A Proposed Scheme of Image Hashing to Address Law Enforcement and Evidence from Child Pornography

Maria Rona L. Perez, Ace C. Lagman, John Benedict C. Legaspi, Kirk Alvin S. Awat
College of Computer Studies, FEU Institute of Technology, Manila, Philippines.
Email: mlperez@feutech.edu.ph; aclagman@feutech.edu.ph; iclegaspi@feutech.edu.ph; ksawat@feutech.edu.ph

Abstract - Governments have tended to tackle online related physical exploitation, in all its forms, and abuse with an emphasis on building the legal and social ‘structure’ to protect or rescue children by: i) establishing legislation, ii) pursuing and prosecuting abusers, iii) raising awareness, iv) reducing access to harm and v) supporting children to recover from abuse or exploitation. These are essential components of a protection response. Many legal jurisdictions, for example, fail to enact legislation sufficient to combat child abuse images or laws to criminalize grooming. There is also a lack of awareness or discomfort among parents and agencies with child protection responsibilities about the real nature of hazards or effective protection strategies. Awareness of online related child abuse and exploitation appears not yet to be organically embedded in the great majority of child protection law and responses. With the use of perceptual Hashing, pHash, algorithms to determine similar images stored on Content-Based Image Retrieval, CBIR, systems, it is possible to integrate awareness of the use of an image that has been part of online related abuse and exploitation into the broader child protection agenda, which should be a priority for policymakers.

Keywords - child pornography, image, pHash, image mining, CBIR

-I. INTRODUCTION

The law on Child Pornography and Cybercrime cannot achieve change without mechanisms in place to implement and enforce them, and services to provide support to victims. Law enforcement agencies are charged with the responsibility of ensuring that laws are applied consistently and effectively and offenders prosecuted and held to account. They therefore have a vital role in challenging sexual exploitation and abuse of children in the online/offline merged environment. Social welfare agencies have a responsibility to promote and protect the best interests of children who have experienced abuse. However, these two agendas may sometimes come into conflict. The challenge is to explore approaches that are both effective at bringing about successful prosecutions while also ensuring that the interests of the individual children concerned remain the paramount consideration.

Children who are the subject of child abuse images or those groomed for sexual exploitation may experience feelings of shame and complicity. Therefore, many victims of Internet crime do not disclose their experiences until the pictures or images are discovered, most typically by law enforcement agencies during an investigation. However, even here matters can be complicated. There have been recorded cases where even though law enforcement officers were in possession of images of a child being abused, the child being victimized has gone into denial and refused to acknowledge that they were featured.

The process of identifying children who appear in child abuse images in order to protect them and offer appropriate psychosocial support can be difficult. Images on the Internet can circulate for many years, so a picture of a 5-year-old girl, for example, may still be online 20 years later.

Hashing has been one of the tool used to identify known illegal photos, like child sexual abuse material. Prior to hashing technology, law enforcement agencies and technology companies were unable to distinguish between already-known images of child sexual abuse from those that are new. In effect, it slowed the process of detecting the illegal photos and identifying victims.

In a mail survey of US law enforcement agencies that identified a sample of Internet child pornography possession cases were no arrest was made has become the dilemmas for law enforcers to determine whether or not images fits the statutory limits and ascertaining the age of children in images has a great impact in arrest outcomes of these cases [1]. These investigations present challenges for law enforcement agents to establish techniques to identify alleged child pornography possession and possibly connect to the depicted child pornography images.

This study attempts to provide solution on the law provisions from Child Pornography and Cybercrime on proving possible evidence on a crime committed. This is important both for the development of a comprehensive case against an alleged perpetrator and for determining the duration and nature of the abuse a child has suffered in order to support recovery.
-II. ACT DEFINING THE CRIME OF CHILD PORNOGRAPHY

In our research we quote sections from the Republic Act No. 9775 [2], enacted by the Senate and House of Representatives of the Philippines in Congress assembled:

Section 1. Short Title. - This Act shall be known as the "Anti-Child Pornography Act of 2009."

