STATISTICAL BYTE FREQUENCY ANALYSIS FOR IDENTIFYING JPEG FILE SEGMENTS

NUR FASIHAH BINTI ABDUL KADIR

A project report submitted in partial fulfillment of the
Requirements for the award of the degree of
Master of Science (Information Security)

Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2015

This project report is dedicated to my family for their endless support and encouragement.

ACKNOWLEDGEMENT

In the name of Allah Most Merciful and Most Gracious. All praises to Allah the Almighty that enabled me to complete my work in a good way. I would like to express my sincere appreciation to my respected supervisor **Prof. Madya Dr. Shukor Abd Razak** for his guidance, support and careful review was valuable. I really benefited professionally and intellectually while working with him. More importantly, I have benefited tremendously from his broad range of experience, technical insights, vision, inspiration and enthusiasm for research. In fact, I appreciate his care, appreciation, directions, patience and dedication in constructively criticizing my research work and thesis as well.

On a final note, I would like to say thanks to my teachers **Dr. Hassan Chizari** for being very cooperative, supportive and guidance in my research work. In the last but the not the least Mr. Junaid Akhtar (UTM), for his support, encouragement and constant efforts which has been my source of patience, inspiration, motivation, courage and strength.

ABSTRACT

File carving is a file recovery technique based on file structure, without the assistance of file system metadata. The important concern here is how file recovery can take place for the file segments that cannot be linked to an existing image header. This project focuses on identifying JPEG file format in hard disk storage. Digital images are broadly used in most industries. It plays a vital role in advertising, education, filming activities, etc. In business world, an image acts as an instant communication to present products and services promptly to the market. Rapid advancements in image processing technology make images more interactive and more modifiable to comply with particular preferences. However, this kind of adjustment will disturb the originality of the raw data. In previous works, researchers mostly focused on recover file segments with assistance of file markers which sometimes might be corrupted. Thus, the statistical byte frequency technique is proposed to provide alternative to address the limitations. In this study, the proposed solution was evaluated based on the accuracy and efficiency performance in identifying the distributed segments. The simulation process involved four different JPEG files format. The simulation indicates that the proposed technique gives a better performance for the files to be carved, in term of accuracy. During the simulation, most of the segments are identified with small gap between all four JPEG files format. The results are gained from k-mean clustering evaluation tool. For computational speed, it takes shorter response time to find file patterns which might due to less number of file segments. The results might be helpful for future reference in file carving program.

ABSTRAK

Identifikasi fail adalah teknik carian berdasarkan struktur dalaman fail, tanpa bantuan dari sistem fail metadata. Perkara yang menjadi perhatian penting adalah bagaimana identifikasi fail mengambil tempat dalam situasi ketiadaan struktur header. Projek ini memberi tumpuan besar dalam mengenal pasti format fail JPEG dalam simpanan cakera keras (hard disk). Imej digital digunakan secara meluas dalam sebahagian besar industri. Ia memainkan peranan penting dalam pengiklanan, pendidikan, aktiviti perfileman dan lain-lain lagi. Dalam dunia perniagaan, imej bertindak sebagai saluran komunikasi serta merta bagi menjana produk dan perkhidmatan secara terus kepada pasaran. Kemajuan pesat dalam teknologi pemprosesan imej menjadikan imej lebih interaktif dan mudah diubah suai bagi memenuhi kehendak tertentu. Walau bagaimanapun, jenis pelarasan ini dibimbangi akan mengganggu keaslian data bagi sesebuah fail. Oleh yang demikian, teknik statistik kekerapan fail adalah dicadangkan untuk menyediakan ruang alternatif bagi menangani limitasi yang wujud. Dalam kajian ini, penyelesaian yang dicadangkan telah dinilai berdasarkan ketepatan dan kecekapan prestasi dalam mengenal pasti serpihan fail yang teragih. Simulasi ini melibatkan empat buah JPEG fail yang berbeza. Hasil kajian menunjukkan bahawa teknik yang dicadangkan menunjukkan prestasi yang lebih baik, dari segi ketepatan operasi. Simulasi menunjukkan sebahagian besar daripada serpihan JPEG berjaya dikenal pasti tetapi dengan jurang bacaan yang kecil di antara keempat-empat file JPEG. Keputusan ini dinilai menggunakan kaedah penilaian secara kelompok, k-mean. Bagi kelajuan pengkomputeran, ia mengambil masa tidak balas yang singkat dalam mengenal pasti variasi fail kesan daripada bilangan serpihan fail yang dianggap sedikit. Dapatan yang diperolehi dijangkakan berguna untuk rujukan di masa hadapan bagi program fail identifikasi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview of the Research

