Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 11 (2013) 86 – 92

Procedia Technology

The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013)

Systematic Literature Review on Data Carving in Digital Forensic

Nadeem Alherbawi*, Zarina Shukur, Rossilawati Sulaiman

School of Computer Science, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,43600 Bangi, Selangor, Malaysia

Abstract

Data carving is a very important topic in digital investigation and computer forensic. Researches are needed to focus on improving data carving techniques to enable digital investigators to retrieve important data and evidences from damaged or corrupted data resources. This paper is the result of a systematic literature review which answer three main questions in data carving filed. The Results fall into four main directions. First it shows the need of realistic data sets for tools testing. Secondly, it points to the need of object validation under fragmented data storage. Thirdly, investigating content based validation and its benefits in digital investigation field. Finally, it points to a new direction for using semantic validation to reduce false positive rates.

Keywords: Digital Forensics; Data Carving; Carving Validation

1. Introduction

Digital or computer forensics is defined as the practice of identifying, preserving, extracting, analyzing and presenting legally sound evidence from digital media such as computer hard drives [1]. Since the past ten years digital forensic has been changed from a technique which was almost solely used in law enforcement to an invaluable tool for detecting and solving corporate fraud. As digital forensic play a vital role in solving digital crimes it become worth to be investigated.

* Corresponding author. Tel.: +6 011-29077753
*E-mail address:*nalherbawi@gmail.com

doi:10.1016/j.protcy.2013.12.165

During a digital forensic investigation many different pieces of data are preserved for investigation, of which bit-copy images of hard drives are the most common way for the process [2]. These images contain the data allocated to files as well as the unallocated data. The unallocated data may still contain information relevant to an investigation in the form of intentionally deleted or automatically make a deletion of temporary files. Unfortunately, this data is not always easily accessible. However, a string search on the raw data might recover interesting text documents, but it would not help getting information present in, for example, images or compressed files. Beside, the exact strings to look for may not be known beforehand. Getting to this information, the deleted files have to be recovered.

There are multiple ways to recover files from the unallocated space. Most techniques use information from the file system to locate and recover deleted files. The advantage of this approach is that its relatively fast and the meta-information, such as last access date, can often be recovered as well [3]. The downside of this approach is that these techniques become much less effective if the file system information is corrupted or overwritten. In these cases, a new technique that works independently without need of the file system information is required. In other words, this can be done by identifying the deleted files and file parts directly in the raw data and extracting them in a verifiable manner [4].

Carving is a general term for extracting files out of raw data, based on file format specific characteristics present in that data. Moreover, carving only uses the information in the raw data, not the file system information. Nicholas Mikus wrote Disc carving is an essential aspect of Computer Forensics and is an area that has been somewhat neglected in the development of new forensic tools [5]. After two years since this thesis the field of carving has evolved considerably, but there are still many possible areas of improvement.

Most notably, there are few different carving techniques and there is no standard method of rating or comparing between them. Also little scientific information on carving and the results of carving tools need to be improved. This means that this field provides multiple possibilities for projects that combine scientific research into fundamental carving issues with practical improvements of carving tools [6].

In 2006, the Digital Forensics Research Workshop (DFRWS) issued a challenge to digital forensic researchers worldwide to design and develop file carving algorithms that identify more files and reduce the number of false positives. Nine teams took up this challenge. The final results of this challenge, and its winners, caused some discussion on how a carving tool should be rated. More above the winning team used manual techniques to recover the deleted files, which, as Metz stated, does not scale for realistic data sizes [7].

2. Systematic Literature Review

In order to review the current state of the art related to data carving in digital investigation point of view, a systematic literature review has been done following the procedures mentioned by [8]. The research questions that need to be raised are listed in the Table 1.

Table 1. Research Questions

ID	The Question
Q1	What are the current measuring methods for carving tool quality?
Q2	What are the different carving techniques and directions?
Q3	What are the current issues facing the researchers in data carving?

The search done on several digital libraries and databases, the language in the searching process was English language. The publishing date was not defined. Focus was only on the articles that are related to computer forensic, or digital investigation on disk area. All other irrelevant area articles were dropped.

