Journal of Computer Engineering
ISSN 10003428

Original Research Article
Vol 10, Issue 10, 2021 | ISSN: 1000-3428

Enhancing Digital Forensics through Hash-Based Block Matching of Image Files

Dr. Elara Vex, Dr. Kaid Rylan

Dr. Elara Vex, Department of Computer Science, University of California, Los Angeles; Dr. Kaid Rylan, Department of Cybersecurity, Georgia Institute of Technology

***Abstract*—**Internet use, intelligent communication tools, and social media have all become an integral part of our daily life as a result of rapid developments in information technology. However, this widespread use increases crimes committed in the digital environment. Therefore, digital forensics, dealing with various crimes committed in digital environment, has become an important research topic. It is in the research scope of digital forensics to investigate digital evidences such as computer, cell phone, hard disk, DVD, etc. and to report whether it contains any crime related elements. There are many software and hardware tools developed for use in the digital evidence acquisition process. Today, the most widely used digital evidence investigation tools are based on the principle of finding all the data taken place in digital evidence that is matched with specified criteria and presenting it to the investigator (e.g. text files, files starting with letter A, etc.). Then, digital forensics experts carry out data analysis to figure out whether these data are related to a potential crime. Examination of a 1 TB hard disk may take hours or even days, depending on the expertise and experience of the examiner. In addition, it depends on examiner's experience, and may change overall result involving in different cases overlooked. In this study, a hash-based matching and digital evidence evaluation method is proposed, and it is aimed to automatically classify the evidence containing criminal elements, thereby shortening the time of the digital evidence examination process and preventing human errors.

I. Introduction

A. Motivation

DIGITAL forensic science has become a rapidly growing scientific discipline, due to increasing number of serious digital crimes in recent years.

It is very important for the process of the case to evaluate the digital evidence quickly while depending on case file. As it is pointed in the Basic Human Rights Reference Guide, the right to a fair trial in criminal and non-criminal proceedings includes the right to a trial without delay or within a reasonable time. A timely trial is including the right to a timely trial [1]. The judgments must be judged in the best way so that evidences have to be perfectly examined in the judiciary can be fair. Digital forensic science also has to investigate all digital evidence. In this sense, it is very important for forensic experts to evaluate digital evidence in the best way and timely for case solutions.

Although there are many digital forensic software and

B. Problem Statement

If one is not using any compression application, one of the most important reasons for the increase in investigation time is that data images created by forensic tools are exactly the same size as the hard disk. Another difficulty of digital forensics is to handle very large amount of data. For example, if a 2 TB hard disk is full of image files with an average size of 100 MB, it takes a considerable amount of time to examine it. Nevertheless, it is clear that making digital evidence reviews by scanning images that experts obtained with forensic software tools causes time loss and skips some important evidence. Therefore, investigation process is costly and inefficient. To develop a smart system instead of human perceptions will reduce examination process and will increase accuracy.

C. Related Work

Recently, some researchers have focused on this topic, and various studies have been made. But, it cannot be said that a fully efficient system has been developed. The block hash method often has been used in studies on digital evidence investigations. However, based on the results up to this time, it is understood that all of the digital evidence reviews cannot be performed using the block hash, but block hash values can be used as an auxiliary method.

In a study, Garfinkel et al. have used block hash values to detect file types as JPEG, MPEG and compressed files, and also to do data carving based on hash values in digital evidences. They have shown that certain file types can be detected via hash values obtained from the metadata of these files [2]. In another paper, Garfinkel et al. studied that partially deleted or destroyed data can be detected using block hash values [3].

Chen et al. showed that spam mails and similar mails can be detected by developing a fast and accurate block-based hash algorithm [4].

In another study, Young, et al. developed a system to detect target files in large disk images using cryptographic hashes on sectors of data rather than entire files. They partitioned database, which contains 128-bit MD5 hash values, into multiple chunks by using the high-order bits in

How to cite: Dr. Elara Vex, Dr. Kaid Rylan (2021). Enhancing Digital Forensics through Hash-Based Block Matching of Image Files. Journal of Computer Engineering, 10(10), 58–64. DOI: https://doi.org/10.5281/zenodo.19343662

the key type [5]. In this study, a method is proposed to save time by using block hash blacklist in digital evidence investigations. Various block sizes for hashing were used, and the results were observed for future studies.

