image, which then gives that particular image a personlised fingerprint (Mckeown & Buchanan, 2023; Struppek et al, 2022; Monga et al, 2004). According to an early study of perceptual hashing by Monga et al (2004), the binary string is based on how the image appears to the human eye and therefore matches with similar images. Shaik et al, (2022) have stated that “generating an image hash does not alter the image content” which is a key strength of perceptual hashing to ensure image

integrity. Once images have been confirmed as being CSAM content by a specially trained human analyst, the images are then saved to a central database and can be automatically cross-referenced to other online images that contain a matching hash (safer, 2022).

Despite perceptual hashing having many strengths which have helped, and continue to help identify CSAM , the technique has also demonstrated that it has weaknesses and limitations.

A strength of perceptual hashing is its robustness to defend against counter-identification techniques which can be used to obfuscate against detection, such as contrast adjustment, noise addition, and scaling (Biswas et al, 2019; Rahalkar & Virgaonkar, 2022). According to Shaik et al (2022) hamming distance is an approach used in calculating whether an image is identical to another or not.
Although there appears to be variations of perceptual hashing algorithms that focus on the distributions of hamming distances to test for accuracy in identifying matching images, a recent study by Mckeown & Buchanan (2023) found that the scaling algorithm used within perceptual hashing achieved 97.9% for exact matching of images.

However, even though Mckeown & Buchanan (2023) reports that scaling achieved 97.9% accuracy in exact image matching, there are several counter-identification techniques which have been reported to expose a weakness within perceptual hashing. Biswas et al, (2019) had stated that “rotating images, mirroring images, and cropping

images, have negative effects on identification performance". It has also become evident that these weaknesses have still not been mitigated against, because Mckeown & Buchanan (2023) reported these same weaknesses during their research of 250,000 images, four years later.

According to research carried out by Struppek et al (2022) that focused on exposing the weakness in perceptual hashing, it was found that the technique is vulnerable to data leakage, hash string bypass, and the potential framing of innocent users. Struppek et al (2022) discovered that it was possible for an adversary to create a fake image and assign it a perceptual hash which matches a hash within the CSAM database, the fake image can then be spread across many devices which could trigger false alerts for devices not containing potential CSAM and therefore, frame innocent users. But, according to Rahalkar & Virgaonkar (2022), users are unable to access the database that stores known CSAM images and their assigned perceptual hashes. Therefore, it seems plausible that the adversary would need to be an inside risk at the database center or closely associated. This, by itself, may partially mitigate the risk of an adversary successfully identifying hashed and verified CSAM content and copying the perceptual Hash.

A current challenge associated with the process of assigning a binary string during Perceptual hashing, is due to secure communications via end-to-end encryption (Negreiro, 2020). The challenge created by end -to-end encryption is that images can be shared and redistributed prior to being filtered and detected as CSAM (Negreiro,

2020). An idea discussed by Rahalkar & Virgaonkar (2022) in their paper titled ‘Content moderation schemes in end-to-end encrypted systems’ suggests the possibility of using perceptual hashing to identify misinformation by storing perceptual hashes in a database on the client side. If the misinformation is sent to another user, then the receiver is informed about the message being misinformation.

Considering the discussions within the recent paper by Rahalkar & Virgaonkar (2022), it may be possible that the challenge of identifying CSAM content in end-to-end encryption communication may be solved sooner than we think. That said, it appears that Whatsapp was able to generate 1,372,696 reports of CSAM in 2021 (NCMEC, 2023). This raises the question, if Negreiro (2020) was concerned about end-to-end encryption causing difficulties with being able to identify CSAM via hashing in 2020, and Rahalkar & Virgaonkar (2022) also discussed possible ways around end-to-end encryption in 2022, then how did Whatsapp manage to submit those reports in 2021 after whatsapp enabled end-to-end encryption in 2016?

