Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 14 (21): 2704-2712, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.2704.2712

An Ordered Selective Imaging and Distributed Analysis Computer Forensics Model

[1,2]Waleed Halboob, [1]Ramlan Mahmod, [1]Nur Izura Udzir,
[1]Mohd. Taufik Abdullah and [1]Ali Deghantanha
[1]Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia
[2]Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia

Abstract: The traditional computer forensics procedures and tools collect and analyze the entire user data. This scenario has been proven to be not appropriate any more due to increased size of user data and storage. Accordingly, selective imaging and distributed analysis concepts have been introduced in the literature to reduce the digital evidences collection and analysis costs (time and resources). Current selective imaging approaches image the relevant data according the order of their selection and not according to their physical offsets order inside the targeted storage. Furthermore, integrating the selective imaging and distributed analysis has not been considered yet. This study proposed a computer forensics investigation process that provides an efficient imaging and scalable analysis. The selected data artifacts are first ordered upon their physical offsets. Then, based on the selected data size and available investigation time, the selected data are imaged into one or more partial forensic image in such a way that the produced images can be analyzed by different investigators and using several machines. An Advanced Forensic File Format 4 (AFF4) is used as a container for the collected relevant data. An experiment study has been used to evaluate the performance of the selected imaging process. The result shows that, even if ordering the selected digital evidences has a small performance negative impact but it has a positive effect on the performance of the selective imaging process itself. A qualitative study has been also used to evaluate the system and management scalability of the distributed analysis.

Key words: Computer forensics, selective imaging, digital evidences, efficiency, distributed analysis

INTRODUCTION

Existing computer forensics solutions collect digital evidences by making a bit-by-bit image from the entire suspect's data storage and later on, at the Computer Forensics Lab (CFL), the bit-by-bit image is analyzed. This procedure has been already proven to be undesirable solution due to the increase in user data and storage size which will increase the investigation cost (time and resources). According to Stuttgen *et al*. (2013), imaging two Terabytes hard disk with an about 70 megabytes per second imaging bandwidth requires about 8 h. This means that the analysis process will need one day more.

For addressing this issue, selective imaging and distributed analysis concepts are proposed. The idea behind the selective imaging concept is to collect only pre-selected relevant data. A pre-analysis step is used for selecting the data artifacts that seem to be relevant to the crime. In addition, distributing computer forensic analysis task among different machines and several investigators is highly required today to come out with evidence in acceptable time (Roussev and Richard III, 2004).

Our study proposed by Halboob *et al*. (2014) is extended here with proposing a computer forensics investigation model that provides an efficient imaging and scalable analysis. The imaging process is based on the selective imaging concept. The selected data artifacts are first ordered according to their offsets on the targeted suspect storage. An Advanced Forensic File Format 4 (AFF4) is used as evidence container. Based on the size of the user data and available investigation time, the relevant selected data are imaged to one or more AFF4 image file. Each AFF4 image will be then analyzed, using a separated machine and by a different investigator.

Corresponding Author: Waleed Halboob, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia

Researchers have tried to address the issue of increase user storage and data size. The first attempt was trying to reduce the required amount of storage space using what is called block based compression (Kloet *et al*., 2008; Garfinkel *et al*., 2006) to the data stream. Another solution is a hash-based disk imaging (Cohen and Schatz, 2010) in which the amount of the collected data is reduced using data de-duplication and reduction technologies.

Recent study on this research consider this issue by separating the digital evidence collection or imaging step from digital evidence analysis. In the former step, the selective imaging concept (Turner, 2005a, b, 2006, 2007; Richard III and Roussev, 2006) is used to image or collect only relevant data to a crime instead of making a physical bit-by-bit image from whole user storage device. Researchers on selective imaging concept have proposed several methods such as risk sensitive digital evidence collection (Kenneally and Brown, 2005) and digital evidence bags (Turner, 2005a, b, 2006, 2007; Richard III and Roussev, 2006). In (Stuttgen, 2011; Stuttgen *et al*., 2013), the first implemented selective imaging model is proposed. This model enables the investigator to use the selective imaging concept in a forensically sound manner.

According to Turner (2006), the relevant data artifacts can be identified or selected through manual, semi-automatic and fully automatic selections. Using the manual selection method, an investigator identifies the relevant data artifacts from, for example, a folder tree. With the semi-automatic selection method, the relevant data artifacts are identified using tools enabled with search engines, for example, searching for data according to attributes such as content, name, extension or signature. Finally, the fully automatic selection uses intelligent methods for deciding which data artifacts are relevant and according to some parameters given by the investigator. The fully automatic selection still has several shortcomings, the manual and semi-automatic selections are now totally supported by existing computer forensics tools (Stuttgen, 2011).

