Digital Evidence Search Kit

K.P. Chow, C.F. Chong, K.Y. Lai, L.C.K. Hui, K. H. Pun, W.W. Tsang, H.W. Chan
Center for Information Security and Cryptography
Department of Computer Science
The University of Hong Kong

Abstract

With the rapid development of electronic commerce and Internet technology, cyber crimes have become more and more common. There is a great need for automated software systems that can assist law enforcement agencies in cyber crime evidence collection. This paper describes a cyber crime evidence collection tool called DESK (Digital Evidence Search Kit), which is the product of several years of cumulative efforts of our Center together with the Hong Kong Police Force and several other law enforcement agencies of the Hong Kong Special Administrative Region. We will use DESK to illustrate some of the desirable features of an effective cyber crime evidence collection tool.

1. Introduction

Cyber Forensics can be defined as the process of extracting information and data from computer storage media and guaranteeing its accuracy and reliability [1]. Forensic techniques for analyzing paper documents are very well established. However, few of these techniques can be applied to electronic documents because the two types of documents are fundamentally different. There are two main problems inherent with electronic documents that make them more difficult to analyze than paper documents. First, they are easy to copy and modify. If a blackmailing letter is stored as a file on a suspect's personal computer, the suspect may argue that the document was planted into their computer after the computer had been seized by the law enforcement agency. Secondly, it can be argued that the document had been modified by the law enforcement agency. One solution to this problem is to use *special purpose computer forensic software tools* to verify the *file system integrity* of the suspect's computer, after it has been seized by the law enforcement agency. Thus, one of the major goals of a software forensic tool is to ensure the validity and reliability of the electronic evidence. Once it has been proven that the tool has been used properly and in compliance with the Evidence Ordinance [2], the suspect's computer can potentially be used to provide evidence in court.

Another problem introduced by electronic documents is the complexity of different file formats and different file system structures. Electronic documents can be generated by various kinds of application programs: word processors, spreadsheet software, database software, graphic editors, electronic mail systems, etc. Electronic documents are usually stored as computer files in the computer's file system. They can be stored as user files in user directories, or as fake system files in system directories, or even as deleted files. When a file

0-7695-2478-8/05 $20.00 © 2005 IEEE

is deleted, the operating system typically only removes the reference to the file in the File Allocation Table (FAT). Although the reference is removed, the data still remain physically on the disk. Deleted data will remain on the disk until another file overwrites them. It is a time consuming task to inspect every possible storage area of the computer's file system to look for potentially useful evidence. In view of this, another important goal of a software forensic tool is to provide an efficient and automatic search function to search for electronic content that may potentially be used as evidence of cyber crimes.

The cyber crime forensic tool - DESK (Digital Evidence Search Kit) - is designed with these two main goals in mind [3, 4]. DESK is a software system developed by the Center for Information Security and Cryptography, The University of Hong Kong, in collaboration with the Hong Kong Police. Apart from achieving the two above-mentioned goals of providing file system integrity checking and effective search functions, DESK is also specifically targeted at the bilingual environment of Hong Kong. DESK is capable of searching word patterns in both English and Chinese (both traditional and simplified Chinese). To the knowledge of the authors, DESK is the first bilingual (Chinese and English) software in the world to specialize in the investigation and reporting of computer crime evidence. This paper provides a brief description of DESK and in doing so demonstrates the general requirements of an effective cyber crime evidence collection tool.

Section 2 gives an overview of the DESK system. In Section 3 the DESK search functions are described, while Section 4 describes file integrity checking. Section 5 describes case management and Section 6 describes a 'diff' function. Finally, in Section 7, a summary is given.

2. Overview of DESK

An overview of the DESK system during operation is depicted in Figure 1. The DESK machine is the computer used by a law enforcement agent and the subject machine is the personal computer of the suspect. The two machines communicate with each other using a serial (RS-232) cable that connects them together. In addition, a floppy diskette containing a part of the DESK software is used to start up the subject machine. In summary, the hardware components of the DESK system include a DESK machine, which is typically a notebook computer with a serial port, and a floppy diskette used to start up the subject machine.

The software components of DESK consist of the DESK client that is installed on the DESK machine and the DESK server that is contained on the floppy diskette to be run by the subject machine. The DESK client is mainly used to provide a user interface for issuing commands to inspect the subject machine. The DESK server component, installed on the floppy diskette, has additional functionality which includes the following:

- To start up the subject machine. This is done in order to prevent modification of the subject machine's system files that typically occurs when the subject machine is started up by its own operating system.
- To 'lock' the subject machine. This is to protect the subject machine from any accidental corruption by the machine's own interrupts in order to ensure that the content found on the subject machine's file system cannot be modified. This operation is very important since it ensures the integrity of the subject machine while various forensic operations are being performed.

0-7695-2478-8/05 $20.00 © 2005 IEEE

- To provide a simple user interface for simple search operations. The user interface is much less sophisticated than that of the DESK client running on the notebook due to the storage limitations of floppy diskettes.

The main operations of DESK are provide by two software components in the DESK system:

- A text pattern file which contains search keywords, in Chinese and/or English, to be searched for on the subject machine, and
- Hash value databases that contain 'fingerprints' of file systems that enable file integrity verification. More details are given below.