Taguchi examined coverage/time trade-offs for different sample sizes when using random sampling and sector hashing for drive triage. He concluded that a 64-kbyte read size was optimal in a wide variety of circumstances [6]. White et al. demonstrated that block hashing can be used to reduce the amount of data that needs to be examined with human perceptions during digital evidence examinations [7].

D. Proposed Approach

In this study, a block hash based preliminary examination method was proposed to provide time savings in digital evidence examinations. With this method, it is aimed to determine the evidence in an image file by using block hash matching.

E. Contribution

It takes a long time to manually obtain the digital evidence by forensic experts using forensic software [8]. In addition to this software, block hash values are used to quickly extract evidences from image files in the case. For this reason, a hash black list was created with the specified block sized hash values of the image files for extracting evidence from the digital data image. When the block hash values of the image files match with the existing hash values, the presence of image files with similar content is proved with high accuracy due to uniqueness of hash values.

F. Paper Organization

Section II describes proposed system, forensic tools and block hashing. In Section III, experimental results are performed. Conclusions are given at the last section.

II. The Proposed System

Today, hash values are very important in the field of digital forensics, but the calculation of block-hash values is not widely used in this area. Generally, the hash values of a hard disc or image files are calculated as a whole.

Although the block hash is not used extensively in the study of digital evidence investigations, there are some studies about it.

In this study, a block-hash based preliminary method is proposed in digital evidence investigation.

For this purpose, the image file is divided into pieces of different block sizes, and experiments are carried out on picture files. So, data can be detected from data pieces whose name is changed, size is changed, cropped, resolution is changed, deleted or even data are partially written on it. In this study, E01 and Raw formats are used as image format. In the E01 format, parts of the hard disk that do not contain data are compressed, but in raw format disk is copied bit by bit. However, no difference was seen in the detection of the data between these two formats.

The block diagram of the proposed system is given in Fig. 1.

A. Hashing

A hash function is a mathematical function that converts a numerical input value into another compressed numerical value. Values returned by a hash function are called message digest or simply hash values. The input to the hash function is arbitrary length, but output is always fixed length. At the end of this process, the resulting data are expected as unique. This means that another hashed file cannot give the same result as the output. Also, hash functions are irreversible functions. It means that we cannot obtain the original data by using hash values. In this case, it can be said that hash value of a datum becomes its signature. There are many algorithms that perform this operation. The most widely used hash algorithms today are SHA1, SHA2, and MD5 algorithms. Depending on the algorithm, the size of the obtained hash value may also vary. Output sizes for some of the most commonly used hash functions are given in Table I.

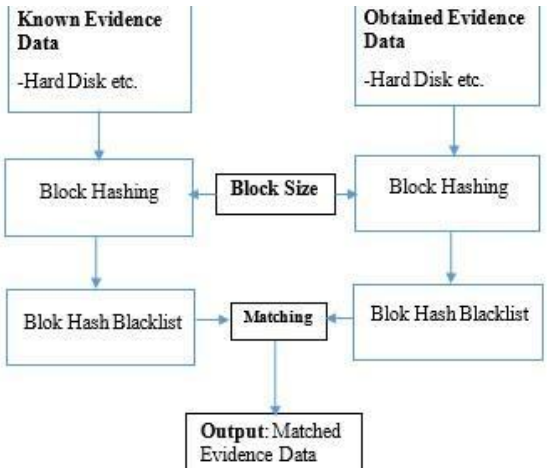

Fig. 1 Block diagram of the proposed system

TABLE I
DIGEST SIZES OF COMMONLY USED HASH ALGORITHMS

Hash Algorithm	Digest Size
MD2, MD4, MD5	128 bits
SHA-1	160 bits
SHA-2 (SHA-224, SHA-256, SHA-384, SHA-512)	224, 256, 384, 512 bits
SHA-3	Arbitrary
RIPEMD, RIPEMD-160	128,160 bits

Due to the irreversibility of the data and uniqueness of the hash value properties, hash algorithms are used

extensively in applications such as storing user data in databases such as user passwords and checking whether a datum has changed or not over time. Digital forensics applications also use hash functions to check whether the evidence is being changed in the time it is obtained or until a later time when it is examined. In a forensic examination process, firstly digital evidence to be examined is imaged, and a hash value of the image is generated. In the following processes, all operations are executed via these data. No operation is performed on the original device if it is unnecessary. The hashing process has a very important place in the field of forensics applications, since, after the evidence has been obtained, even if 1 bit of data has changed, it will lose the quality of the evidence.

