Annual ADFSL Conference on Digital Forensics, Security and Law 2010

May 19th, 10:30 AM

A Framework to Integrate the Data of Interview Investigation and Digital Evidence

Fahad Alshathry

De Montfort University, Software Technology Research laboratory (STRL), Leicester, UK, famalsha@gmail.com

A Framework to Integrate the Data of Interview Investigation and Digital Evidence

Fahad Alshathry
De Montfort University
Software Technology Research laboratory (STRL)
Leicester, UK
famalsha@gmail.com

ABSTRACT

The physical interview process in crime investigation produces an extremely large amount of data, particularly in big cases. In comparison, examiners of digital evidence have enormous amounts of data to search through whilst looking for data relating to the investigation. However, the links between their results are limited. Whilst investigators need to refute or support their hypothesis throughout, digital evidence examiners often use search based keywords. These keywords are usually created from evidence taken from the physical investigation reports and this basic method has been found to have many shortcomings and limitations. This paper proposes a highly automatic framework to integrate anything suspicious that victims or witnesses have said in their interview with their digital data. The proposed model applies to both physical crime investigations and digital evidence examinations.

Keywords: Computer Forensics, Digital Evidence Analysis, Crime Investigation**.**

1. INTRODUCTION

One of the most frequently discussed and emerging topics in law and criminology is computer forensics. Computer forensics is one of several expressions about whole procedures that deal with digital evidence when involved with digital crime (e.g. hacking) or physical crime (e.g. terrorism). It supports the investigation hypothesis to get answers for these five questions: what, where, when, why and who did the crime.

The most critical phase in both digital crime and physical crime is the data analysis. A number of analysis tools are available in the market; however, they are designed for **digital crime** and use simple matching techniques. Consequently "the investigator becomes inundated with data and wastes valuable investigative time scanning through noisy search results and reviewing irrelevant search hits"[1]. Thus, "existing general purpose computer forensic analysis tools are rapidly becoming inadequate for modern analysis workloads"[2].

Although a number of models in digital crime are introduced, the physical crime investigation requirements are diverse and should have a specific framework to compare the conclusion of field investigators and the content analyst regarding the psychological characteristics and motives of suspects of concern.

There are several reasons why this framework is needed. Firstly, "the majority of digital evidence that is processed by law enforcement today is on computers used as instruments of traditional crimes"[3]. Secondly, these days most people have at least one form of digital storage e.g. cell phone. Historically, the amount of physical crime is enormous compared with digital crimes which have come to the fore in the last two decades. In addition, *it is possible to reduce digital crime by improving security applications and awareness, but preventing criminals using computers as instruments in physical crime is impossible.* Finally, when a number of pieces of digital evidence have been involved in a traditional crime, particularly in big cases, questions emerge: *who is supposed to investigate it or lead the investigation team? A person whose background is criminology (a normal investigator) or whose background is computing (a technician)?* While the analysis in traditional crime depends on

disciplines that have continued for hundreds of years, using computers in physical crime has only arisen recently. Therefore, confusion among technicians and normal investigators is common and has led to a lowering of the standards of proficiency.

Consequently, the loss of correlation between the statements of the accused or victims and their digital evidence data possibly generates truncated evidence which does not support one another. This can be particularly obvious when the case is manipulated by numerous investigators, different agencies or when the media is examined by different technicians. This issue is repeated in questions and recommended research in number of the Digital Forensic Research Workshop (DFRWS) 2001[4], 2004 [5] and 2005 [6].

The objective of this work is to introduce a framework to integrate these two types of evidence and to improve the coordination between the two investigatory teams.

This paper is organized as follows: the second section describes the methodology; the third section is the research hypothesis. The fourth section is the proposed investigation framework. The fifth section shows a scenario. The sixth section depicts related work. The seventh section presents the discussion. The final section gives the conclusion and future work.