Section 2. Declaration of Policy. - The State recognizes the vital role of the youth in nation building and shall promote and protect their physical, moral, spiritual, intellectual, emotional, psychological and social well-being. Towards this end, the State shall:

1) Guarantee the fundamental rights of every child from all forms of neglect, cruelty and other conditions prejudicial to his/her development;
2) Protect every child from all forms of exploitation and abuse including, but not limited to:
   a) the use of a child in pornographic performances and materials; and
   b) the inducement or coercion of a child to engage or be involved in pornography through whatever means; and
3) Comply with international treaties to which the Philippines is a signatory or a State party concerning the rights of children which include, but not limited to:

Section 3. Definition of Terms.

1) "Child" refers to a person below eighteen (18) years of age or over, but is unable to fully take care of himself/herself from abuse, neglect, cruelty, exploitation or discrimination because of a physical or mental disability or condition.

For the purpose of this Act, a child shall also refer to:

a) (1) a person regardless of age who is presented, depicted or portrayed as a child as defined herein; and
b) (2) computer-generated, digitally or manually crafted images or graphics of a person who is represented or who is made to appear to be a child as defined herein.

2) (b) "Child pornography" refers to any representation, whether visual, audio, or written combination thereof, by electronic, mechanical, digital, optical, magnetic or any other means, of child engaged or involved in real or simulated explicit sexual activities.

-III. DETECTING SIMILAR AND IDENTICAL IMAGES USING PERCEPTUAL HASHES

A hash function that is commonly used as a hash value, hash code, hash sum, or simply hash is the process of taking a big volume of data and reducing it in size by assigning a unique numerical identifier to a file. The number assigned on a hash is created based on an algorithm that is then applied to the characteristics of the data set.

The most commonly used algorithms, which are known as MD5 and SHA-1, generate number IDs that are so distinct that the chance that any two data sets are given the same hash value is less than one in one billion. A hash table is useful in hash when program to look up for data quickly. Hash functions speed up the search process in a table or database by looking for duplicated records in a large file.

Image hashing works if two pictures look practically identical but are in a different format, or resolution (or there is minor corruption, perhaps due to compression) they should hash to the same number. Despite the actual bits of their data being totally different, if they look practically identical to a human, they hash to the same thing. This is important on image authentication and possibly finding information about the image.

Perceptual hashes are a different concept compared to cryptographic hash functions like MD5 and SHA1. With cryptographic hashes, the hash values are random. The data used to generate the hash acts like a random seed, so the same data will generate the same result, but different data will create different results. If the hashes are different, then the data is different. And if the hashes are the same, then the data is likely the same. (Since there is a possibility of a hash collision, having the same hash values does not guarantee the same data.) In contrast, perceptual hashes can be compared which can give a sense of similarity between the two data sets.

Every perceptual hash algorithm that has the same basic properties: images can be scaled larger or smaller, have different aspect ratios, and even minor coloring differences (contrast, brightness, etc.) and it will still match similar images.

This method is described by Dr. Neal Krawetz on the HackerFactor blog [3]. According to Dr. Neal, a better approach for image similarity detection is pHash. Here is how pHash is computed:

1) Reduce the image to grayscale.
2) Resize the image to 32x32.
3) Compute the Discrete Cosine Transform (DCT) of the image. The DCT separates the image into a collection of frequencies and scalars.
A. Human Recognition of Low-Resolution Images

In a paper by Torralba [4], a minimal image resolution is determined which retains useful information about the visual world. In order to do this, a series of human experiments is performed on: i) scene recognition and ii) object recognition.

Studies on face perception have shown that only 16×16 pixels are needed for robust face recognition. This remarkable performance is also found in a scene recognition task. However, there are no studies that have explored the minimal image resolution required to perform visual tasks such as generic object recognition, segmentation, and scene recognition with many categories. In computer vision, existing work on low-resolution images relies on motion cues.

The study provides experimental evidence showing that 32×32 color images contain enough information for scene recognition, object detection and segmentation (even when the objects occupy just a few pixels in the image). A significant drop in performance is observed when the resolution drops below 322 pixels. Note that this problem is distinct from studies investigating scene recognition using very short presentation times.

-IV. DESIGN SCHEME PROPOSITION

This section describes the design to be used for the development of the prototype.