In term of the digital evidence analysis, the proposed solutions consider the cost of both the required time and storage. The effective and efficient analytical concept is proposed by Beebe (2009). Here, a bit-by-bit image is made from the whole user storage device during evidence collection and then the collected image of data is analyzed selectively or in a distributed manner. Researchers using the effective and efficient analytical concept have applied distributed evidence analysis (Roussev and Richard III, 2004), data mining search process (Beebe and Clark, 2005), file classification (Sanderson, 2006) and clustering text-based search (Beebe and Clark, 2007).

As discussed above, several research efforts have been directed to resolve the problem of imaging and analysis cost in term of the required and time and resources. But, studying how the partial (or selective) forensic image can also be analyzed in a distributed manner is still a research gap.

PROPOSED MODEL

The proposed model has two main modules namely “selective imaging” and “distributed analysis” modules as illustrated in Fig. 1. The digital evidence identification and preservation steps are also consider here but as sub-steps from the selective imaging module. This means that the selective imaging module includes digital evidence identification, collection and preservation. The following sub-sections deal with the selective imaging and distributed analysis modules in more detail.

Selective imaging module: The selective imaging module has three main steps shown in Fig. 2. These steps are

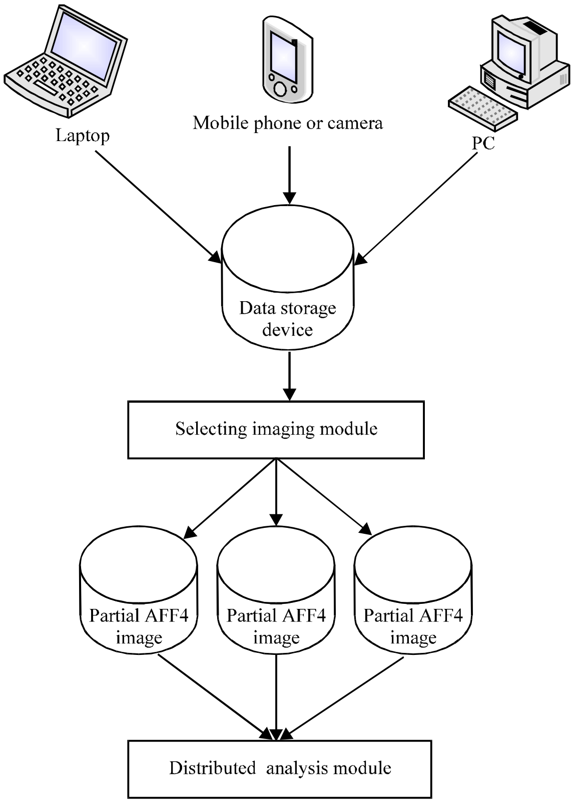


Fig. 1: Architecture of the proposed model

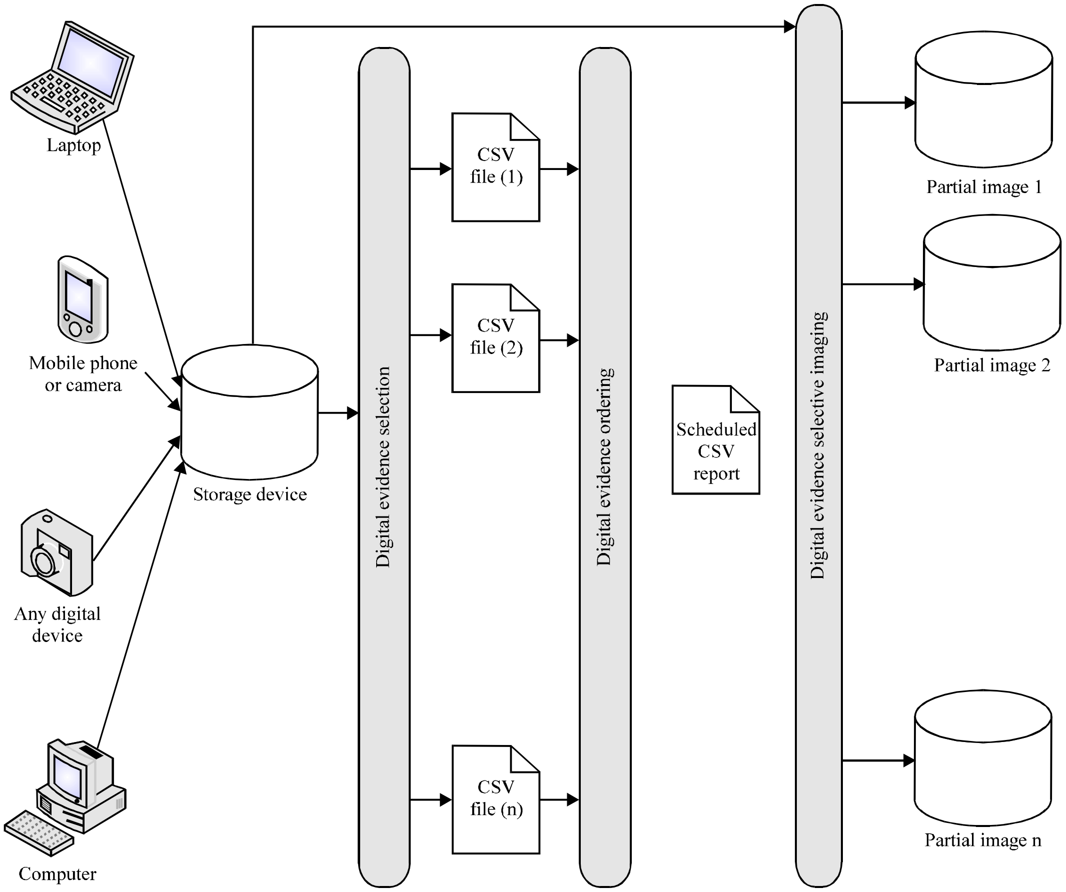


Fig. 2: Selective imaging process

digital evidence selection, digital evidence ordering and digital evidence imaging. These steps are discussed in the following sub-sections.

- **Digital evidence selection:** Usually this step is used for identifying anything related to the happened crime from the crime scene. In case of using a bit-by-bit imaging method, this step is used for identifying any user storage (hard disk, flash drive, CD-Room, etc.) that may contain an evidence. But, if the selective imaging concept is used, this step must also identify which data artifacts inside the user storages are relevant. This is because the selective imaging method uses a pre-analysis step to select the relevant data items, not only relevant storages. Unlike the bit-by-bit imaging process, in which the irrelevant user data are filtered out during the analysis step which comes after the entire user data are collected. With the selective imaging, the filtering task can be done in two different phases and times. First, a pre-analysis is used before imaging for selecting only the data that seem to be relevant to the crime. Later on/and at the analysis phase, the collected relevant data are analyzed and more irrelevant data are also filtered out to come out with the required digital evidence (e.g., files, facts, timeline and so on)