A field of computer science determine image as an exact replica of an object, which stored inside the storage device. Computer storage represent image in a digital format. Digital images play a vital role to the resolution of advertising, education and filming activities. In business world, image roles as an instant communication to present products and services prompt to the market. It shows that images are very useful in most industries. With the high density of image processing technology, make image more interactive to be modified, to comply the certain preferences. However, this kind of adjustment will disturb the originality of raw data.

The raw image can be recognized with minimum processed data on it which is produced by optical device such as digital camera. The raw data is significance in digital investigation process in order to fight for justice towards any criminal activities. In fact, data organization inside storage also reflects the quality of image itself. Image experience “compress” and “decompress” of bit string to save the storage space. Furthermore, it performs data segmentation on chunk space of the memory. The data segmentation phenomenon extends challenge to researchers on

identifying these segmented image files using several potential techniques or signatures.

In this study, further analysis will take place on JPEG file format. JPEG files image is also known as JFIF format. JPEG is a standard image encoding besides GIF, BMP, and PNG etc. It is used widely in digital storage equipment (J van den Bos & Van Der Storm, 2011). The JPEG file format always comes with maximum quality and possesses a larger file size. However, for some purposes such as email, web pages and memory cards, they require small file size. Most of the JPEG files are designed with lossy compression features by purposely eliminate certain percentage of original data to address those demands. In this case, the encoded bytes distribution which belongs to each JPEG file should be different from others due to its unique encoding pattern.

This study highlights the nature of JPEG files and proposes alternative solution to file recovery technique in order to address the files deleted and segmented issue.

1.2 Problem Background

File carving technique is considering three main phases within the process which are pre-processing, collation and same file reassembling, respectively (Pal & Memon, 2009). Pre-processing is a phase to ensure that disk is in their original structure. It involved decompressed and decrypted process if necessary. The reassembling phase is where the proper order of segments is obtained, to find original file. In this phase, it used file header to identify the segmented points.

