Global Science Research Journals

Global Journal of Business and Management

Extended Abstract Open Access

Reassembly and Clustering Bifragmented Intertwined Jpeg Images Using Genetic Algorithm and Extreme Learning Machine

Rabei Raad Ali

UTHM Foundation, Malaysia

Abstract

Statement of the Problem: File carving tools are essential element of digital forensic investigation for recovering evidence data from computer disk drives. Today, JPEG image files are popular file formats that have less structured contents which make its carving possible in the absence of any file system metadata. However, completely recovering intertwined Bifragmented JPEG images into their original form without missing any parts or data of the image is a challenging due to the intertwined case might occur with non-JPEG images such as PDF, Text, Microsoft Office or random data. In this research, a new carving framework is presented in order to address the fragmentation issues that often occur in JPEG images which is called RX_myKarve. The RX_myKarve is an extended framework from X_myKarve, which consists of the following key components: (i) an Extreme Learning Machine (ELM) neural network for clusters classification using three existing content-based features extraction to improve the identification of JPEG images content and support the reassembling process; (ii) a genetic algorithm to reconstruct a JPEG image from a set of deformed and fragmented clusters in the scan area. The RX_myKarve is a framework that contains both structure-based carving and content-based carving approaches. The RX_myKarve is implemented as an Automatic JPEG Carver (AJC) tool in order to test and compare its performance with the state-of-the- art carvers such as RevIt, myKarve and X_myKarve. It is applied to three datasets namely DFRWS (2006 and 2007) forensic challenges datasets and a new dataset to test and evaluate the AJC tool. These datasets have complex challenges that simulate particular fragmentation cases addressed in this research. The RX-myKarve framework is an adoption from the X_myKarve framework. It consists of structure-based and content-based carving approaches as depicted in Figure. The structure- based approach consists of three main components are the Address database (ADB), Validated JPEG Markers (VJM) and Automated Work Queue (AWQ). On the other hand, the content-based carving approach consists of three main components which are Identification, Frame validation and Reassembling. The first component in the structure-based carving approach is the ADB. ADB is a database that is used to store addresses of validated headers in order to be used for distinguishing patterns of the JPEG image/thumbnail(s). The second component in the structure-based carving approach is VJM list. It contains JPEG image and thumbnails markers that are recovered along with their index according to the ADB index. Both components will be used and play important roles in the third component for generating work instructions in the AWQ. The main difference between the VJM list and ADB is that the VJH list is only used to initiate the process in the AWQ. The third component in the structure-based carving approach is AWQ. It works as an automated carving of the structure-based carving approach depends entirely on reading hexadecimal values with the predefined patterns of JPEG image's frame in an AWQ. Also, it can automatically carve a JPEG image with/without thumbnails as shown. Therefore, the AWQ will decrease the time needed to check for reconstruction of the fragmented image in the scan area as well as determine whether fragmentation point in the scan area belongs to the first or second file. The AWQ, it works by reading a validated header addresses from the index store in ADB index and then the maker's name from the index store in the VJM list are accumulated until they match one of the predefined patterns AWQ otherwise it will be discarded. Thus, the AWQ will be generated when a successful match is found. The next step for the RX_myKarve

framework is an image reconstructing process starts, when the AWQ process ends, all accumulated markers in AWQ are decoded into image files to determine the image complete recovered or not. However, thumbnail(s) may also be included in the images decoded in case of complete or non-completed image recovered. Finally, the fragmented images are further processed in the content-based carving approach to address the fragmentation problem which is the final step in RX_myKarve Framework.

As mentioned earlier, a content-based carving approach has three components which are Identification, Validation and reassembling. The Identification is used to find information related to the JPEG image clusters in the scan area. There are two techniques applied in the Identification. The first identification technique the Define Restart Interval (DRI) of the header information to identify the existence of a RSTm in the JPEG image. If the image with RSTm, it used the RSTm pattern as guidance to reassembling the image contents. If the image without RSTm, then the technique removes the clusters that contain RSTm. The second identification technique applies the ELM binary classifications to distinguish between JPEG clusters and non-JPEG clusters. The ELM provides feasible measures for file type identification accuracy by looking at a set of interrelated features or attributes that are extracted from the scan area as explained earlier. The ELM classification technique is performed based on three features which are Entropy, BFD and RoC. The data of the identification along with the extracted data from the Structure-based approach are used to support the validation and reassembling processes. The Validation component combines and checks the collected evidence about the image from different resources. This process removes the conflict in the data and organizes the data to be ready for the reassembling process. The main task of this component is to check the image file in order to guarantee that the file contains all necessary markers and tables to complete the decoding process later on. The reassembling component includes a carving procedure that works based on three conditions. They are the cleaning intertwined and fragmented clusters. The first condition entails cleaning some data in the scan area. This condition removes unknown clusters from the scan area that do not contain relevant decoded image pixel data and thumbnail(s). There are two different techniques to solve the cleaning problem. The first technique analyses the possible occurrence of the RSTm. The second technique computes the actual MCUs of the image specification then decodes the image. The second condition is the existing of intertwined JPEG images. When this condition is satisfied then the reassembling algorithm applies a Coherence of Euclidean Distance matric. The third condition is the existing of non-JPEG clusters or fragmented clusters. It has been mentioned previously that an ELM algorithm classifies the clusters into JPEG-file clusters and non- JPEG-file clusters based on content-based features of entropy, BFD and RoC. This classification identifies the unrelated clusters to be removed before conducting the reassembling operation. The final results show that the AJC with the aid of the RX_myKarve framework outperform the X_myKarve, myKarve and RevIt.

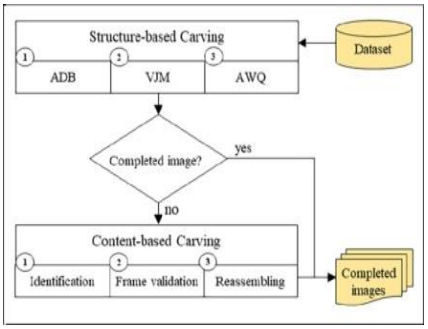

Figure: The flow diagram of the RX_myKarve framework.