2. METHODOLOGY

The research area is very critical particularly in the collection of the data. Therefore, an ad hoc methodology is applied. To adapt to the issue; however, we divided our research plan into three phases. The author worked at a computer forensic research centre for years. Throughout this experience, he observed the process of an organization technique when dealing with digital evidence. He found that the gap between investigators and technician examiners is a serious problem. It is considered to be the main cause which prevents the investigators from using digital evidence in the investigation process. Therefore, this experience is taken as a valid base for the research. Secondly, a comprehensive revision of previous studies in the research domain is conducted looking for a solution in the literature. Thirdly, the framework has been realized and is currently being tested.

3. THE HYPOTHESIS

The examiner in computer forensics formulates a hypothesis to start the analysis. This hypothesis relies on what the investigator needs. A common search technique is based on keywords, or metadata. However, physical investigation is more than a search by keywords, and requires an accurate observation of the psychological behaviour throughout numerous questions, e.g. does X work on his computer and when? Does X navigate the internet and why? What the user wrote and why? Does X have a relationship with Y and at what level? And for how long?

This work assumes that digital evidence is a number of activities which have been added and accumulated in the past in digital storage. These activities may show the users psychology, behaviours, plans or social relationships. This work, also assumes that the physical crime investigation process is like a spiral cycle. Throughout this cycle, the investigator attempts to answer hypothetic questions. Of course, from this perspective, the entities (e.g. involved persons) in a case profile are increasable. Hence, the investigation continues until the answers are achieved. We assume, also, that when the technician could correlate these entities and activities, answers may be provided to the investigation questions. However, computers may contain thousand of files and include hundreds of errors in various ways, or even have several parts which may be missed. Therefore, a number of questions may be left without answers; the users of digital evidence who are the suspects, victims, or witnesses, of course, have these answers. In criminology, the fastest way to extract these answers is through interrogation. Although investigation data and digital evidence may answer these questions, the barrier is how can we integrate these two sets of data? However, when we achieve this, it is possible to apply this technique to digital crime investigation as well.

4. PROPOSED INVESTIGATION FRAMEWORK

This model divides the whole process into four main blocks: Crime Scene Phase, Inspection Phase, Interview investigation Phase and Integration Phase. Any main phase has sub-phase[s] as shown below: (see Figure 1)

4.1 Crime Scene Phase

Although this phase is not fully technical with the exception of some circumstances, it provides a solid base for other phases. This phase exists in most current models and has sub-phases explained below:

- **Identifying an Incident and Crime Scene**

The investigators receive the initial information about the incident and issue a written search warrant. The level of formal authorization varies considerably depending on the crime type and its level. At this stage, all equipment, and investigators assistants must be prepared. Although identification is not a technical task, it forms the basis of the case and if there is a mistake the case will have weaknesses which may destroy the legality of the evidence.

