A New Approach to Multimedia Files Carving

Weidong Qiu
School of Information Security Engineering
Shanghai Jiao Tong University
Shanghai, 200240, P.R. China
qiuwd@sjtu.edu.cn

Run Zhu
Development Center
China Construction Bank
Shanghai, 200120, P.R. China

Jie Guo, Xiaoming Tang, Bozhong Liu, Zheng Huang
School of Information Security Engineering
Shanghai Jiao Tong University
Shanghai, 200240, P.R. China

***Abstract*—Traditional file recovery methods rely on file system information, which are ineffective when file system information isn't available. File carving is a file recovery method that recovers files according to their structure and content without file system information, which is widely used in digital forensics. As the important carriers of digital information, multimedia files are important digital evidence. In this paper, a new multimedia file carving approach is proposed to improve the recovery accuracy of high entropy file fragments. The fragmented files can be recovered by a hierarchical carving process, including file header identification via entropy, file fragment type classification, and file reassembly via parallel unique path approach. A new file type classification method is constructed based on support vector machine, by using the features of BFD (byte frequency distribution) and ROC (rate of change). Four different datasets, such as DFRWS 2006/2007 challenge datasets, dataset simulating actual disk, dataset with randomly disordered fragments, and dataset with biomedical images, are employed in our experiments. The results show that JPEG recovery accuracy is improved greatly compared with that of PhotoRec tool. Our method performs best in the situation where the order of fragments is completely confusing.**

Keywords—file carving; file fragment; file type classification; SVM; biomedical image;

I. INTRODUCTION

With the development of information technique, the number of digital devices like smartphones and tablets are increasing rapidly. Because of device failure, human error, deliberate destruction and computer criminal activities, files in the digital devices can be damaged. How to recover these damaged files is very important. As the network bandwidth improves, multimedia files are used widely. Multimedia files can carry more information, which is necessary during digital investigations [1].

Traditional file recovery methods are based on file system information, which is useful for continuous saved files. When file system structure is damaged, deleted or not existed, traditional file recovery methods can't work. File carving is a file recovery method which recovers files according to their structure and content without file system information. The method can be divided into two processes: classification of file fragments and reassembly of file fragments. There are three techniques for file type classification: extension-based classification, signature-based classification, and content-based classification [2, 3]. If files aren't fragmented, extension-based classification and signature-based classification perform well by identifying headers, footers and even intermediate blocks with signatures. However, file fragmentation is common. What's worse, blocks can be out of order and file types have no useful signature. Thus, content-based classification is focused. Statistical features are useful for identifying some file types. However, such features are limited to analyze the files with similar statistics. Therefore effective classification methods are needed to improve the classification accuracy. File reassembly is a procedure of detecting a fragmentation point of each unrecovered file and the next fragment's starting point. This process is repeated until a file is built or determined to be unrecoverable. Many existing carving methods fail when undertaking fragments are in non-sequence order.

In this paper, a new multimedia files carving approach is proposed to solve the above limitations. The main contribution of this paper is divided into three parts. Firstly, we propose a new file fragment type classification method based on supervised learning and support vector machines. Both the byte frequency distribution and the rate of change of file fragments are taken into consideration [4]. Secondly, a hierarchical file carving approach is designed. The signature and statistical features are used for the preliminary classification of file fragments. Then, a new file type classification method is applied for depth classification. Finally, the parallel unique path (PUP) approach is implemented for the reassembly of file fragments. The hierarchical carving algorithm can improve the file recovery accuracy effectively. Thirdly, a JPEG carving tool is developed to verify the proposed method. Four different datasets are chosen in our experiments, including a public dataset, a simulation disk dataset, a disordered fragmentation dataset and a biomedical dataset. The first three datasets aim at general JPEG files, while the final dataset undertakes JPEG files with biomedical content. The result indicates that the proposed method can correctly recover JPEG files from datasets without file system information. Our method has proper feasibility and effectiveness.

This paper is organized as follows. In Section 2, a brief overview of related work is given. In Section 3, the proposed method is described. Experiments and results are presented in Section 4. In Section 5, the conclusion is drawn.

