Asian Journal of Information Technology 15 (24): 5137-5144, 2016
ISSN: 1682-3915

A Survey on Data Carving in Digital Forensic

Nadeem Alherbawi, Zarina Shukur and Rossilawati Sulaiman
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. And for that reason 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 study 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 in data carving such as in-place data carving, bulk extractor and using semantic validation in data carving. Finally, a set of potential areas of interest are pointed out that needs further research and investigation.

Key words: Digital forensics, data carving, carving validation, systematic, realistic

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 (Povar and Bhadran, 2011). 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. The following section describes this role of file recovery in a forensic setting.

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 (Garfinkel, 2010). 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 (Pal and Memon, 2009). 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 (Veenman, 2007).

Motivation: 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”. In the 2 year since this thesis the field of carving has evolved considerably but there are still many possible areas of improvement.

Corresponding Author: Nadeem Alherbawi, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

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 which needs 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.

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 and Mora stated, does not scale for realistic data sizes. Finally, most current carving tools focuses on data recovery rather than evidence search which results in many lost potential evidences that could be used in court of law for that reason a study of literature is needed to discover needs and gaps.

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 Yusof (2011). The research questions that need to be raised are in Table 1.

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 DFRWS (Digital forensics research conference). Study shows search strings used in above mentioned sources.

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?

Search strings:

- String
- Data recovery and digital forensic
- Verification of data carving
- Validation of data carving
- File structure based carving
- Fragmented data Carving
- Digital forensic and data carving
- File carving
- Image carving
- Data carving

MATERIALS AND METHODS

The initial search ran in October 2011. Table 2 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 I f there was any doubt about the inclusion of potential publications the full study would be obtained for further assessment.

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 I and based on the clearance of their objectives and methodology.

Data extraction: Table 3 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 3 with Fig. 1.

Table 2: 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

Table 3: A sample of data extraction form

Publication	Methodology	Conclusion	Q ID	Key
Carving contiguous fragmented files with fast object validation (Garfinkel, 2007a, b)	Developing algorithm that validate carved data for JPEG and Microsoft documents	Internal file structure is very important in the process of carving results	Q1	K0 K8 K9
Reconstructing corrupt Deflated Files (Brown, 2011)	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 (Povar and Bhadran, 2010)	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 (Zha and Sahni, 2011)	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 (Pal and Memon, 2009)	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 (Guo and Slay, 2010)	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 year (Garfinkel, 2010)	Literature study that suggests and predicts new directions for the coming researches in digital forensic area	This study 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 (Sencar and Memon, 2009)	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 study discusses two main issues the time spent in trying to match. Fist, 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
Forensic corpora: a challenge for forensic research (Garfinkel, 2007a, b)	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 study 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 anonymousness tools which can remove all private data using improved models to create simulated data	Q3	K4 K5
Digital media triage with bulk data analysis and bulk_extractor (Garfinkel, 2013)	Proposing a new model for bulk extraction based on unique features that can be used to identify specific potential evidences such as GPS coordination	In this study the researchers propose a novel model for extracting evidences from data storage's in form of information rather than files that could help digital investigators in the process of searching for evidences	Q2 Q3	K0 K3