In addition to ensuring the integrity of the data, the hashing is also used during the examination of the digital evidence. Large hash lists are used to ensure that data known to be suspicious can be found quickly or to eliminate data known to be clean during the digital evidence examination process. These black hash lists are created by listing the hash values of data known to be harmful. For example, files that are only related to the running of an operating system can be excluded from forensic examinations by using hash lists created by forensics experts, so the examination phase can be accelerated. However, hash values are not used as detailed evidence search methods in forensic studies. When a hash value is calculated by changing only the name of a file, the value obtained will be different from the original hash value. For this reason, it is not an efficient method to search the evidence by calculating the hash value of data as a whole.

One of the reasons why the hash values are not directly used as a criminal indicator in the digital evidence reviews is the possibility of collision of hash values. Collision is that two different data have the same hash value. MD5, one of the most well-known and widely used hash algorithms, is considered to be less secure by the National Institute of Standards and Technology (NIST) since collisions are detected in hash values produced by it [9], [10]. As a secure hash algorithm, SHA group hashes are shown (SHA-1, SHA-2, SHA-3). However, the safest method is not always preferred. MD5 is used in forensic examination software such as Forensic Explorer, FTK, Encase because of factors such as faster results, less memory consumption and low collision rate. For this reason, MD5 hashing algorithm is used in this study.

B. Block Hashing

Block hashing is the name given to the generation of hash values of file or devices that contain many data by separating them into small pieces [11]. In some studies, the hash list has been created by taking the disk sector size as the block size. The use of 512 and 4096-byte block sizes is proposed in [2], [3]. This is because the harddisks have a standard sector size of 512 and 4096 bytes.

There are several tools available for creating a block hash list. It is possible to create these lists with open source tools like md5deep [15], while block hash list can be extracted by creating scripts in licensed software such as Forensic Explorer, Encase [12], [13]. A script has been prepared for use in the Forensic Explorer software, which creates a hash list by dividing the image file into blocks of the specified size. The results have also been confirmed with the MD5deep tool. In the experiment, a sample digital evidence analysis environment is prepared. An 8 GB flash memory was used as digital evidence for this. We use some of the random images from the standard test images dataset [14] and random images from Internet, average 1 Mb in size as evidence. The hash lists of these pictures have been created and recorded in different block sizes. After formatting the flash memory, some of the sample images were copied into the flash memory by changing name, resizing, rotating, cropping, and unaltered. And the files that are not in the suspicious data set are also copied into the flash memory. With these data, the image of flash memory is taken with FTK Imager image acquisition software.

An important issue in block hash-based research is the determination of the block size. While there is no exact standard for this issue, parameters such as speed, storage area and reliability must be considered when determining block size.

Experiments were carried out for block sizes in 1 byte, 2 bytes, 8 bytes, 16 bytes, 64 bytes, 256 bytes, 512 bytes, 4 kbytes, 8 kbytes, 16 kbytes sizes. As the block size decreased, the accuracy rate and the processing time increased. It has been observed that when the hash list is constructed according to 512-byte block size, which is the sector size of the flash memory, average performance is obtained in terms of speed and performance. The graph depicting the results of the experiment is shown in Fig 2.