II. RELATED WORK

Foremost is the first open-source file-carving tool, which was developed by Kris Kendall and Jesse Kornblum in USA

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.31

Air Force Office [5]. In 2005, the structures of several common file formats were analyzed to improve the performance of foremost [6]. Golden et al developed Scalpel based on foremost [7]. Scalpel performs well when there are signatures in files and files aren't fragmented. PhotoRec can recover various file formats and is widely used to recover files with file system damaged [8]. Files can be carved by searching for a specific header and footer, which is not suitable for fragmented files [9]. Garfinkel introduced a fast object validation technique for the recovery of bifragmented files. Some files don't have signature in the footer, but the length information is available in metadata. So, files can be recovered by using file header and length [10]. Joachim et al. introduced Smart Carving, which has no limitation of file system and the number of file fragments [11]. Some features, such as entropy, keywords and fingerprints, are extracted to classify file fragments. In [12], the file carving is considered to be a process of estimating a mapping function between the bytes stored in the devices and the recovered file. If a recovered file is corrupt, the result is fed back to mapping function generator. In this way, incorrect mapping functions are excluded and the carving process can be improved.

Content-based file carving method is an active area. In [13], three algorithms were proposed, including Byte Frequency Analysis (BFA), Byte Frequency Cross-correlation (BFC), and File Header/Trailer (FHT) analysis. Karresand et al were the first people to study the classification of individual data clusters instead of entire files [14]. They extended the original Oscar method by measuring the rate of change of the byte contents. Li et al. used a support vector machine with Byte Frequency Distribution (BFD) feature vectors to classify high entropy file fragments [15]. Simran et al. applied natural language processing (NLP) techniques for file fragment classification [16]. Amirani et al. deploys principle component analysis (PCA) and neural networks for an automatic feature extraction and classification. However, how to classify high entropy file fragments is a difficult problem and little related work is done [17]. By using statistical approaches like Shannon entropy statistics, the classification of high entropy file fragments was inefficient with low accuracy. So it is still challenging to classify high entropy file fragments.

III. PRELIMINARIES

A. Feature Extraction

1) Entropy: Multimedia files are of high entropy because they are compressed or encrypted. Shannon entropy can be used to distinguish multimedia files from other files [18]. For fragments which are not the header or footer, entropy-based method can identify multimedia fragments preliminarily for follow-up classification.

Each computer byte value in the fragment is between 0 and 255, which we let be i ($0 \le i \le 255$). Let $p(i)$ be the probability of occurrences of byte value I, $N(i)$ be the number of occurrences of byte value I, $length1$ be the size of a file fragment. Then $p(i)$ is described by:

$$p(i) = N(i) / length1 \tag{1}$$

Thus the entropy of fragments can be described by:

$$Entropy = -\sum p(i) \log_b^{p(i)},\ 0 < p(i) \leqslant 1 \tag{2}$$

2) BFD: The Byte Frequency Distribution (BFD) feature is commonly used in file fragment classification. The BFD of a file fragment can be represented by a feature vector. The feature vector consists of 256 basic features and is defined as $v\_bfd\ (b_0, b_1, \ldots, b_{255})$. b_i is defined as:

$$b(i) = p(i) \tag{3}$$

From the above definition, it is seen that the BFD feature only considers the byte value without the order of the bytes.

3) ROC: In order to consider the order of the bytes, rate of change (ROC) was put forward. It is defined as the absolute value of the difference between two consecutive byte values in a data fragment[14]. Let two consecutive bytes in the fragment values to be v_i and v_{i+1}. ROC is described by:

$$ROC = |v_i - v_{i+1}| \tag{4}$$

It is obvious that the byte change of two consecutive values can be positive or negative. Then, the range of the difference value is from -255 to 255. In order to be consistent with the number of BFD feature vectors, rate of change (ROC) is calculated by using absolute value instead of signed value. Let $length2$ to be the size of ROC sequence. Let $N(i_r)$ to be the number of occurrences of ROC value i_r. The ROC byte frequency of the file fragment is defined as $p(i_r)$.

$$p(i_r) = N(i_r) / length2 \tag{5}$$

B. Support vector machine

Support vector machine is selected for its powerful supervised learning technique for data classification. This algorithm is able to predict the file type by using the generated training model. There are four basic kernel functions: linear, polynomial, radial basis function (RBF), and sigmoid. In this paper, the RBF is selected as the nonlinear kernel of the SVM classifier. Let P be the input vector and C_i be the prototype of the input vectors. The output of each RBF unit is as follows:

$$R_i(P) = \exp\left[-\| P - C_i \|^2 / \sigma_i^2\right] \tag{6}$$

Where $\| \ \|$ is the Euclidean norm on input space. σ_i is the width of the ith RBF unit.