Sources of digital libraries and databases that have been searched were IEEEXplore, Springer link, Scopus, Science Direct, ACM, and Digital Forensics Research Conference (DFRWS). Table 2 shows search strings used in above mentioned sources.

Table 2. Search Strings

No	String
1	Data recovery and digital forensic
2	Verification of data carving
3	Validation of data carving
4	File structure based carving
5	Fragmented data Carving
6	Digital forensic and data carving
7	File carving
8	Image carving
9	Data carving

The initial search ran in October 2011. Table 3 presents all findings related to each source. The selection of study involves multiple phases. First potentially relevant studies were identified using search strings, then screening made on the title and abstract of the publications. As a result a large number of publications were excluded based on their irrelevance to the research questions. On the other, hand if there was any doubt about the inclusion of potential publications the full paper would be obtained for further assessment [8].

Table 3. Representing the Number of found Publication

Database	Number of Papers	Filter Based on Title	Filter Based on Abstract
Springer link	785	136	26
Scopus	48	36	19
Science Direct	785	85	24
IEEE	141	18	14
Association for Computing Machinery (ACM)	128	5	5
DFRWS 2006-2011	12	12	12

In term of the quality of publications, a full text scanning has been made on the final set of the journals. Mendeley software has been used to manage all publications and citations. As a result a set of publications have been included in the review based on its relevancy to the research questions mentioned in Table 1 and based on the clearance of their objectives and methodology.

2.1. Data extraction

Table 4 represents sample of data extraction form that consist of five sections. Namely publication title, methodology used by the author, questions answered by publication depending on Table 1, and finally Tag which relates the content of Table 4 with Figure 1.

Table 4. A Sample of data extraction form

Publication	Methodology	Conclusion	Q ID	Key
Carving contiguous fragmented files with fast object validation [11]	Developing algorithm that validate carved data for JPEG and Microsoft documents	Internal File Structure is very important in the process of carving data and the process of validating results	Q1	K0 K8 K9

Reconstructing corrupt DEFLATED Files [12]	Bit-stream pattern search and try /error	Recovering data from corrupted archive file by examining the file structure and trying to reconstruct lost or damaged parts	Q2	K9
Forensic Data Carving [1]	Multiple Methods for contiguous data carving based on file header/footer and also file structure, with validation proposal	Discussed different methods for file carving and representing results related to these methods and limitations	Q1 Q2	K2 K3 K8
Fast in-place file carving for digital forensic [13]	Scalpel uses Boyer-Moore pattern matching algorithms to find headers and footers. The author uses Aho-Corasick multi-pattern search with asynchronous read to reduce time taken in searching for patterns.	Scalpel uses two phases to carve data by eliminating unwanted meta data from phase one (which is called in-place carving). The process time and results will be more accurate and relevant.	Q1 Q2	K0 K6
The Evolution of File Carving [3]	Analytical study to show the benefits and problems of current methods and trends in file carving.	A study that discusses in detail current methods used in file carving without the need of the file system meta data and its drawbacks and advantages.	Q2 Q3	K1 K2 K3 K8 K9
Data Recovery Function Testing For Digital Forensic Tools [14]	Developing validation and verification framework for Forensic Tools	It discusses mapping the fundamental functions of digital forensic disciplines for the purpose of validation and verification of the tools. It also demonstrates data recovery function.	Q1 Q2	K1 K2 K3
Digital forensic research: the next 10 years [2]	Literature study that suggests and predicts new directions for the coming researches in digital forensic area.	This paper points out current forensic research directions and the crises of researches nowadays. It also discusses different proposed solution for it.	Q3	K0 K4
Identification and recovery of JPEG files with missing fragments [15]	Bit pattern construction to reduce the number of pattern searches. It also uses pseudo header by trying to get info from relatively similar picture which may be taken from same source such as cameras or websites.	This paper discusses two main issues. Fist, the time spent in trying to match blocks in order to test if it matches the Huffman table for the JPEG file. Secondly, the recovery of files with missing headers by trying to regenerate a header from relatively similar images	Q2	K7 K8 K9
Measuring and Improving Quality of File Carving Methods [9]	Using identifying quality measures and attributes to developing carving tool.	This thesis explains in details different techniques used by major carving tools. Measurement methods for tool quality has been developed and used as comparison basis among them.	Q1	K4
Forensic corpora: a challenge for forensic research [10]	Proposing large scale corpora that meet a defined seven criteria. The First is to be representative which means to be accepted in the use of court of low. Secondly, complex which represents all kinds of complex data presentations on disks. Thirdly, Heterogeneous as specific pattern should be used to create the corpora. Also it should be Annotated and available. Moreover distributed in open file format and being maintained.	In this paper the author proposed seven factors that must meet in the process of creating any corpora. He also mentioned the current situation and reasons of the lack of realistic corpora. Among those reasons, privacy issue which limits the number of data sources. And the industry which is not leading this process will slow down or even prevent creating realistic corpora. At the end he proposed a set of solutions on how to develop realistic data set such as the use of anonymization tools which can remove all private data using improved models to create simulated data.	Q3	K4 K5