As a result, here the digital evidence selection step is used for identifying or selecting the relevant data artifacts, not only storage devices to the digital crime. In this study, the semi-automatic selection method is assumed to be used as this method is the most practical solution as discussed in litterature review. An investigator runs a forensic recovery tool (or tools) from an external machine to scan and pre-analyze the targeted storage device in a read-only mode. The targeted device must be scanned to determine all existing data even active or deleted. The targeted device is accessed in a read-only mode to ensure that the content of the targeted device will not be altered. After that, the investigator starts the pre-analysis task for selecting the relevant evidence and saves the search results in a report(s).

Several computer forensics tools-such as Winhex (X-Ways), FTK Access Data and CnWRecovery have been tested and used. These tools enable the investigator to search and select the relevant data artifacts by using different search methods. Additionally, the investigator can report the search results in a standard file format mostly in a Common Separated Values (csv) file format report. On the other hand, the output of this step is several csv files for different search tries in which each csv file contains several hits for one search result. The csv reports contain metadata (file name, path, size, offset, etc.) of all found data (existing and deleted artifacts).

```
Input:   CSV REPORTS_LIST
Output:  Ordered csv report
     Let Ordered csv = null;
     For each csv report in CSV REPORTS_LIST
          Let i = 1;
          If data_artifact[i] not found in Ordered csv report then
          Begin
               Add data_artifact[i] to the Ordered csv report
          i = i+1;
          End Else i = i+1;
     End For
     Let i = 1;
     Let min = data_artifact[i];
     While i<n
          Let j = i+1
          While j< = n
               If data_artifact[i] position > data_artifact[j] position then
               min = data_artifact[j]
          End While
          Replace data_artifact[i] with min
     End While
```

Fig. 3: Proposed digital evidence ordering algorithm

- **Digital evidence ordering:** After identifying the relevant evidence into csv files, the search results (found inside the csv files) are passed to the digital evidence ordering algorithm which simply (Fig. 3) first merges the csv files together into one csv file called ordered csv report and then orders them according to their position inside the targeted user storage
- **Digital evidence imaging:** The evidence imaging step, shown in Fig. 4, is used for imaging all relevant