A reason why this question may be of huge importance is because, being able to identify CSAM content, especially when end-to-end encryption has been deployed, has become a priority for the European Union (European Commission, 2020). The European commission suggests that the use of end-to-end encryption has made it almost impossible to identify perpetrators, and therefore, the EU have stated that they will make it a priority to develop tools that are able to detect CSAM (European

Commission, 2020) built on the same proven model as used by the National Center for Missing & Exploited Children.

Reports by safer (2023) claim that Electronic Service Providers (ESPs) had provided over 88.3 million files to the National Center for Missing and Exploited Children's CyberTipline in 2022. According to NCMEC (2023), ESPs had reported 65.46 Million in 2020, 84,99 Million in 2021 and 88.37 Million in 2022 with “Online Enticement of children for sexual acts” almost doubling from 44,155 thousand in 2021, to 80,524 in 2022.

However, although these numbers illustrate an increase of reported CSAM images being detected, it may not directly correlate to mean that CSAM had increased at that rate. The correlation of perceptually hashed CSAM being detected, could be a result of ESPs onboarding to implement detection and share hashes with the centralised database. NCMEC (2023) stated that “in 2020, 42 new ESPs registered with CyberTipline to send reports”. 2021 also saw 149 new ESPs register and a further 68 ESPs in 2022 (NCMEC, 2023).

This increase in reporting highlights the importance of perceptual hashing within the topic of: how can technology be used to raise awareness of child sexual abuse material (CSAM) between content platform providers.

Conclusion

To conclude, there are two areas that can be discussed for future research. Although some of these ideas may seem far-fetched at this moment in time, the advances in AI/ML and technology have made significant gains in the last decade and therefore, allow the following thoughts to become a reality. But first, we will briefly summarise the importance of perceptual hashing.

This literature review has identified the strength and importance of perceptual hashing in identifying and detecting CSAM content. This review has also highlighted some of the weaknesses and limitations associated with this form of hash matching. Although there are vulnerabilities with using perceptual hashing that appear to have been identified by Biswas et al, (2019) and still not presently mitigated, we can not deny the importance of this technique. 563,461 reports of CSAM content being discovered each week (Thorn, 2023), demonstrates and supports the importance of perceptual hashing.

There has been a lot of data referenced and shared in this review regarding the quantities of CSAM detected and identified. What has become apparent during this research is that the data is still very segmented. There appears to be more than one database for submitting and storing hashes for referencing. To reduce any confusion, a “Global CSAM database” could be established with an organisation such as the United Nations to ensure a global law that stipulates all media platforms must register if they wish to operate.

After reviewing the relevant research papers and related content, there are two ideas which could potentially be helpful in identifying and detecting associated CSAM activities.

Firstly, Looking towards the future, there could be an opportunity to combine the recent discovery and research findings by Alzayer & Zhang (2023) of radiance field reconstruction. These findings may allow for the use of eye reflections and the CSAM perceptual hashing imagery database to be combined for forensic investigations via automation and machine learning.

The paper by Alzayer & Zhang (2023) demonstrates that the reflection seen within the eye of a photographed individual can be used to reconstruct the environment. This combination may eventually be able to then correlate reflections in children's eyes with their surroundings to see if some of the abuse is at the same location for multiple victims as well as location discovery. Ideally, CSAM will be identified by perceptual hash matching, radiance field reconstruction will be automated to create a 3D image of the environment, the environment image will be hashed and that perceptual hash could crawl social media postings for a match.

Secondly, Europol often requires help in identifying clothing which either belongs to children who have been discovered in CSAM or belongs to the abusers. To date, 23 children have been identified and removed from harm due to identifying clothing (Europol, 2023). Perceptual hashes could play a role in identifying the garments of

clothing by ensuring that all clothing manufactures photograph and submit their products to a centralised manufactures database, each upload is then hashed and can be queried for hash matching. This may allow some garments to be geographically narrowed or identified to aid in investigations.

Not all children will have the possibility to look into the camera or to have their face visible, for those that do look into the camera or have their face visible, then there could be an increased opportunity to gain valuable forensic information for further analysis.