2.2. Analysis of the results

In this section, an analysis of the results of the systematic review will be handled. Fig 1 represents a general illustration of the answers founded regarding the research questions mentioned previously. Consequently, an elaborative analysis will follow in the next paragraphs.

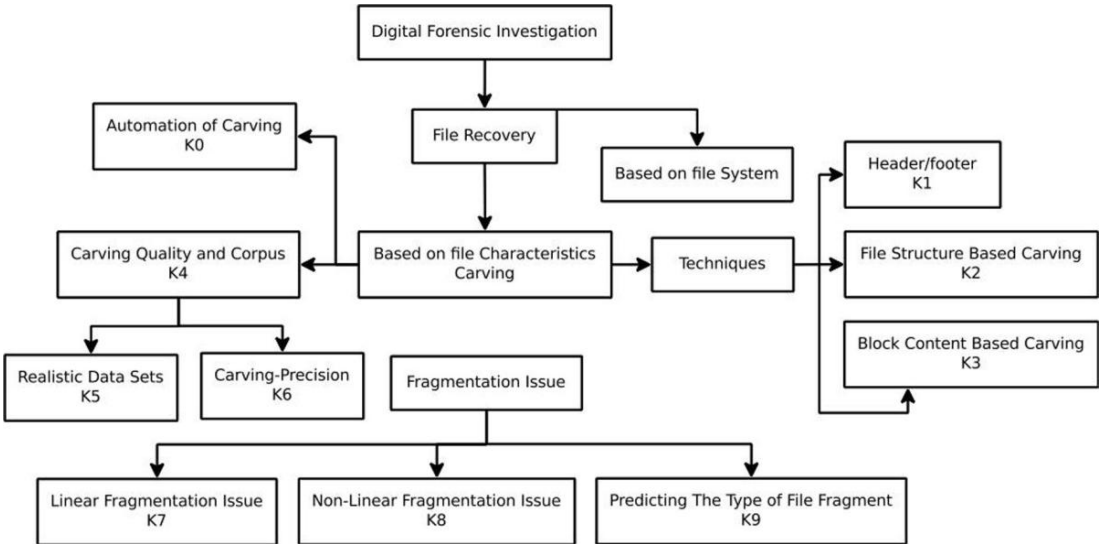

Fig 1. Data Recovery Research Area Mapping

The techniques used in file carving, answer one of the research questions. Fragmentation is considered as a serious issue and because of that, techniques was developed to be considered. For contiguous data it is usually easy to be carved using header/footer techniques which use header of specific file type and its footer as a unique identification flag. After that all the data between the header and footer will be considered as a file data section. Most of the standard formats have their own unique headers and footers which will be used in carving process to identify and recover data.