C. Parallel unique path

After the type classification, the file fragments of the same type may be stored in disorder. Parallel unique path (PUP) algorithm is a variation of single source shortest path algorithm [19], which can be applied to recover the order of file fragments.

Suppose that there are k file headers, described as $H\{h_1, h_2, \ldots, h_k\}$. Every file has its reconstruction path Pi, and the file header should be stored in the starting block of the reconstruction path. Let S_i be the current processed block of the ith file. Obviously, S_i is initialized with h_i. The best greedy matches for the current processed blocks are stored in a set T,

where t_i is the best match for S_i. Then, t_i is added to the reconstruction path P_i, and S_i is replaced with t_i. The above process is repeated until all files are recovered. The steps are as followed.

1) Add t_i to ith file's reconstruction path P_i (i.e. $P_i = P_i \| t_i$).

2) Replace ith file's current block in SET S (i.e. $S_i = t_i$).

3) Update the best match SET T for SET S.

4) Choose the overall best matching from SET T.

5) Repeat 1 until all files are recovered.

IV. PROPOSED METHOD

In this paper, an algorithm is proposed to carve multimedia files based on a new fragment classification method, as shown in Fig.1. Firstly, the file header and footer are classified via the signature. Considering the high entropy value of multimedia files, entropy-based approach is adopted for the preliminary classification of the remaining fragments. A supervised learning technique is applied for the accurate classification of file fragments. Support Vector Machine is adopted for its powerful effect in the field of supervised learning. After fragments classification, PUP algorithm is used to reassemble the fragments of the same file type. Through classification and reassembly, a certain type file is recovered.

The detailed steps are as follows:

1) Choose certain file type as target file type to recover.

2) Determine if it is the end of original disk image. If it is, then go to step 9. Otherwise, a block of given size is read from image.

3) Determine whether the block is the file header based on the signature of target file header. If it is, go to step 8. Otherwise, go to step 4.

4) Determine whether the block is the file footer based on the signature of target file footer. If it is, go to step 8. If not, go to step 5.

5) Calculate the entropy of the block. If the entropy of the block is larger than the deadline, the block is judged to be multimedia block. Then, go to step 6. Otherwise, go to step 2.

6) Calculate the BFD of the block and predict the block's type base on the SVM training model. If the predicted type is the target type, go to step 8. Otherwise, go to step 7.

7) Calculate the RoC of the block and predict the block's type base on the SVM training model. If the predicted type is the target type, go to step 8. Otherwise, go to step 2.

8) Add current block to the target dataset.

9) Use PUP algorithm to reassemble all the blocks in the target dataset. At last, the target type files are recovered.