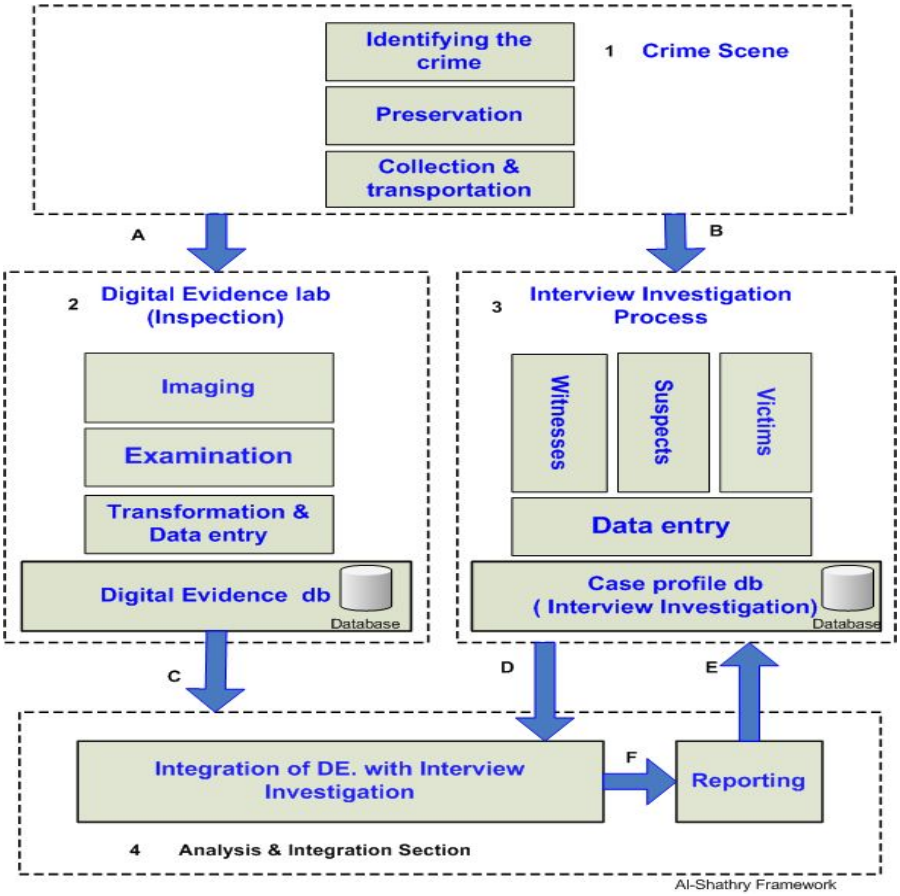

Figure 1 the proposed framework

- **Preservation of and Surveying a Crime Scene**

Preservation is the activity in a crime scene to protect the evidence that appears to be interesting for investigation. However, because digital evidence is changeable and can be modified, this requires extra care, e.g. investigators take the hash value then do a full copy image of the irremovable media immediately to avoid possible spoilage. In addition, they document every action by taking many

pictures of evidence; determining their location and using labels which are related to each other, describing the computer, its manufacturer, the data storage, its serial numbers, its interface and so on. Once the evidence has been secured, the [normal] investigator makes a survey and formulates hypotheses to explain what happened. Although the investigator may suppose a hypothesis relating to the visible objects, it is impossible to correlate that with digital evidence until the data has been extracted and displayed from the hardware.

- **Collection and Transportation**

This phase is not only dedicated to digital evidence, but also includes people who may be involved in the crime. The people, according to their relationship with the crime, may be sent to hospital, to their homes or to the investigation department. On the other hand, digital evidence would be sent to the digital evidence laboratory. The collecting and transporting of digital evidence requires extra attention to prevent any loss or damage. Furthermore, it is important to insure that transport does not affect the integrity of the evidence.

4.2 Inspection Phase

This phase is fully technical; therefore, people who work in this phase must have good experience in computing e.g. file systems, data structures, and encryption. However, they should not engage in data comparison or entities relationships. This phase consists of three sub-phases explained as below:

- **Bit-stream Imaging**

The best evidence is the original storage media. However, keeping the original evidence in the laboratory is not permitted. The insight behind the operation on a full copy of evidence is to keep the original without any modification. Therefore, getting the hash value is significant to ensure that evidence is not modified. Imaging means a bit-by-bit copy of the original evidence. In other words, the exact data stored in the duplicated media must be the exact data in the original. Once the image is performed, the original evidence should be kept in a secure place out of the laboratory building.

- **Low Level Data analysis**

In this sub-phase, data analysis is involved with manipulating all text data to be active for the system and readable for the examiner; that includes recovery of deleted files, hidden data, decryption of encrypted data and translation of human language if different. Data analysis that belongs to a network layer should be illustrated in this phase as well, and then the results should be sent back to case profile. In addition, whole picture files, video and audio should be given comments explaining their contents e.g. if a persons name or mobile number has appeared in a video file, these entities must be extracted into a text file and stored as analyst comments.