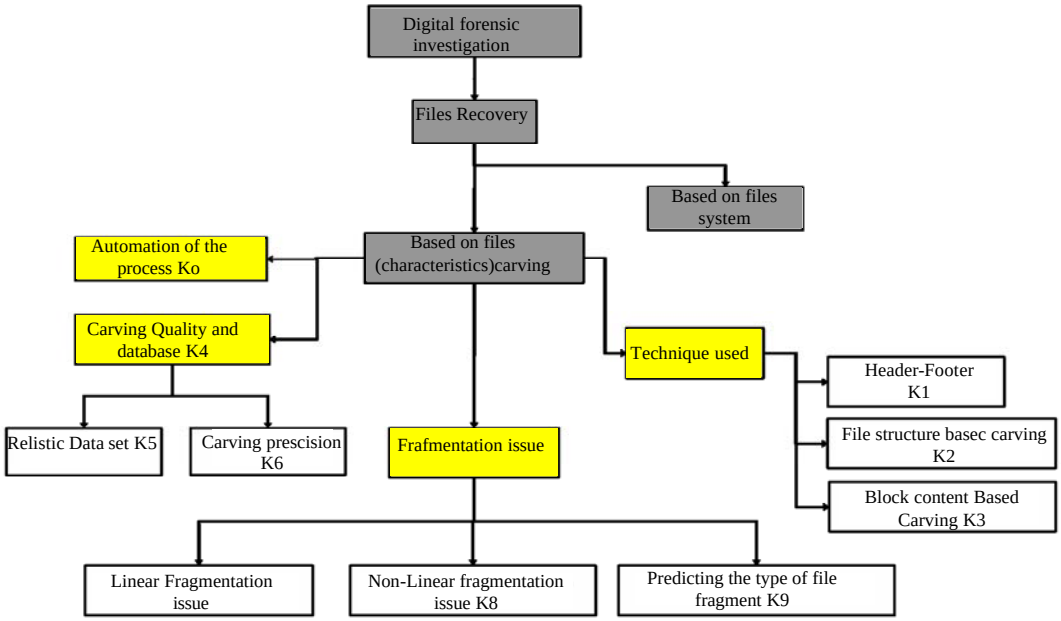

Fig. 1: Data recovery research area mapping

RESULTS AND DISCUSSION

In this study, an analysis of the results of the systematic literature review will be shown. Fig. 1 represent a general illustration of the answers for the research questions mentioned previously in Table 1. Consequently, an elaborative analysis will follow in the next paragraphs.

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 consider it. 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 header and it is respective footer maybe not be sequentially ordered and accordingly another file footer may exist in between. As a result if the previous technique 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 (Sencar and Memon, 2009). 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 format different technique will be used to recover them (Povar and Bhadran, 2011).

The above technique used to recover fragmented data that 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 (Park *et al*., 2009).

The previous paragraphs forms as an introduction of the traditional form of data carving and its issues. On the other hand in the following sections a review of a new non traditional data carving techniques will be covered. First section covers In-place carving. The second section covers forensic feature extraction And bulk extractor. While the last section covers the topic of object validation in data carving and datasets.

In-place carving: 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

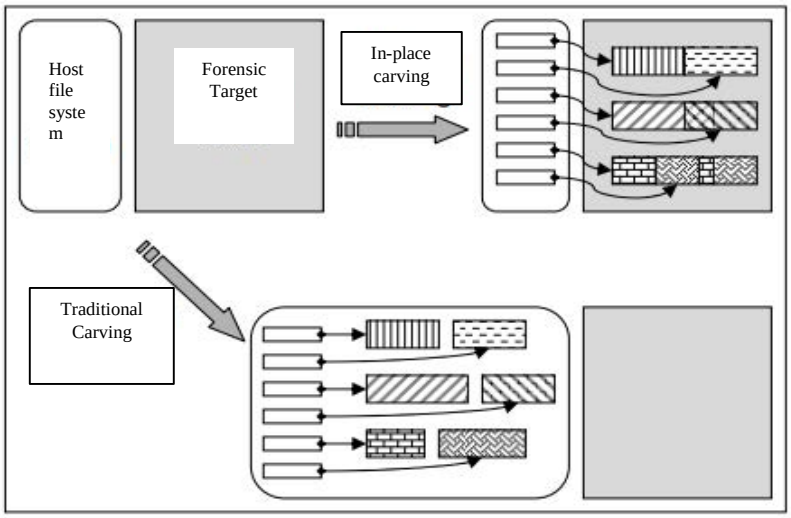

Fig. 2: In-place carving versus traditional carving

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.

The issue of the current practice of file carving is recovering data into new files which holds a big performance cost that is inherent and cannot be solved by optimizing the file carving software. The only argument for this approach is that virtually all the carvers used to view or process the recovered data need a file based interface to the data. A new approach is needed that add a file system interface to the output of the carver without actually creating carved files. Particularly, if a file system interface is arranged to candidate files without physically recreating files, existing file carvers can still be used without creating new files which many of those files will likely be invalid. This approach called "in-place" file carving. The technique is similar to that used by current file systems, except that file system metadata is stored outside the target.