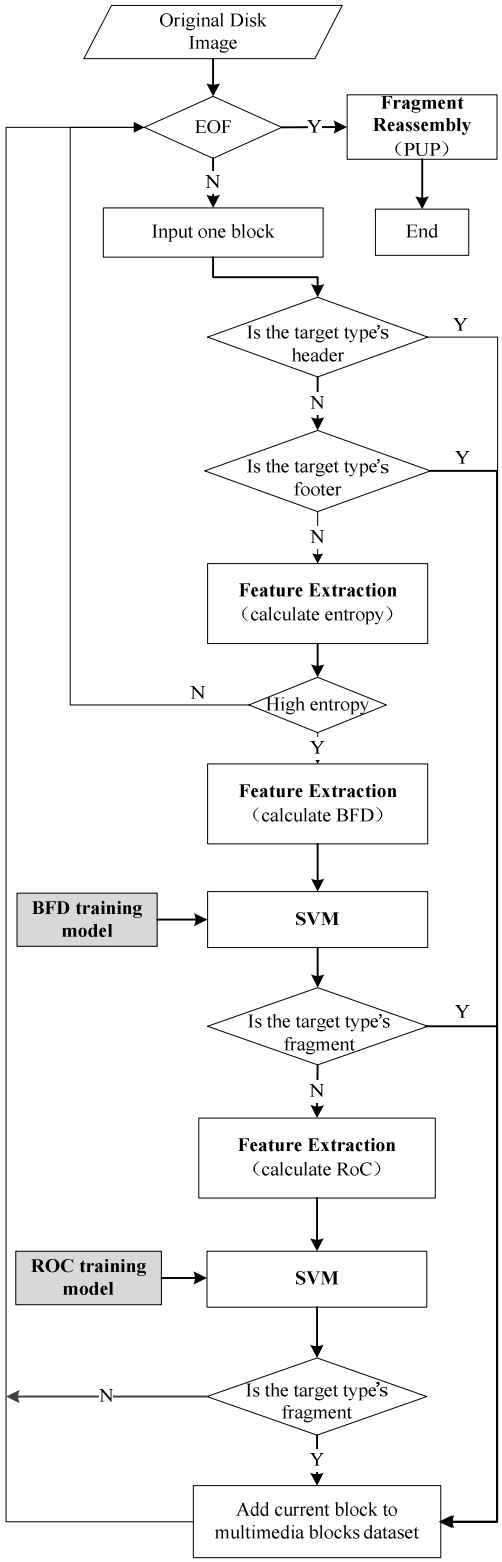

Fig.1. Multimedia Files Carving Algorithm

V. EXPERIMENT

A. Datasets

In order to evaluate the performance of our scheme, four different datasets are chosen in our experiments. The target file type is JPEG, but there are other file types in the datasets, including PNG, XML, HTML, PDF, GZ, ZIP, Office, MP3, TXT and so on. The four datasets are as follows:

1) Public dataset1: The datasets are selected from DFRWS 2006 Forensics Challenge and DFRWS 2007 Forensics Challenge. DFRWS 2006 Forensics Challenge file whose size is 50M includes non-fragmented JPEG and fragmented JPEG in sequential order. DFRWS 2007 Forensics Challenge file whose size is 256M includes fragmented JPEG in non-sequential order and incomplete JPEG missing fragments. As our method aims at recovering complete files without missing fragments, we don't focus on this scenario. The number of JPEG files in these two challenges is 28.

2) Dataset 2: R. Poiseland and S. Tjoa use a script to simulate the actual image saved in the disk [1]. We use the script to generate our dataset, including JPEG, PNG, XML, HTML, SVG and TXT.

3) Dataset3 with disordered fragments: The file types in the dateset3 include JPEG, PDF, GZ and PNG. When generating this dataset, we divide all files into fragments with the size of 512 bytes. The last fragment may not be 512 bytes, but we still treat it as 512 bytes. The dataset is created by reading fragments randomly from files of different type. The order of the fragments in the dataset is unpredictable. Compared with above two datasets, the order of the fragments is more random.

4) Dataset4 with biomedical images: In this part, we focus on the JPEG files with special content. Biomedical images are with high noise and low SNR (Signal to Noise Ratio), which is difficult to be classified and reassembled. The fragments are completely disordered and the file types in the dateset4 include JPEG, PDF, GZ and PNG.

TABLE I. SPECIFICATIONS OF DATASETS USED IN EXPERIMENTS

File Type		Number of files
JPEG	*Dataset 1*	28
	Dataset 2	6
	Dataset 3	11
	Dataset 4	11
Other type files		254

Table I shows the distribution of testing files. The number of JPEG files in each dataset and the number of other type files in all datasets are indicated. In [4], the SVM training dataset are derived from the freely available corpus of forensics research data [20]. We can guarantee that the testing datasets aren't the same files. In this way, the result would not be biased.

B. JPEG File Carver

Take the Dataset4 for example, the steps of our experiment are presented, as shown in Fig.2. There are 11 biomedical images in Dataset4. Four PDFs, four GZs, and four PNGs files are added to generate the disordered fragments together.