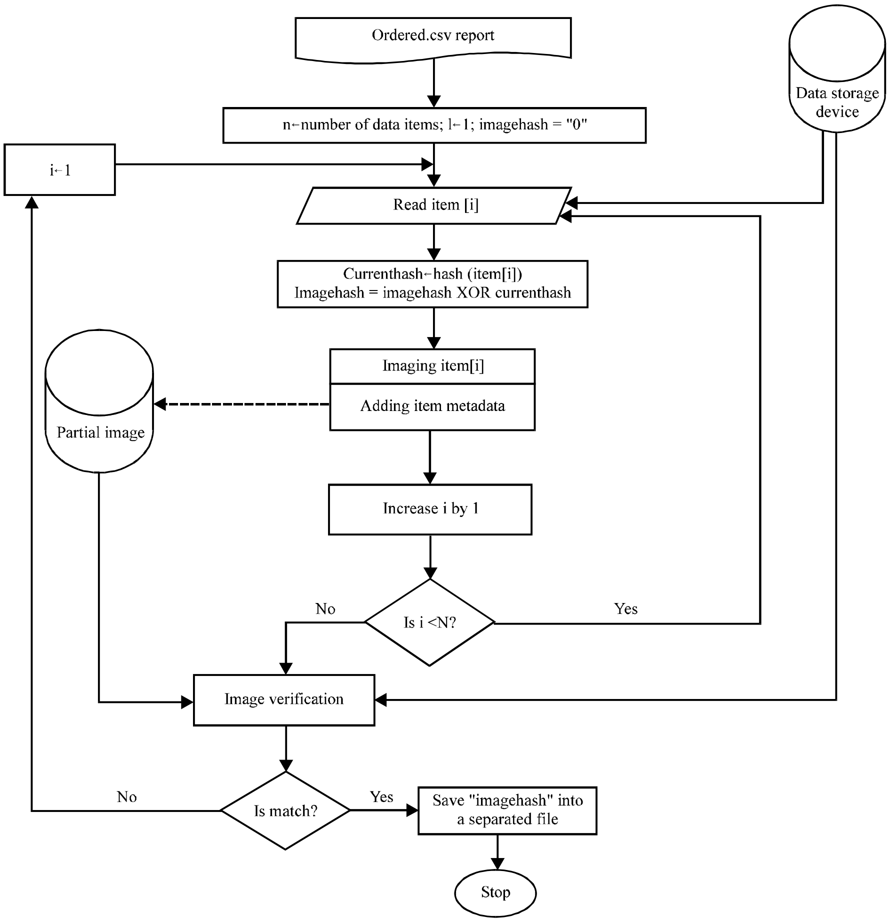


Fig. 4: Execution of the selective imaging process

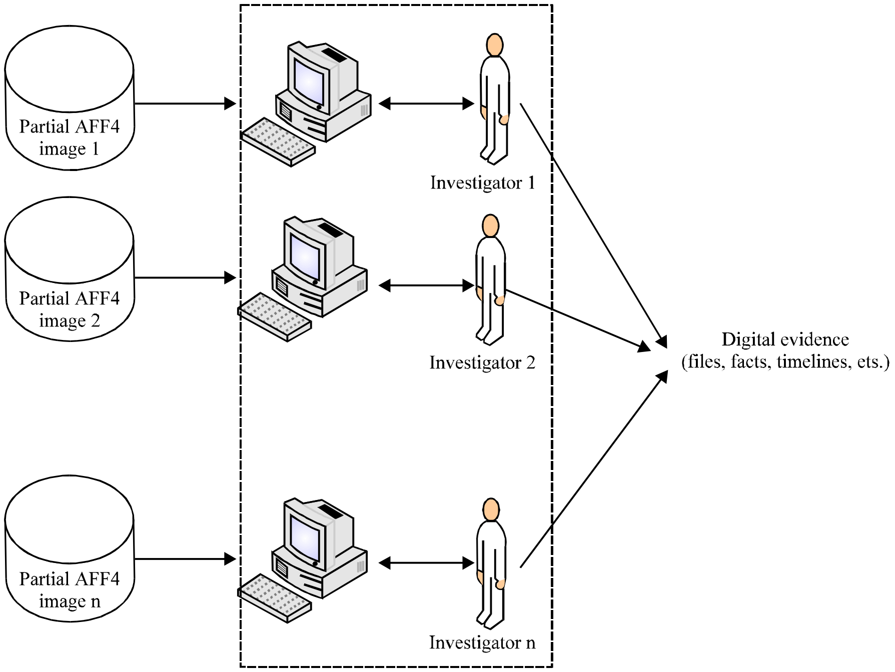


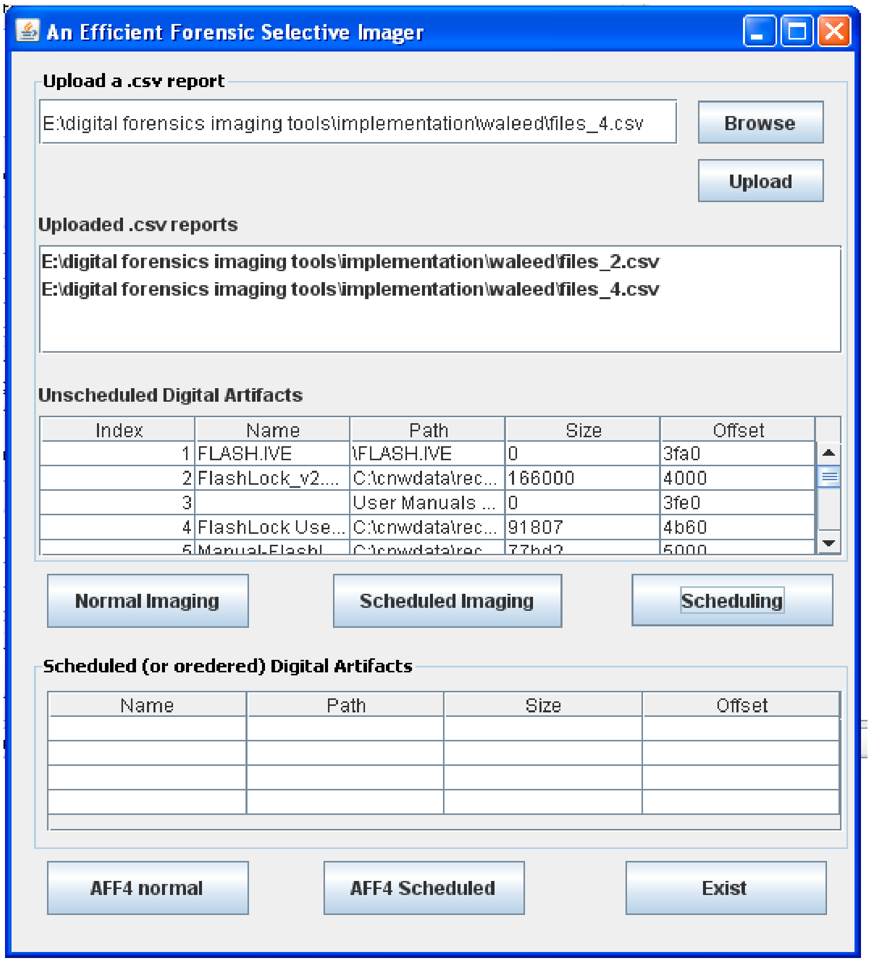
Fig. 5: Distributed digital evidence analysis

data artifacts to one or more partial forensic images. The partial forensic image contains all relevant data artifacts along with their metadata (hash value, imaging time, size, name, extension, address, last modified, etc.). Each relevant data artifact is read from the user's storage device, hashed and then the data artifact with its metadata is written into the partial image

In terms of the partial forensic image used, the selective imaging process requires a forensic image that supports at least two futures which are: (1) Multi objects streams as each data artifact which is selectively imaged and as a result, needs to be stored inside an image as a separated object stream and (2) Storing metadata of each object stream. Several existing forensics images have been used in the literature such as RAW/DD, SGZIP, EO1, AFF3 and AFF4. In our model, the AFF4 is used as it provides the required futures (Stuttgen, 2011). Finally, the integrity of the collected partial AFF4 image is ensured by comparing the hash value of the imaged data artifacts inside the image with the hash value generated during reading data artifacts from the data storage device. Then, the hash value is signed with a public key cryptography.

Distributed analysis module: First and during the selective imaging process, the relevant data can be classified before collecting them based on their types (e.g., photos, videos, documents, etc.) or location (such as hard drive partition). Therefore, the relevant data can be imaged into several AFF4 forensic images and each AFF4 image can separately analyzed with different investigator. In other words, at the Computer Forensics Lab (CFL), the selective AFF4 images can be distributed among several investigators to be analyzed using different forensic computers and tools. As illustrated in Fig. 5, it is clear that the analysis workload is easily distributed since, the collected data is already distributed among different AFF4 images.

Implementation: The proposed model is implemented using Java programming language on NetBeans IDE 6.9.1. Some additional java Application Programming Interfaces (APIs) are used. In other words, the Javacsv 2.1 free java package is used for processing the csv files. The Advance Forensic File Format (AFF4) java package-called 'Truezip1.6.1' which is used for creating the AFF4 partial image and writing selected data artifacts into it. The SHA-1 message digest is used for hashing the data artifacts. Figure 6 shows a screenshot from the implemented prototype. Whereas, the "Scheduled" or "Scheduling" words refer to ordered or ordering, respectively.