- **Data Transformation and Store**

Once the file is readable as text, it should be stored in a digital evidence database. It should have a few details about the case such as the digital evidence's owner, the crime description, and the information that is required. Therefore, the evidence data should be linked with these details and stored as object files with their hashes inside a directory. Authorization to modify this data is restricted. Users are allowed to read-only the data and copy files when needed. To enable search and comparison later, the file format of all files should be illustrated in this phase. The text file format should be stored as a plain text.

4.3 Interview investigation Phase

Interview investigation (or interrogation) requires correlating all entities with the activities that have been gathered about the persons involved as well as at the crime scene to determine suspects. In addition, it requires documentation for any action, i.e. questions and answers. This data is required to be filtered and analyzed to explore any strange and relevant phrases or words. Therefore, the proposed system requires that the data should be stored in a database, to automate the investigation process and

to allow information to be compared with other resources.

4.4 Integration Phase

This phase is central to the whole approach of the solution when we can determine the two targets, interview investigation as digital texts and digital evidence data, and integration will be possible. The matching of these sides should be presented in a view screen with referencing of each hit and word context to allow an analyst to give their technical opinion as comment. The advantage of our system is the dynamic in exchange information between investigators and technicians. This phase consists of three sub-phases as explained below:

- **Text Mining**

In this sub-phase, it is expected that the system shows not only similar words, patterns, but also it should present more inferences, e.g. are these patterns significant for investigation, is this word a persons name or does it mean something different. This technique is called semantic text mining.

In addition, the system uses a number of techniques to extract the knowledge such as machine learning, text chunking based on entities and their context, association rules, clustering entities and ontology. These techniques allow users running a number of representations (e.g. generating a series of questions automatically, presenting Information visually (e.g. entities with relationships).

The proposed system has two approaches to present its results. The first form is dedicated for the normal user, who wants to run the search manually. The second form is to show the dynamic search results. Using the dynamic search shows any matches in real time during the interview investigation.

- **Decision Support**

In respect of crime detection, the decision in crime investigation is very sensitive. As a consequence, the system results are not considered as the final decision until they are validated by an analyst. However, time is significant in an investigation and the expected number of hits is enormous. Thus, the system should provide an abstract level of decision to allow the analyst to start commenting on hits that have high priority rather than hits that may be further from the core profile of the case. There are two methods to evaluate matching words to assess if they are significant or not: 1) comparison with a lexical of suspicious words and phrases (e.g. hate words) judged by their context; 2) the case history data is used as the foundation of investigation judgments and it is used widely in a normal process; hence, the system decides the entities that are recorded in the case history database as significant words for investigation.

- **Comment and Reporting**

Throughout the text analysis, the expert checks all hits and adds the value of hits into the interview investigation database as new questions. These questions should be answered by the suspects, victims or witnesses. Their answers will be searched again until the hits are completed. The comments may also be sent back to the analyst to provide an explanation for any ambiguous technical points. The system stores comments in the digital evidence database as new events associated with the case profile; the events are shown to the technician to allow him to replay their answers. Therefore, the reporting stage is considered like a switch for the results.

5. SCENARIO

This case shows how to apply the framework to the analysis of digital evidence and the integration of its data with the interview investigation. Because it is very difficult to present a real case, the case study is fictional, but is based on frequent cases, which have occurred when a crime has been associated with digital evidence.

A suspicious car was being tracked by an investigator before it had an accident and the driver escaped. The investigator found a computer laptop, three CD’s and two small locks of hair inside the car.

In addition, a witness called (Wit1) described the driver at the accident scene.

Three persons were involved in a drug case number 6544x three weeks ago. This case is considered to be a class C drug activity, and there are two persons who have already been arrested. These people are mentioned as: Sus1 and Sus2. They only know the third person by his photograph because he uses a nickname. According to the witness' statement, the investigator thought that the suspicious car was being driven by the third person. Therefore, it is necessary for the driver to be arrested and sent to the counter-drug section, to give his statement about the accusation.