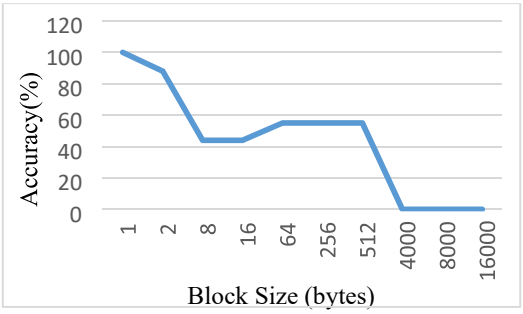

Fig. 2 Change in accuracy according to block size

III. CONCLUSIONS

As a result of this study, it has been shown that a fast preexamination system can be established in the digital evidence examinations with the block hash matching method. It has been seen that the most important parameter in the block hash matching method is the block size and that an optimization process must be applied in order to make the system available, usable, and reasonable. It is thought that the rate of performance changes very rapidly in the experiment because the memory used and the test dataset are not large.

A disadvantage of the method is the increase in the storage space required along with the reduction in block size. The reason is that hash algorithms produce fixed-length output regardless of input. For example, if 1 byte (8 bits) of data is sent as an input to the MD5 algorithm, an output of 128 bits is produced as output. As a result, a memory that is 16 times larger than the size of the input data must be stored. So, the storage space is an important parameter that influences the optimization of the block size.

Another difficulty of implementing the block hash matching method in forensic examinations is that the examinations are usually performed on disk images. When block hash operation is implemented on disk images, whole disk image is fragmented into pieces. However, our hash blacklists are created logically. That is, each file has its own individual piecewise hashes. For this reason, as the block size increases, the match rate decreases.

Further studies will be carried out on large datasets and a real case study. Programs that can perform logical operations on image files and scripts for licensed software will be developed. Thus, it is aimed to achieve matching at a higher rate. Existing software does not have this feature yet. The development of such software is thought to be an important step in accelerating the digital evidence examination process.

REFERENCES

[1] United Nations Human Rights Office of The High Commissioners “Basic Human Rights Reference Guide”, CTITF Publication Series, Published by the United Nations, New York 13-11484—March 2014.
Last accessed on 15.08.2017. https://www.un.org/counterterrorism/ctitf/sites/www.un.org.counterterro rism.ctitf/files/FairTrial.pdf.

[2] Garfinkel, S., Nelson, A., White, D., & Roussev, V. (2010). Using purpose-built functions and block hashes to enable small block and subfile forensics. digital investigation, 7, S13-S23.

[3] Garfinkel, S. L., & McCarrin, M. Hash-Based Carving: Searching media for complete files and fragments with sector hashing and hashdb, Digital Investigation, Elsevier, Volume 14, Supplement 1, August 2015, Pages S95-S105.

[4] Chen, L., & Wang, G. (2008, January). An efficient piecewise hashing method for computer forensics. In Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on (pp. 635638). IEEE.

[5] Young, J., Foster, K., Garfinkel, S., & Fairbanks, K. (2012). Distinct sector hashes for target file detection. Computer, 45(12), 28-35.

[6] Taguchi, J. K. (2013). Optimal sector sampling for drive triage.

[7] White, D. (2008, February). Hashing of file blocks: When exact matches are not useful. In Presentation notes, American Academy of Forensic Sciences 60th Anniversary Meeting. http://www.nsrl.nist.gov/Presentations.html, Accessed on 15.08.2017.

[8] S. L. Garfinkel. Digital forensics research: The next 10 years. digital investigation, 7:S64–S73, 2010.

[9] IFAC Proceedings Volumes, 42(1), 45-50.

[10] Pamula, D., & Ziebinski, A. (2009). Hardware implementation of the MD5 algorithm.

[11] Salgado, R. P. (2005). Fourth Amendment Search and the Power of the Hash. Harv. L. Rev. F., 119, 38.

[12] Forensic Explorer User Manual,2015, Link: http://www.forensicexplorer.com/forensic-explorer-user-guide.en.pdf, Accessed on 15.08.2017.

[13] EnCase Forensic Version 6.11 User's Guide, 2008, Link: http://www.thecybercrimeinvestigator.com/crj455/EnCase%20Forensic %20Version%206.11%20User%27s%20Guide.pdf, Accessed on 15.08.2017.

[14] Link: http://www.imageprocessingplace.com/root_files_V3/image_databases.h tm , Accessed on 28.07.2017.

[15] Kim, Y., & Ross, S. (2012). Digital forensics formats: seeking a digital preservation storage container format for web archiving. International Journal of Digital Curation, 7(2), 21-39.