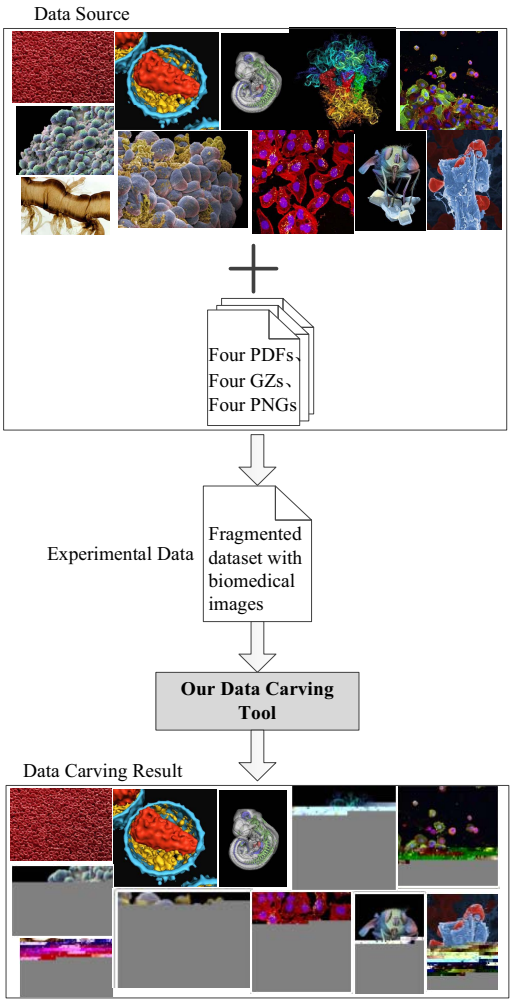

Fig.2. Experimental process of biomedical file carver

Firstly, the header and footer of JPEG files need be determined. The signature of the header is "FF D8 FF" and the signature of the footer is "FF D9". Libmagic library is chosen to access the signatures of JPEG.

Secondly, the file type of the fragments, which aren't header or footer, need be identified. According to our method, the deadline of JPEG entropy value is set to be 5.6. If the entropy of the testing fragment is larger than 5.6, the fragment can be judged to be a suspicious JPEG file. Then, the identification moves onto the next step. As "FF 00" is the signature for JPEG fragments, we then use this byte sequences to identify fragments. If there is "FF 00" in the fragment, the fragment is identified as JPEG fragment. Otherwise, SVM is employed. We choose the LIBSVM package to implement SVM classifier. The version of LIBSVM is 3.16, and we use the python interface on Ubuntu platform. The best parameters have to be chosen for higher classification accuracy. We choose radial basis function as the kernel function and use the grid.py for optimal parameter selection. The feature vectors are

scaled to increase training speed and classification accuracy. Referring to the SVM training model and predicting model [4], the BFD features of fragments are extracted to identify its file type. If the file type is predicted to be JPEG, fragment classification is finished. Otherwise, the ROC features are extracted for SVM prediction. By using SVM predicting model, the fragment classification accuracy can be greatly improved.

At last, all fragments identified as JPEG are collected. Based on PUP method, the fragments are reassembled and all JPEG files are recovered.

C. Results

Many file type fragments are input in our method. As our method is based on type-x recognition and target file type is JPEG, we can only obtain JEPG files. PhotoRec 6.14[8] is a common file recovery software designed to recover different type files with good performance. So we compare our results with that of PhotoRec. For each dataset, JPEG recovery accuracy is decided by the number of all recovered JPEG files. The results are summarized in Table II.

TABLE II. RECOVERY ACCURACYF OF DIFFERENT DATASETS

Dataset	Method	Number of Recovered Files	Accuracy
Dataset 1	***PhotoRec***	21	75%
	Our method	24	85.7%
Dataset 2	***PhotoRec***	2	33.3%
	Our method	5	83.3%
Dataset 3	***PhotoRec***	0	0%
	Our method	10	90.9%
Dataset 4	***PhotoRec***	1	9.1%
	Our method	4.5	40.9%

It is shown that our method performs better than PhotoRec in four situations. In general, the recovery accuracy is improved by no less than 10%. For the image similar to the actual disk image, our method improves the recovery accuracy by 50%. Since biomedical images are similar and difficult to classify and reassemble, the accuracy of our method is 40.9%. It is also better than PhotoRec. In particular, our method works best for dataset 3, whose recovery accuracy is close to 100%. In that case, PhotoRec can't recover any JPEG files. That means our method is good at undertaking the situation where the order of file fragments is more confusing.

As we know, SVM can be very powerful in differentiating data fragments from different types of files. If the training set is effective and parameter values are optimal, the predicting accuracy can be increased. The predicting phase doesn't rely on the order of file fragments and PUP can handle fragments in non-sequence order very well. Thus, in the 3th situation, no matter how confusing the fragments' order is, our method can identify JPEG fragments well and reassemble the disordered fragments.

VI. CONCLUSION

In this paper, a new method to recover multimedia files is proposed. Signatures and entropy are selected as features for fragments classification firstly. Next, BFD and ROC features were extracted for the supervised learning and classification. At last, PUP approach was used to reassemble the fragments of the same file type. The experiments are performed with four different datasets in four different situations. The datasets include public challenge files, simulating files, disordered fragments and biomedical images. Considering its high recovery accuracy, PhotoRec tool is chosen for the comparison of experimental results. It is proven that our method provides better performance than PhotoRec. Our method performs best when the fragments are completely out of order.

Further study is needed to discuss type-all recognition models and missing fragments. With type-all recognition models, many files of different types can be recovered at the same time. By studying the structure of files, files with missing fragments can be recovered. Besides BFD and ROC features, more distinctive features can be defined to improve classification accuracy. Apart from SVM, more supervised learning algorithm can be considered in the future.