An Efficient Forensic Selective Imager

Upload a .csv report

E:\digital forensics imaging tools\implementation\waleed\files_4.csv

Uploaded .csv reports

E:\digital forensics imaging tools\implementation\waleed\files_2.csv
E:\digital forensics imaging tools\implementation\waleed\files_4.csv

Unscheduled Digital Artifacts

Index	Name	Path	Size	Offset
1	FLASH.IVE	\FLASH.IVE	0	3fa0
2	FlashLock_v2....	C:\cnwdata\rec...	166000	4000
3		User Manuals ...	0	3fe0
4	FlashLock Use...	C:\cnwdata\rec...	91807	4b60
5	Manual-Flashl...	C:\cnwdata\rec...	77bd2	5000

Scheduled (or oredered) Digital Artifacts

Name	Path	Size	Offset

Fig. 6: A screenshot from the implemented prototype

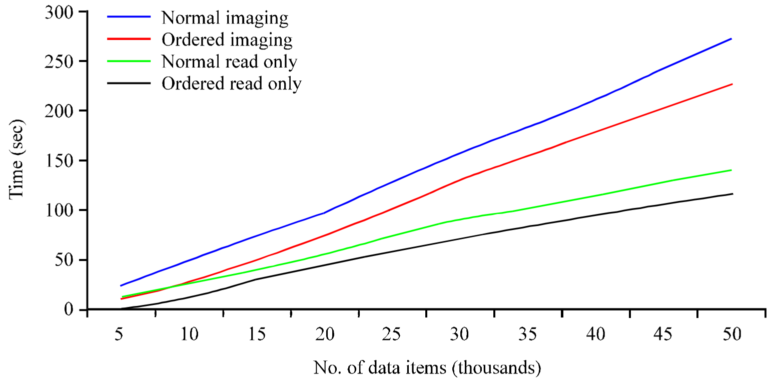


Fig. 7: Performance result of the hard disk direct selective imaging

RESULTS AND DISCUSSION

To evaluate the proposed model process, a performance study is used to evaluate the imaging process efficiency. The distributed analysis system and management scalability is evaluated using a quantitative analysis. The following sub-sections present our result and discussion.

Performance study: The performance of the proposed model is measured to evaluate its efficiency. To run the experimental study, Enron dataset is used. Two computing storages are considered hard disk and flash drive. Due to the limited capacity of the flash drives, only 50,000 files from the Enron dataset are used. The exact size of the used Enron files is 124 MB. These files are divided into 10 groups. Each group contains 5000 files and copied into the hard disk and flash drives with other random data until the hard disk and flash drives are occupied. Our experiment is executed on Microsoft Windows XP, Dell computers with Intel Core Quad CPU (4 CPUs), 2.83 GHz speed and 4096 MB memory.

Then, the performance of the imaging process is measured in normal and ordered cases. With the ordered case the digital evidence ordering algorithm is applied first to the relevant data artifacts before imaging. This is to measure the implication of the relevant data positions or offsets on the imaging process efficiency. Two types of current selective imaging methods are considered. First, the relevant data artifacts are directly imaged or copied from the user storage to the investigator used storage file by file. The second method uses a forensic partial image format. To date, only the AFF4 image format supports the selective imaging concept. The cost of reading/writing and reading only the relevant data is measured. This is to clearly identify the impact of the relevant data position. In any case, in both selective imaging methods, the metadata and hash value of the imaged data are considered. In addition, the performance is measured with two different computing storages: Hard disk and flash drive.

- **Ordering algorithm performance:** Here, the performance cost of the digital evidence ordering algorithm is measured. This algorithm is proposed to merge and order the relevant data for improving the efficiency of the imaging process. So, the cost of the merging and ordering process should be less than the outcome of using it; in other words, its negative impact must be less than its positive impact on the imaging process. However, ten csv. files are used where each file contains metadata of 500 data artifacts from Enron dataset. There are 5000 files only about 82 of which are duplicated. The costs of merging and ordering tasks are measured independently

We found that the cost of merging the ten files takes only 573 msec (mile seconds) while the cost of ordering the merged csv files takes only 452 msec (mile second) also in average. The cost of our ordering algorithm is only about 1.025 sec. The performance of the imaging process is measured in terms of required time within two different cases. The first case is a non-ordering imaging and when the relevant data artifacts are randomly imaged without considering their position in the user storage. The second case is a ordered imaging, where the relevant data are ordered first according to their position before imaging using and our ordering algorithm. Each case is evaluated using two imaging approaches. The first approach is a normal or direct imaging by collecting relevant data artifacts into storage. The second approach is AFF4 imaging. Only 50,000 files from the Enron dataset are used during this evaluation.