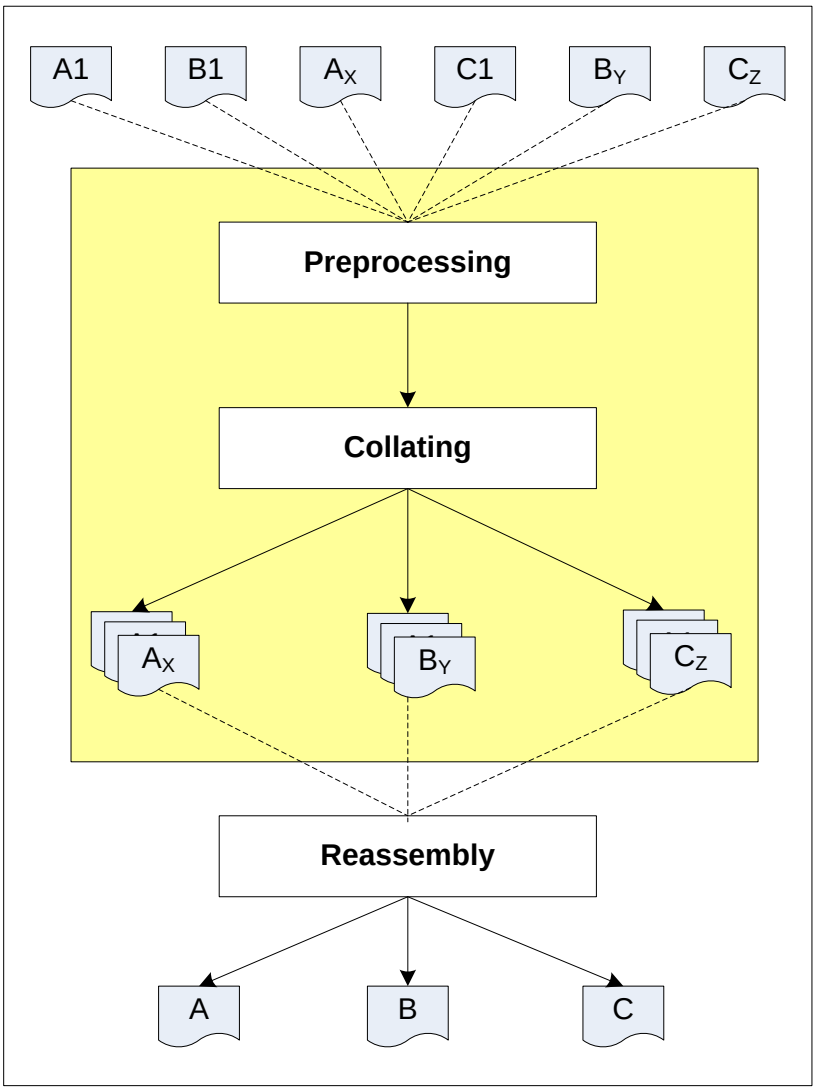

Figure 1.1: Generic structure of file carving technique

This study will focus on collation phase which identifying and classifying different types of segments (refer Figure 1.1). Most of the researchers skipped this phase as they assumed that the segments are already in the same file type. Pixel Matching (PM), Sum of Differences (SoD) and Median Edge Detection (MED) (K. M. Mohamad, Patel, & Deris, 2011; Pal, Shanmugasundaram, & Memon, 2003); these are the approaches of JPEG segment collation. However, those methods require segments to be in the same size, which not always happen. It should be noted that PM, SoD, and MED techniques is not concerned with how much segments are being

stored. It is also a big concern to determine whether two or more segments are within a correct group. This analysis is binding under segment joint validation, which at least requires one Restart (RST) marker.

The available tool related to non-segmented JPEG files recovery is Foremost 0.69 which operates in one pass over disk image and later Scalpel (Roussev, 2005) come to improve the memory usage and performance of Foremost which operates in two passes. Their method capable to support 10 megabytes (MB) reading process of chunk boundary on disk image. Even though Scalpel dominated the performance of Foremost 0.69 but still could not satisfied the accuracy of file carving tool as it does not carve files with missing footers.

Other alternative is done in (Ming & Shule, 2009) by using sequential prediction or sequence validation approach. This method is suitable for the segments that are stored in contiguous arrangement form. However, file segments could be out of order, not sequential or even missing (Merola, 2008). In fact, file carving works well when dealing with non-segmented JPEG files, as the content is not scattered and are still in original form. Other related approaches for segment identification and classification could be obtained in (Cohen, 2008; Garfinkel, 2007; A Pal, H. T. Sencar & N. Memon, 2008; H. T. Sencar & N. Memon, 2009).

1.3 Problem Statement

File carving analysis takes a huge portion in digital investigation. From last few decades, several existing techniques have been tried to deal with deleted files from a hard disk compartment. However, most of them least noted the biggest issues in data carving which somehow deleted files were embedded in a form of segments.

Even though, if they are afforded to render solution for that problem, even then, the carving procedures still depend on file system metadata which does not guarantee to be always available (damaged or corrupted). The researchers need to realize that the damaged or deleted data might be due to intentional behaviour like criminal activities.

1.4 Project Aim

This project aims to provide solution on how to determine deleted JPEG files if the files stored in the segmented form within huge size of hard disk using statistical byte frequency analysis.

1.5 Objectives of the Project

The objectives of this project are:

1. To analyze the existing segmented JPEG file recovery techniques.
2. To propose an alternative technique for identifying segmented segments of JPEG files from hard disk using statistical byte frequency analysis.

1.6 Scope of the Project

The scopes of this project are:

1. This project primarily focuses on JPEG files format.
2. The identification or collation of segmented JPEG files across the hard disk will be the main concern of this project.
3. An extra effort will be undertaken on statistical byte frequency analysis.

1.7 Significant of the Project

This project has the potential to solve some part of file carving techniques. The statistical byte frequency analysis might be possible to address the issue of deleted files images in segmented form. This method will be an alternative solution to the limitations in the existing file carving tools. Those tools can carve data files that are contiguous and also the carvers do not perform extensive and precise validation and create many false positives. Besides that, they tend to depend on file system metadata which sometimes might not available.