ACKNOWLEDGMENT

This paper has been supported by New Century Excellent Talents in University of Ministry of Education under Grant NCET-12-0358, Technology Innovation Research Program, Shanghai Municipal Education Commission under Grant 12ZZ019, the Supporting Program of the "Twelfth Five-year Plan" for SCI and Tech Research of China under Grant 2011BAK13B05 and Shanghai Municipal Science and technology project under Grant 13JG0500400.

REFERENCES

[1] R. Poisel and S. Tjoa, "Roadmap to approaches for carving of fragmented multimedia files," in Availability, Reliability and Security (ARES), 2011 Sixth International Conference on. IEEE, Vienna, pp. 752–757, August 2011.

[2] D. J. Hickok, D. R. Lesniak, and M. C. Rowe, "File type detection technology," in the 38th Midwest Instruction and Computing Symposium, Eau Claire WI, pp. 73-76, April 2005.

[3] M. C. Amirani, M. Toorani and A. Beheshti, "A new approach to content-based file type detection," in IEEE Symposium on Computers and Communications (ISCC '08), Marrakech, Morocco, pp.1103-1108, July 2008.

[4] R. Zhu, W. D. Qiu, Y. M. Li, B. Z. Liu, "Content-based Type Classification of High Entropy File Fragments," in Computer Science and Automation Engineering (CSAE), 3rd International Conference on IEEE, Guangzhou, China, vol. 3, pp. 1345-1349, August 2013.

[5] A. Merola, "Data Carving Concepts," http://www.sans.org/reading_room/whitepapers/forensics/data-carving-concepts_32969, 2008.

[6] N. Mikus, " An analysis of disc carving techniques," Master's thesis, Naval Postgraduate School, Monterey, California, Mar. 2005.

[7] G. G. Richard III, V. Roussev, "Scalpel: A Frugal, High Performance File Carver," in the 5th Annual Digital Forensic Research Workshop (DFRWS'05). New Orleans, Louisiana, pp. 1-10, August 2005.

[8] C. Grenier, “PhotoRec,” http://www. cgsecurity. org/wiki/PhotoRec, 2007.

[9] D. Dickeman, “Advanced Data Carving,” in the DFRW 2006 Data Caving Challenge, 2006.

[10] S. L. Garfinkel, “Carving contiguous and fragmented files with fast object validation,” Digital Investigation, vol.4, pp. 2-12, 2007.

[11] J. Metz, RJ. Mora, “Analysis Of 2006 Dfrws Forensic Carving Challenge,” http://sandbox.dfrws.org/2006/mora/dfrws2006.pdf, 2006.

[12] M. I. Cohen, “Advanced carving techniques,” Digital Investigation, vol.4, no.3, pp.119-128, 2007.

[13] M. McDaniel and M. H. Heydari, “Content based file type detection algorithms,” in System Sciences, Proceedings of the 36th Annual Hawaii International Conference on. IEEE, Washington, DC, USA, pp. 10-20. January 2003

[14] M. Karresand and N. Shahmehri, “Oscar-file type identification of binary data in disk clusters and ram pages,” Security and Privacy in Dynamic Environments. Springer US, pp. 413-424, 2006.

[15] Q. Li, A. Ong, P. Suganthan, and V. Thing, “A novel support vector machine approach to high entropy data fragment classification,” in Proceedings of the South African Information Security Multi-Conference (SAISMC 2010), Port Elizabeth, South Africa, pp.236-247, May 2010.

[16] S.Fi tzgerald, G. Mathews, C. Morris, and O. Zhulyn, “Using nlp techniques for file fragment classification,” Digital Investigation, vol. 9, pp. S44-S49, 2012

[17] M. C. Amirani, M. Toorani, and S. Mihandoost, “Feature‐based Type Identification of File Fragments,” Security and Communication Networks, vol.6, no.1, pp.115-128, 2013.

[18] M. M. Shannon, “Forensic relative strength scoring: ASCII and entropy scoring,”International Journal of Digital Evidence, vol. 2, no. 4, pp. 1-19, 2004.

[19] E. W. Dijkstra, “A note on two problems in connexion with graphs,” Numerische mathematik, vol.1, no.1, pp.269-271, 1959.

[20] S. Garfinkel, P. Farrell and V. Roussev, “Bringing science to digital forensics with standardized forensic corpora,” Digital Investigation, vol.6, pp.2-11, 2009.