Figure 2 illustrates the differences between traditional and in-place carving. The host file system and forensic target can be thought of as "input" to both traditional and in-place carving. In traditional carving, both the metadata and the data for carved files are dumped into the host file system and the target has no significant role after the carving operation completes. In the case of in-place carving, a database of metadata is inserted into the host file system, indicating where potentially interesting files are located in the target . In order to use the in-place technique and save time and space, a multi-level system is proposed. Suggest an in-place carving architecture, the first part of the proposed architecture ScalpelFS. ScalpelFS comprises three main elements, the first one is Scalpel v1.60 which provides a new mode called preview, made of a custom FUSE file system that is the second element for the purpose of providing a standard file system view of carved files. The third element is the Linux network block device, for the purpose of carving of remote disk targets (Fig. 2).

Dutch National Police Agency has proposed Another similar approach named as the carved path zero storage Library and filesystem (CarvFs). They develop a library that provides in the low-level needs of zero storage carving. It does this by providing an interface to hierarchically ordered fragment lists and allowing these fragment lists to be converted to and from virtual file paths. These particular virtual file paths can be used in conjunction with the CarvFS filesystem, a pseudo filesystem build using fuse and LibCarvPath (Zha and Sahni, 2010).

Finally 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 min extra when a traditional carver Scalpel is used (Marziale *et al*., 2007).

Forensic feature extraction and bulk extractor: Forensic investigator becomes the victims of their achievement. Since, digital storage devices in all different shapes are such valuable sources of information, they are now routinely seized in many digital investigations. As a result, investigators do not have the time to investigate all

the storage devices that comes across their desks. When the investigator is available, the contents of the device are copied to a working storage drive to maintain chain of custody. This bit to bit copy of the drive is then opened or mounted using a forensic tool, after that the investigator can perform variety of analysis such as string searches or manually explore the image. When the analysis is finished, the copy is removed from the system and the investigator handles to the next drive.

The previously mentioned approach has multiple drawbacks as been pointed out by (Garfinkel, 2006). First, it has priority issue related to which has to come first the resources and storages or the attention of the examiner on the value of information that the storage media contains. The second issue is related to the lost potential correlation among data from various storages, files and objects which can help in connecting all the dots related to the case on the hand. Finally, traditional forensic tools focus on recovering documents while the traditional approach neglect data on the drive that cannot be reconstructed to be filed. Forensic tools should enhanced and adapted to be evidence focus rather than documents and files recovery.

Currently, two general techniques are common in the processing of digital evidence that balances each other. File-based approaches and bulk data analysis. The file-based technique is widely used by digital forensic investigators and many popular tools such as EnCase and AccessData's FTK implement such approach. This kind of approach operates by finding, identifying, extracting and processing files pointed by file system metadata (Garfinkel, 2013). This has multiple advantages among other techniques since its easy to understand and it integrates well with most legal systems since extracted files can be easily utilized as evidence. On the other hand, it suffers from ignoring data that are not contained within files or not pointed out by metadata entries (Garfinkel, 2010).

On the other hand, bulk data analysis technique examine data storage and identify potential evidances based on content then processed and extracted without returning or using of file system metadata. An example of this approach is file carving but it has limitation which is ignoring Bulk data that cannot be assembled into files. Both methods file based and bulk data analysis complement each other. In file-based approach, the results are easier to put in context and to be explained to an individual who does not have technical knowledge. On the other hand bulk, data analysis applies on all kind of computer systems, file system and file types since it does not rely on the metadata of the file system. Additionally, it also can be applied on damaged or partially overwritten storage media.

Feature extraction technique is a new model for bulk analysis that work by first scan and search for pseudo-unique features which are an identifier that has sufficient singularity such that within a giving data it is highly unlikely that the identifier will be repeated by chance and then store the results in an intermediate file. An example of both feature extraction and pseudo-unique features is and email extractor which can recognize RFC822-style email addresses via unique identifiers which are the Message-ID value (Garfinkel, 2006).