Additionally fragmented data has a different story. The previous technique will not work since the header and it's respective footer maybe not be sequentially ordered, and accordingly another file footer may exist in between. As a result if the previous technique is being used, then the carver will recover a bad corrupted file. In this way, a general approach called 'file structure based carving' has been introduced. For each type file or category of files a different technique is needed since the carver need to check and use the structure inside the data blocks to decide if these blocks of data are consistent and consider as one coherent unit in a file [15].

To clarify, if we take JPEG file format, the carver uses the Huffman code table to identify file fragments by comparing the table results with the results of matching blocks, which may or may not have fragments of that file. Additionally, another file format has its different way of identifying file fragments and many researches done on this field considering many different file format including zip files, PDF files, PNG, and XML based documents such as DOCX, for each one of these file formats different technique will be used to recover them [1].

The above technique used to recover fragmented data still produces high false positive rates. Since file structure of file which used to identify fragments may get missing, or altered, or corrupted, carvers produce higher number of potential files which lead to double or triple the storage size of carved data [16].

As stated in the previous paragraphs, clear summary and review of methods, techniques and introduction of current status of data carving have been discussed. And in this part In-place carving will be the topic to be investigated. In-place carving is one type of data carving in which it reduces the amount of recovered data which may get multiplied hundred times of the original media size. For example in one case carving of a wide range of file types from 8 GB target results in a total carved files which was over 250 GB of storage [17].

As observed, carving process generate many false positives and for that Richard et al. [17], proposed in-place carving which initially generate metadata of the data that will be carved and then these meta data will be examined by an expert who will eliminate all unwanted results. After that, a work of Xinyan et al. [13], introduced an automated in-place file carving which first identifies the location of specified headers and footers and then applies some additional constraints which is defined by the user and then the carver will recover the files that only meet their defined constraints. In-place carving helps digital investigator to reduce the numbers of carved files which need to be analyzed and examined for evidences which reduce the time needed by the investigator. Also in-place carver is in many times faster than regular carvers. For instance 16 GB storage needs 30 minutes extra when a traditional carver Scalpel is used [18].

Object validation in carving is also considered as an issue, Garfinkel [11] defined object validation as the process of determining which sequence of bytes represent a valid JPEG or PNG or any kind data. Object validation is subset of file validation since some files may contain multiple objects and for that carver may recover these objects separately.

Another important topic in object validation is using content validation, which we will focus on in our development of an enhanced in-place carver. In general, content validation tries to validate file based on content such as using semantic validation that uses human languages in the process of validation. That kind of validation works well with document type files. Over and above content validation can be used as a part of in-place carving to identify specific files based on it is content. This approach can be more beneficial in the digital investigation process. For instance, if the investigator wants to find out any evidence of any malicious act, he can use in-place carving with focus on searching in the content part of files to scrutinize any kind of malicious code. If the carver found such a code it will carve that file.

Testing the carving tools is another major issue. It deals with how to measure the tools performance, accuracy, and its false positive/negative rate. In this matter Garfinkel points out the need to realistic Dataset, which can be used to test and validate files that have been recovered [2]. This will enable researchers to figure out weaknesses in the developed tools and increase their quality. The same author developed the most used corpus for testing carving tools, which was used by DFRWS challenge in 2006. Developing a realistic data set is not an easy goal since researchers need a huge amount of disks and also permissions from users who own these disks to be able to use them for research purposes [10].

The last issue is to use aspects of languages such as English as validation indicators. Many authors suggest semantic validation for the results of carving tools to reduce false positive rates. More works need to be done for the purpose of automating this approach and supporting of many languages [19].

3. Conclusion

Four main areas have been defined. The first one is the need to real dataset or corpus that will be used to better test the carving tools and the results. There are few realistic dataset which can be used for testing purposes, but those current ones do not reflect the real complexity and openness. To achieve this, a framework for developing automated solution to make realistic dataset is needed.

Secondly validation In fragmented file is necessary especially in the domain of digital forensic point. For example if we have a sequence of bytes, then process of validation has to produce a valid file. To clarify, for JPEG file, the process validation will depend its internal structure, i.e., the entries of Huffman table. Since each file type has different internal structure more researches are needed to cover all kind of data types which need its own way of validation.