1.8 Thesis Organization

The thesis is organized as follows. Chapter 1 highlights the background, problem statement, objectives and scopes of the project. Chapter 2 describes the tools and techniques used to extract the information from image file system based on previous related works. It also highlights the need of such carving techniques in

various case studies. Chapter 3 provides a description of adopted methods. In Chapter 4, researcher reviews the model design of the research. While in Chapter 5, researcher reviews the findings or testing of the model and Chapter 6 concludes the research.

REFERENCES

Bellamy, J. C. (1995). Digital network synchronization. *Communications Magazine, IEEE, 33*(4), 70-83. doi: 10.1109/35.372197

Cohen, Michael I. (2008). Advanced JPEG Carving.

Damashek, Marc. (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. *Science, 267*(5199), 843-849.

Farhood Norouzizadeh Dezfoli, Ali Dehghantanha, Ramlan Mahmoud, Nor Fazlida Binti Mohd Sani, Farid Daryabar. (2013). Digital Forensic Trends and Future. *International Journal of Cyber-Security and Digital Forensics (IJCSDF), 2*, 48-76.

Garfinkel, S L. (2007). Carving contiguous and segmented files with fast object validation. *Digital Investigation, 4*, S2-S12. doi: DOI 10.1016/j.diin.2007.06.017.

Hsiang-Cheh, Huang, Wai-Chi, Fang, & Shin-Chang, Chen. (2008, 13-15 Dec. 2008). *Copyright Protection with EXIF Metadata and Error Control Codes.* Paper presented at the Security Technology, 2008. SECTECH '08.

Krawetz, Neal. (2008). digital image analysis and forensic. 1-43.

McDaniel, M., & Heydari, M. H. (2003, 6-9 Jan. 2003). *Content based file type detection algorithms.* Paper presented at the System Sciences, 2003. Proceedings of the 36th Annual Hawaii.

Memon, N., & Pal, A. (2006). Automated reassembly of file segmented images using greedy algorithms. *Image Processing, IEEE Transactions on, 15*(2), 385-393. doi: 10.1109/TIP.2005.863054.

Merola, Antonio. (2008). Data Carving Concepts.

Ming, Xu, & Shule, Dong. (2009, 18-20 Jan. 2009). *Reassembling the Segmented JPEG Images Based on Sequential Pixel Prediction.* Paper presented at the Computer Network and Multimedia Technology, 2009. CNMT 2009.

Mohamad, K. M., Patel, A., & Deris, M. M. (2011, 20-23 March 2011). *Carving JPEG images and thumbnails using image pattern matching.* Paper presented at the Computers & Informatics (ISCI).

Mohamad, Kamaruddin Malik, & Deris, Mustafa Mat. (2009). Segmentation Point Detection of JPEG Images at DHT Using Validator. 173–180.

Pal, A, Sencar, H T, & Memon, N. (2008). Detecting file segmentation point using sequential hypothesis testing. *Digital Investigation, 5*, S2-S13.

Pal, A., & Memon, N. (2009). The evolution of file carving. *Signal Processing Magazine, IEEE, 26*(2), 59-71. doi: 10.1109/MSP.2008.931081.

Pal, A., Shanmugasundaram, K., & Memon, N. (2003, 6-9 July 2003). Proceedings. 2003 International Conference on *Automated reassembly of segmented images.* Paper presented at the Multimedia and Expo, 2003. ICME '03.

Qiming, Li, Sahin, Bilgehan, Chang, E C, & Thing, V L L. (2011). Content based JPEG segmentation point detection. *International Conference on Multimedia and Expo (ICME), 2011 IEEE*.

Roussev, Golden G. Richard III; Vassil. (2005). Scalpel: A Frugal, High Performance File Carver. *Digital Forensic Research Workshop (DFRWS), New Orleans, LA*, 1-10.

Sencar, H. T., & Memon, N. (2009). *Identification and recovery of JPEG files with missing segments*, Montreal, QC.

Van Den Bos, J., & Van Der Storm, T. (2012) Domain-specific optimization in digital forensics. *5th International Conference on Theory and Practice of Model Transformations, ICMT 2012: Vol. 7307 LNCS* (pp. 121-136). Prague.