The bulk extractor is An example of a tool that applies the two previously mentioned approaches above. The program operates on multiple disk images, file or a directory and extracts useful information without returning to the file system metadata (Garfinkel, 2013). The results can be easily checked, determined or processed with automated procedures. Bulk extractor also creates histograms of the occurrence of features that it finds since features that are more common in the media tend to be more important.

The bulk extractor has multiple scanners that run sequentially on target digital evidence and each scanner record extracted features in a certain mechanism and then the tool perform post-processing for the extracted features and then exit. The bulk extractor has two types of scanners basic and recursive. An example of a basic scanner is an email scanner that searches for email addresses RFC822 Headers and other recognizable strings in the email message. A recursive scanner as it implies can decode data and pass it back for re-analysis for further scanning an example of this kind of scanner is zip-scanner, since compressed file may contain multiple types of other data form and files.

Object validation in data carving and datasets: Object validation in carving is also considered as an issue, Garfinkel 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 (Garfinkel, 2007).

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. 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 (Poisel and Tjoa, 2011).

Finally, 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 (Garfinkel, 2010). 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 (Garfinkel, 2007).

CONCLUSION

Throughout the whole process illustrated above 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. Another potential point is to Investigate new ways for feature bulk analysis is essential for encoded data such as MP3 files and JPEG images, since current models and tools search for features from text based files such as docs and text files.

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 it self. 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.

REFERENCES

Brown, R.D., 2011. Reconstructing corrupt DEFLATEd files. Digital Invest., 8: 125-131.

Garfinkel, S., 2007a. Forensic corpora: A challenge for forensic research. Electron. Evidence Inf. Center, 2007: 1-10.

Garfinkel, S.L., 2007b. Carving contiguous and fragmented files with fast object validation. Digital Invest., 4: 2-12.

Garfinkel, S.L., 2006. Forensic feature extraction and cross-drive analysis. Digital Invest., 3: 71-81.

Garfinkel, S.L., 2010. Digital forensics research: The next 10 years. Digital Invest., 7: 64-73.

Garfinkel, S.L., 2013. Digital media triage with bulk data analysis and bulk-extractor. Comput. Secur., 32: 56-72.

Guo, Y. and J. Slay, 2010. Chapter 21: Data recovery function testing. Ifip Intl. Fed. Inf. Process., 2010: 297-311.

Marziale, L., G.G. Richard and V. Roussev, 2007. Massive threading: Using GPUs to increase the performance of digital forensics tools. Digital Invest., 4: 73-81.

Pal, A. and N. Memon, 2009. The evolution of file carving. IEEE. Signal Process. Mag., 26: 59-71.

Park, D.G., S.J. Park, J.C. Lee, S.Y. No and S.Y. Shin, 2009. A file carving algorithm for digital forensics. Proceedings of the International Conference on Computational Science and its Applications, June 29-July 2, 2009, Springer, Berlin, Germany, ISBN:978-3-642-02454-2, pp: 615-626.

Poisel, R. and S. Tjoa, 2011. Roadmap to approaches for carving of fragmented multimedia files. Proceedings of the 2011 6th International Conference on Availability, Reliability and Security (ARES), August 22-26, 2011, IEEE, New York, USA., ISBN:978-0-7695-4485-4, pp: 752-757.

Povar, D. and V.K. Bhadran, 2011. Forensic data carving in lecture notes of the institute for computer sciences. Social Inf. Telecommunications Eng., 53: 137-148.

Sencar, H. T. and N. Memon, 2009. Identification and recovery of JPEG files with missing fragments. Digital Invest., 6: 88-98.

Veenman, C.J., 2007. Statistical disk cluster classification for file carving. Proceedings of the 3rd International Symposium on Information Assurance and Security (IAS 2007), August 29-31, 2007, IEEE, New York, USA., pp: 393-398.

Zha, X. and S. Sahni, 2010. Fast in-place file carving for digital forensics. Proceedings of the International Conference on Forensics in Telecommunications, Information and Multimedia, November 11-12, 2010, Springer, Berlin, Germany, ISBN:978-3-642-23602-0, pp: 141-158.