Thirdly semantic validation, which uses languages in the process of validating files is urgent issue. For instance if we have a text file or a document the content of the file should contain valid words, further more the file can be known as invalid if the carved file has nonsense words that does not have meaning. Therefore that file is carved incorrectly. Using the above approach will reduce false positive rates. Accordingly, more investigation is needed regarding semantic validation.

Finally enhancing carving validation process to enable it to detect injected codes, hidden data or potential evidences are needed by digital investigators. Most of the validation process focuses on testing the file structure as indicator of file validity but not concentrating on the content of the file itself. For example, if we have a picture

recovered correctly by the carver, and within the data blocks of the picture malicious code were hidden, this kind of information is very important in the field of digital investigation. For that reason content based validation from digital forensic point of view is essentially needed.

Acknowledgments

Supported by Universiti Kebangsaan Malaysia. We thank Dr. Yahya Amin for proof reading and valuable comments.

References

[1] D Povar, Forensic Data Carving, Digital Forensics and Cyber Crime, pp. 137-148, 2011.
[2] S. L. Garfinkel, Digital forensics research: The next 10 years, Digital Investigation, vol. 7, p. S64-S73, Aug. 2010.
[3] A. Pal and N. Memon, The Evolution of File Carving, IEEE Signal Processing Magazine, no.March, pp. 59-71, 2009.
[4] C. J. Veenman, Statistical disk cluster classification for file carving, Information Assurance and Security, 2007. IAS, pp. 393–398, 2007.
[5] N. Mikus, An analysis of disc carving techniques, Computer, no. March, 2005.
[6] L. Aronson and J. V. D. Bos, Towards an Engineering Approach to File Carver Construction, 2011 IEEE 35th Annual Computer Software and Applications Conference Workshops, pp. 368-373, Jul. 2011.
[7] B. J. Metz, Shrinking the gap : carving NTFS-compressed files, October, no. October 2009, 2009.
[8] M. M. Yusof, SOFTAM : Systematic Review Hands-on Workshop, Review Literature And Arts Of The Americas, pp. 1-12, 2011.
[9] S. Kloet and others, Measuring and Improving the Quality of File Carving Methods, Almere, Niederlande: Eindhoven University of Technology, pp. 4–79, 2007.
[10] S. L. Garfinkel, Forensic corpora: a challenge for forensic research, Electronic Evidence Information Center, April, pp. 1-10, 2007.
[11] S. Garfinkel, Carving contiguous and fragmented files with fast object validation, Digital Investigation, vol. 4, pp. 2-12, Sep. 2007.
[12] R. D. Brown, Reconstructing corrupt deflated files, Digital Investigation, vol. 8, p. S125-S131, Aug. 2011.
[13] X. Zha and S. Sahni, Fast in-Place File Carving for Digital Forensics, Forensics in Telecommunications, Information, and Multimedia, pp. 141–158, 2011.
[14] Y. Guo and J. Slay, Chapter 21 DATA RECOVERY FUNCTION TESTING, Ifip International Federation For Information Processing, pp. 297-311, 2010.
[15] H. T. Sencar and N. Memon, Identification and recovery of JPEG files with missing fragments, Digital Investigation, vol. 6, p. S88-S98, Sep. 2009.
[16] D.-gyu Park, S.-joon Park, J.-chan Lee, and S.-young No, A File Carving Algorithm for Digital Forensics, Order A Journal On The Theory Of Ordered Sets And Its Applications, pp. 615-626, 2009.
[17] Golden Richard, V. Roussev, and L. Marziale, IN-PLACE FILE CARVING, in Advances in Digital Forensics III, vol. 242, P. Craiger and S. Shenoi, Eds. Springer Boston, 2007, pp. 217-230.
[18] L. Marziale and G. R. III, Massive threading: Using GPUs to increase the performance of digital forensics tools, digital investigation, 2007.
[19] R. Poisel and S. Tjoa, Roadmap to Approaches for Carving of Fragmented Multimedia Files, 2011.