Consequently, the digital evidence was sent to the laboratory to be examined along with a letter containing these details: the car plate number, owner's name, case number in addition to MD5 values[1] of the evidence and their size. The investigators wait to provide them with any information that can supply further knowledge to the case or to the suspect's relationship to the other citation.

Because the two accused as well as the car owner have already given their DNA and the results have been saved in the database at the forensic laboratory it was possible to show that the DNA analysis of the two little locks of hair were similar but did not match with these persons. The time was very critical after the investigation received a threatening letter.

5.1 Computer Forensic Process (Using traditional Framework)

After the examiner had taken a bit-by-bit image, he began to analyze the file format of the evidence and to recover deleted files for laptop data. The size of its hard drive was 120GB and real size was 81GB. Then, the examiners created all entities based on the provided information in the case and reported these keywords i.e. Sus1, Sus2, Wit1, the driver's nickname and the plate number.

The examiner used a known tool to extract these entities. He created these as keywords manually. Creating keywords manually, of course, requires time, and the creation of wrong words is common. After he had run the search, thousands of hits were matched and presented in the examiners computer and had to be examined carefully. Checking these entities took around three days.

5.2 Computer Forensic Process (by Proposed System)

Once the examiner had completed the imaging, during which he analyzed file format and recovered deleted files, the data was sorted and delivered to his assistant. His assistant transformed the data into text file format by an ad hoc tool and then saved these files in the digital evidence database as unstructured plain text data. Interview investigation with suspects was started immediately. All questions and answers were stored in the investigation database in short order.

The system (proposed system) is running dynamically. Therefore, it started extracting whole entities and making the integration. The system presented the matches with some supporting information concerning its relevance. The text analyst reviewed the results, made comments and then sent them to the investigator as new questions to be put to the suspects. Nearly, all human names are recognized and all strange words are determined.

5.3 Observation

Although the technical methodologies in both frameworks were similar (i.e. thousands of hits were produced in both methods), the time and way of processing are different. In the normal framework, the examiner did not know what was significant for investigation. Examiners were under stress after the investigators asked the laboratory supervisor to speed up the results because they had received a letter threatening that a man involved with drugs may be planning an imminent attack. Therefore, the examiners decided to send a report every hour. This report included hundreds of pages written in MS-Word. Both the examiners and investigators could not deal with this situation because system integration was missing. Therefore, a number of examiners and investigators had to combine to work as one team. This operation took more than 72 hours of hard work but within 48 hours of receiving the threatening letter a drug fighter was killed. After 62 hours of hard work, this team could integrate a

[1] MD5 is a hash function used in a wide variety of forensic applications to check the integrity of files.

deleted Doc. file containing the full name of the driver.

By using the proposed framework, the process runs smoothly, i.e. there is no need to search for people's names, the system extracts them automatically; there is no need to create keywords to start the search, and the search is working dynamically. Therefore, the prop-osed system is highly recommended to deal with these types of crimes.

6. RELATED WORK

Despite there being several digital forensic frameworks that have been proposed and have shown the importance of analysis in computer forensics, to date, there are no appropriate results covering the integration between the technician and normal investigators in physical crime. For instance, although Casey and Palmer model [7] gives full details to follow, they did not give a demonstration for their proposed steps. DFRWS[4] report does not explain the model steps in full detail particularly the analysis step. Reith, et al model [8] is considered as an extension to the DFRWS model and the approach is a linear process. However each phase cannot send a feedback to the previous phase; and that is inappropriate for physical crime investigation when involved with digital evidence. Finally, Carrier and Spafford model [9] uses the physical crime process as a successful methodology investigation, however, the results should be passed through 17 phases, which might lead to a time delay. Also, they did not mention how the analysis is going to be accomplished. Particularly, these models do not mention the integration of interview investigation or interrogation with their steps regarding digital evidence. In addition, by reviewing the literature, little data was found on the association between interview investigation and digital evidence. Therefore, this study sets out with the aim of assessing the importance of integration between the data from interrogation and digital evidence.