References

Alzayer, H., Zhang, K., Feng, B., Metzler, C. & Huang, J. (2023) Seeing the World through Your Eyes. *University of Maryland.* Available from: https://arxiv.org/pdf/2306.09348.pdf [Accessed on 9th July 2023]

Biswas, R., González-Castro, V., Fidalgo, E. & Alegre, E. (2020) Perceptual image hashing based on frequency dominant neighborhood structure applied to Tor domains recognition. *Neurocomputing* 383: 24-38. Available from: https://doi-org.uniessexlib.idm.oclc.org/10.1016/j.neucom.2019.11.065 [Accessed on 29th July 2023]

Cybertip. (2016) Child Sexual Abuse Images On The Internet. *Canadian Centre for Child Protection*. Available from: http://s3.documentcloud.org/documents/2699673/Cybertip-ca-CSAResearchReport-2016-En.pdf [Accessed on 7th July 2023]

IWF. (2022) Annual Report 2022. *Internet Watch Foundation Report*. Available from: https://annualreport2022.iwf.org.uk/ [Accessed on 7th July 2023]

Mckeown, S. & Buchanan, W,J. (2023) Hamming distributions of popular perceptual hashing techniques. *Forensic science international: Digital investigation* 44(301509). DOI: 10.1016/j.fsidi.2023.301509 [Accessed on 8th July 2023].

Monga, V., Banerjee, A., & Evans, B,L. (2004) Clustering algorithms for perceptual image hashing. *11th IEEE Signal processing education workshop.* Available from: https://ieeexplore-ieee-org.uniessexlib.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=1437959 [Accessed on 9th July 2023]

NCA. (2023) Extended Reality technologies, such as augmented and virtual reality, have been identified as evolving threats. *National Crime Agency*. Available from: https://nationalcrimeagency.gov.uk/nsa-child-sexual-abuse [Accessed on 31st July 2023]

NCMEC. (2023) U.S. Department of Justice CY 2022 Report to the Committees on Appropriations. *Office of Justice Programs.* Available from: https://www.missingkids.org/content/dam/missingkids/pdfs/OJJDP-NCMEC-Transparency_2022-Calendar-Year.pdf [Accessed on 29th July 2023]

Negreiro, M. (2020) Curbing the surge in online child abuse. *European Parliamentary Research Service*. Available from: https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/659360/EPRS_BRI(2020)659360_EN.pdf. [Accessed on 7th July 2023]

Rahalkar, C. & Virgaonkar, A. (2022) SoK: Content moderation schemes in end-to-end encrypted systems. *School of Computer Science: Georgia Institute of Technology*. Available from: https://arxiv.org/pdf/2208.11147.pdf [Accessed on 30th July 2023].

Safer. (2023) Safer's 2022 impact report. Available from: https://safer.io/resources/safers-2022-impact-report/ [Accessed on 8th July 2023]

Shaik, A.S., Karsh, R.K. & Islam, M. (2022) A review of hashing based image authentication techniques. Multimed Tools Applications 81: 2489-2516. DOI: https://doi-org.uniessexlib.idm.oclc.org/10.1007/s11042-021-11649-7 [Accessed on 29th July 2023].

Struppek, L., Hintersdorf, D., Kersting, K. & Neider, D. (2022) Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash.*Technical University of Darmstadt*. Available from: https://www.ml.informatik.tu-darmstadt.de/papers/struppek2022facct_%20hash.pdf [Accessed on 6th July 2023]

Thorn. (2023) Child pornography is sexual abuse material. Available from: https://www.thorn.org/child-pornography-and-abuse-statistics/ [Accessed on 5th July 2023]

Yuenan, L., Dongdong, W, & Jingru, W. (2018) Perceptual image hash function via associative memory-based self-correcting. *The institution of engineering and technology* 54(4): 208-210. DOI: https://doi.org/10.1049/el.2017.4189 [Accessed on 27th July 2023]