Figure 7 shows the performance result of the normal and ordered imaging in a hard disk device. Here, the relevant data are imaged from one hard disk device to another. It is clear from the result that the ordered imaging is more efficient than the normal imaging in both cases (reading/writing and reading only). The cost of reading and writing 50,000 files normally requires about 272.2243 sec while with the scheduled imaging requires only about 226.7867 sec. As a result, the cost is reduced by about 17%. In term of the reading only the cost is also reduced by 17% also and from 140.5423-116.1765 sec.

The performance result of flash drive is illustrated in Fig. 8. The relevant data inside the flash drive is imaged into hard drive storage. The cost of normal imaging is reduced by 29% (419.958-325.3977 sec) when the ordering algorithm is used in read/write imaging. While with the reading only, the cost is reduced by about 4% only and from 188.374-181.279 sec.

The performance result of imaging the relevant data into an AFF4 partial image from hard and flash drive is shown in Fig. 9 and 10, respectively. The normal and ordered imaging cases are considered to study the impact of relevant data positions on the imaging efficiency. With the hard drive storage, the cost of imaging is reduced by 23% (189.349-144.66 sec) when the ordering algorithm is used. With the flash drive, the cost is reduced by only 4% (218.3715-208.486 sec).

Distributed analysis scalability: The scalability strategy used for digital evidence analysis is distributing the system and management workloads into several machines and investigators, respectively. Regarding the system scalability, analyzing the AFF4 images will be executed without facing any bottlenecks (such as management bottleneck, centralized resources bottleneck, etc.) that may cause the analysis process to scale down. Since, the

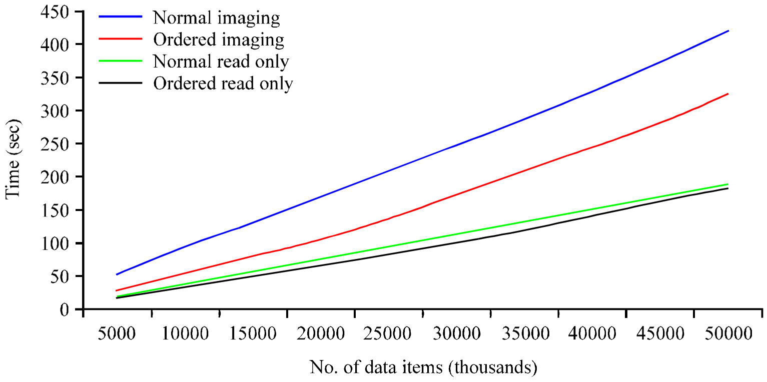


Fig. 8: Performance result of the flash bit-by-bit stream partial image

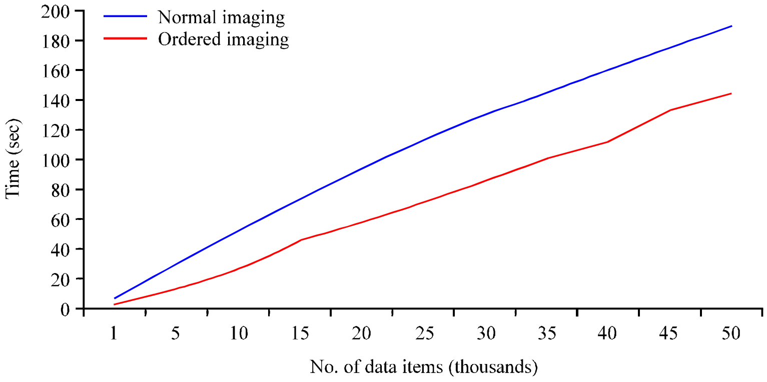


Fig. 9: Performance of the hard disk AFF4 partial image

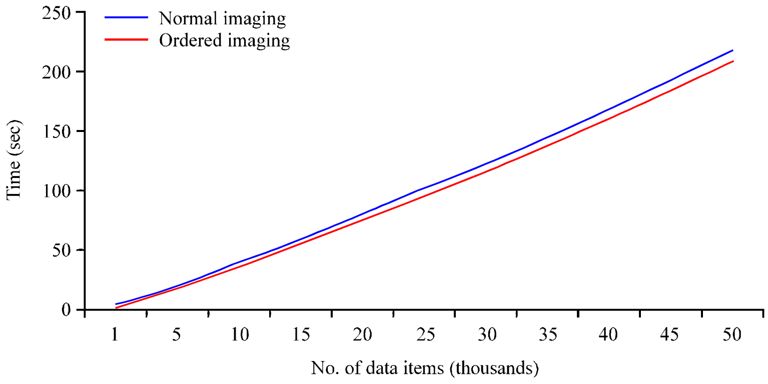


Fig. 10: Performance of the flash drive AFF4 partial image

relevant data is stored in several AFF4 images and analyzed with different machines. In the same manner, the management workload is distributed among several investigators and the analysis result can be reported faster.