Casey 2nd Model [7]	**DFRWS Model** Palmer[4]	**Reith, et al Model** [8]
…	Identification	Identification
Identification		
Preservation, recovery	Preservation	Preparation
Harvesting, reduction	Collection	Approach Strategy
Organization and search	Examination	Preservation
Analysis	Analysis	Collection
Reporting	Presentation	Examination
Persuasion & testimony	Decision	Analysis

Table 1 phases of principle models

7. DISCUSSION

This (on-going) research proposes an automated framework for the integration of the information obtained in the non-digital phase of investigation with the analysis of digital evidence with the aim of improving the quality of digital evidence analysis and interview investigation process.

It uses the advantages of these models[4, 7, 8] as a valid base to build the framework, particularly in the crime scene phase and the examination phase. While these two initial phases are similar in digital crime and physical crime, the investigation hypothesis and the requirement evidence are different. For instance, the analysis phase in this framework has been specified for mining the data to extract entities (e.g. person names, location, organizations) text clustering and linking entities. Thus, the integration

phase is central.

The contribution of this framework can be used for extracting knowledge in different forms e.g. generating questions, presenting entities visually by applying ontology, give reasoning about events by applying association rules, analyze psychological characteristics and motive using cognitive behavior or the creation of automatic keywords based on crime domain. Thus, it will improve the results of Casey model [7] DFRWS Model [4] and Reith, et al. [8] if we add the integration phase (of our framework) instead of the analysis phase in these models.

8. CONCLUSION AND FUTURE WORK

Computers have been involved with crimes on two sides, as a communication tool in digital crime or as an instrument tool (data storage and organizer) in physical crime. Investigation in both of these types is complex. Although there are no major differences in the examination of computers in both of them, the investigation length, investigation methodology, and analysis type are variant. This paper proposes a high automatic framework to integrate the conclusions of the field physical investigators and content analyst of digital evidence and improving the coordination between these two investigatory teams.

The introduced integration model makes the investigation process straightforward. Simultaneously, it means that several cases cannot be handled easily. One such case arises when the file formats are not supported to transformation or data contains text files for two human languages. Another challenge is when we want the system to decide the relationship between words as human names; it requires pre-processing to compare them with other databases such as dictionary names using AI. All these tasks require deep analysis as well as testing on real data. The research of these requirements and other possible improvements of the model presented in this paper are left for future work.

9. REFERENCES

[1] N. L. Beebe and J. G. Clark, Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, Digital Investigation, 4S, , S49, 2007

[2] D. Ayers, A second generation computer forensic analysis system, Digital Forensic Research Workshop (DFRWS). 2009 http://www.dfrws.org/2009/proceedings/p34-ayers.pdf

[3] T. Stallard and K. Levitt, Automated analysis for digital forensic science: semantic integrity checking, Computer Security Applications Conference Proceedings. 19th Annual, 2003

[4] G. Palmer, A Road Map for Digital Forensic Research, Report from the first Digital Forensic Research Workgroup, Utica, New York, Volume 8, 2001.

[5] DFRWS, Workshop Report and Findings, The Digital Forensic Research Workshop Baltimore, Maryland, 2004

[6] DFRWS, Workshop Report and Findings, Digital Forensic Research Workshop, New Orleans, 2005

[7] E. Casey; and G. Palmer., The investigative process, in Book, Digital Investigation and computer crime, 2004.

[8] M. Reith, C. Carr, and G. Gunsch, An Examination of Digital Forensic Models, International Journal of Digital Evidence, 3, 2002.

[9] B. Carrier and E. H. Spafford, An event-based digital forensic investigation framework, Digital Forensic Research Workshop, 2004.