A. Image Mining

Image mining has two main themes. The first is mining large collections of images and the second is the combined data mining of large collections of image and associated alphanumeric data. The data mining objective might be to find if there is some pattern that exists for an object.

This study intend to build Content-based image retrieval system. One premise behind supporting object-based queries in a CBIR system is to eliminate the need for manual indexing of image content.

B. CBIR

Content-based retrieval is characterized by the ability of the system to retrieve relevant images based on the visual and semantic contents of images. [3] Content-based image retrieval, uses the visual contents of an image such as color, shape, texture, and spatial layout to represent and index the image. CBIR is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases. The CBIR system works as follows: The visual contents of the images in the database are extracted and described by multi-dimensional feature vectors. The feature vectors of the images in the database form a feature database. To retrieve images, users provide the retrieval system with example images or searched figures. The system then changes these examples into its internal representation of feature vectors. The similarities distances between the feature vectors of the query example and sketch and those of the images in the database are then calculated and retrieval is performed with the aid of an indexing scheme. The indexing scheme provides an efficient way to search for the image database. [3] The objective of this process is to find and remove content quickly so detectives can focus on images that are new and are of children who have not yet been identified.

C. Data Warehouse

Database of victim photographs are valuable resources. They can save a great deal of police time and reduce the need for officers to look at the images directly. This latter aspect is particularly important. To help identify victims, the police organization and some national law enforcement agencies need to produce databases of child abuse images. By applying sophisticated image analysis software like ImageMiner, the police can assess whether an image of a child contained within, for example, a collection that has just been seized, is identical to ones that have already been discovered by law enforcement agencies, and included in a database of known images. These software can also help identify children who have been abused over long periods of time, whose physical appearance may have changed dramatically as they grew older.

-V. CONCLUSION

Investigations into online criminal activity are complex and time-consuming. They often involve coordination across jurisdictions and concern a huge network of offenders. There are a number of constraints to effectively carrying out such investigations. The first is limited specialist expertise. Tackling online/ offline child sexual abuse and exploitation requires combined expertise in policing, computer and Internet technology and child protection. Even where staff do possess the requisite skills, the technology to investigate such crimes may not be available. Consequently, many law enforcement officers are at a disadvantage in detecting, investigating and prosecuting online related crime due to lack of resource and identifying the information of the child that experience such abuse.
The objective of this study was to investigate the utility of several methods that can incorporate to hashing technology to find child exploitation content on the Internet. Although this is a preliminary study and that more investigation is needed before any firm conclusions can be drawn, the paper has pointed to the need for explicit ethical policies that clarify how the images are used, who has access to them, in what circumstances, and the rights of victims to information about where and how the images are held. Police units and hotlines that handle child abuse images will typically have in place protocols that govern the amount of time and the locations for viewing and storage of images. The harrowing nature of the images may pose a challenge for those who work with them. It is not uncommon to find that counselling services are made available to police officers and staff who work in hotlines where viewing abuse images is an unavoidable part of the job.

From the stated proposition, the strengthened use of hashing technology to fight child sexual exploitation can be of help. This can be an essential tool for government if utilized to identify victims with the help of the technology industry to disrupt the distribution of images of child abuse. This study demonstrated the different ways that innovative technology with the use of hashing to enhance the ability to protect children online.

ACKNOWLEDGMENT

This research would not have been possible without the support of FEU Institute of Technology, College of Computer Studies. As first author, I am grateful to all of those with whom I have had the pleasure to work during this and other related projects. Each of the members of this research, my co-authors have provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. I would especially like to thank Dr. Ace C. Lagman, the director of our department. As my teacher and mentor, he has taught me more than I could ever give him credit for here.

REFERENCES

[6] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng “Fundamentals of Content-Based Image Retrieval”
[19] Klinger, Evan; Starkweather, David. "pHash.org: Home of pHash, the Open Source Perceptual Hash Library". Retrieved 2018-07-05. pHash is an open source software library released under the GPLv3 license that implements several perceptual hashing algorithms, and provides a C-like API to use those functions in your own programs. pHash itself is written in C++.