CONCLUSION

This study proposes an efficient and scalable computer forensics model based on selective imaging and distributed analysis concepts. The proposed model includes a digital evidence ordering algorithm to study the impact of the relevant data offsets on the efficiency of the selective imaging process. The proposed digital evidence ordering algorithm merges and orders the relevant data artifacts based on their position on the user data storage. Furthermore, the relevant data can be imaged into one or more AFF4 images. The proposed model is implemented and its efficiency is measured. The result shows that the ordering algorithm has a small negative impact on the imaging process but it has a good impact on the efficiency of the whole imaging process. The ability of the proposed model to image the relevant data into several AFF4 images leads to having a scalable digital evidence analysis. Actually, this study is a part of a larger research project that proposes an efficient, scalable and flexible computer forensics framework that preserves the privacy of user during the computer investigation process and using the selective imaging concept.

REFERENCES

Beebe, N.L. and J.G. Clark, 2005. Dealing with terabyte data sets in digital investigations. Proceedings of the IFIP International Conference on Digital Forensics, February 13-16, 2005, National Center for Forensic Science, Orlando, FL., USA., pp: 3-16.

Beebe, N.L. and J.G. Clark, 2007. Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. Digital Invest., 4: 49-54.

Beebe, N., 2009. Digital forensic research: The good, the bad and the unaddressed. Proceedings of the 5th IFIP WG 11.9 International Conference on Digital Forensics, January 26-28, 2009, Orlando, FL., USA., pp: 17-36.

Cohen, M. and B. Schatz, 2010. Hash based disk imaging using AFF4. Digital Invest., 7: 121-128.

Garfinkel, S.L., D.J. Malan, K.A. Dubec, C.C. Stevens and C. Pham, 2006. Disk imaging with the advanced forensic format, library and tools. Proceedings of the 2nd Annual IFIP WG 11.9 International Conference on Digital Forensics, January 29-February 1, 2006, Orlando, FL., USA.

Halboob, W., K.S. Alghathbar, R. Mahmod, N.I. Udzir, M.T. Abdullah and A. Deghantanha, 2014. An Efficient Computer Forensics Selective Imaging Model. In: Future Information Technology, Park, J.J., I. Stojmenovic, M. Choi and F. Xhafa (Eds.). Springer, Berlin, Germany, pp: 277-284.

Kenneally, E.E. and C.L. Brown, 2005. Risk sensitive digital evidence collection. Digital Invest., 2: 101-119.

Kloet, B., J. Metz, R.J. Mora, D. Loveall and D. Schreiber, 2008. Libewf: Project info. http://www.uitwisselplatform.nl/projects/libewf/

Richard III, G.G. and V. Roussev, 2006. File system support for digital evidence bags. Proceedings of the IFIP International Conference on Digital Forensics, January 29-February 1, 2006, National Center for Forensic Science, Orlando, FL., USA., pp: 29-40.

Roussev, V. and G.G. Richard III, 2004. Breaking the performance wall: The case for distributed digital forensics. Proceedings of the Digital Forensics Research Workshop, August 11-13, 2004, Baltimore, MD., USA., pp: 1-16.

Sanderson, P., 2006. Mass image classification. Digital Invest., 3: 190-195.

Stuttgen, J., 2011. Selective imaging: Creating efficient forensic images by selecting content first. Master's Thesis, University of Mannheim, Mannheim, Germany.

Stuttgen, J., A. Dewald and F.C. Freiling, 2013. Selective imaging revisited. Proceedings of the 7th International Conference on IT Security Incident Management and IT Forensics, March 12-14, 2013, Nuremberg, Germany, pp: 45-58.

Turner, P., 2005a. Digital provenance-interpretation, verification and corroboration. Digital Invest., 2: 45-49.

Turner, P., 2005b. Unification of digital evidence from disparate sources (digital evidence bags). Digital Invest., 2: 223-228.

Turner, P., 2006. Selective and intelligent imaging using digital evidence bags. Digital Invest., 3: 59-64.

Turner, P., 2007. Applying a forensic approach to incident response, network investigation and system administration using digital evidence bags. Digital Invest., 4: 30-35.