The pattern file and hash value databases can be configured to satisfy the specific requirements of individual forensic investigations. Both components can be modified to cater to different crime scenes. For example, if the original pattern file contains only 5 search keywords, an additional search keyword can be added anytime when necessary.

3. DESK search methods

The first important feature of DESK is the search function. It is used to search for files on the subject machine that contain pre-defined search keywords. Pre-defined search keywords are words that are relevant to a particular crime case. For instance, in a bank corruption crime, the pre-defined text patterns may contain names of different banks. The patterns can either be in English or in Chinese, or combinations of both. For Chinese patterns, different encodings of Chinese, such as Big5, GB (2312) and Unicode UTF16, are supported.

There are three main kinds of search operations: physical search, logical search, and deleted file search. Physical search performs a search of the patterns of each physical sector of the subject machine's storage system. By using a physical search, cyber crime evidence purposely stored in unused sectors in the storage system can be discovered. Moreover, it provides a way for searching files independent of the specific file system. The disadvantage is that physical search, due to its lack of knowledge about the file system, can only search data within individual physical disk sectors.

Logical search makes use of the information about the file system. Conceptually, a file is a continuous sequence of bytes and the file system takes care of placing portions of the sequence into different sectors (not necessarily contiguous) while maintaining the logical contiguity of the contents of a file. A file can have a size larger than that of a disk sector. Sometimes a search pattern for a file may be split across two sectors. In these cases, the pattern cannot be found by a physical search, but can be found by a logical search.

The third kind of search is the deleted file search. In most file systems, file deletion is typically accomplished by modifying only a few bytes of the file system. The contents of a deleted file are still in the storage system provided that it has not been overwritten. Therefore, patterns in a deleted file can still be found until "deleted" disk sectors are overwritten by other new files. DESK is able to search the sectors of files that have been deleted but not yet overwritten.

The above DESK search functions can be repeated with different search keywords simply by modifying the search keywords in the pattern file. Figure 2 shows a sample screen capture of the DESK system when using the search function.

4. DESK file system integrity checker

DESK provides a file system integrity checking function that enables users to examine the ‘fingerprints’ of the files. Technically, the ‘fingerprints’ are called *hash values*. A hash value [5] is computed by a *hash function*, which is a well-known, openly published algorithm that takes a stream of bytes (such as an electronic file) as input and calculates a fixed-size binary data item as output. The two most popular hash functions are MD5 [6] and SHA-1 [7]. MD5 will take a file and produce a 128-bit binary data while SHA-1 will produce a 160-bit binary data. The mathematical theory of hash functions guarantees the following properties:

- If a file *F* gives a hash value *H1*, then every single bit of *H1* is a function of all bits of *F*.
- If a file *F* gives a hash value *H1*, then modifying *F* by a single bit will result in a totally different hash value *H2*.
- If a file *F* gives a hash value *H1*, and given another hash value *H2* not equal to *H1*, it is computationally impossible to purposefully modify parts of *F* (such as modifying the last 10 bytes) such that the newly modified file will produce *H2* as the hash value.
- The chance of two randomly selected files having the same hash value is extremely small. For example, the chance of two files have the same MD5 hash value, which has 128 bits, will be 1/(2^128), roughly equal to 1/(3.4 × 10^38), or roughly the chance of one in 340 billion billion billion billion. To put this number in perspective: the published chance of winning the first prize in the Hong Kong Mark Six (the lotto game in Hong Kong which randomly picks 6 numbers between 1 and 49) is one in 13,983,816. Therefore, the chance of having two files with the same MD5 hash values is much less than that of winning the first prize in the Hong Kong Mark Six. The chance of two files having identical SHA-1 hash values is even smaller because SHA-1 hash values are 160 bits long.

Loosely speaking, the above properties guarantee that the hash value of a file is similar to the fingerprint of a human being. If we want to find the existence of a file *F1* in a huge file system in a subject machine, we can first compute the hash value of *F1*, and compare this hash value with the hash value of every file in the subject machine’s file system. If a match of hash value occurs, then most likely *F1* is in the file system. This is like the matching of fingerprints to identify a person.

Another usage of the hash value is to check whether some standard files have been modified. One of the objectives of cyber crime investigation is to check whether criminal records are stored under the name of some popular software (e.g. ‘winzip.exe’). To achieve this, the hash values of all of the files in a standard software distribution are computed, and the results are stored in a hash value database. The hash value of every file in the subject machine will be computed by the DESK software, and then matched with that of the corresponding file (i.e. the file with the same file name) in the database. If the standard file in the subject machine has been modified, the hash values will not match, and the DESK software will detect this. This is like checking a person’s fingerprint with the stored record to see whether he/she is that person.

For the DESK tool implementation, each file in a directory system has two hash values that are computed by using the 2 different hash functions, MD5 and SHA-1. These hash values are stored in the database. The basic steps for using the File System Integrity Checker are shown below:

Step 1: Create a hash value database for a collection of files.
Step 2: Use database file.
Step 3: Process the file system of the subject machine covering

- disk drive(s) or
- a directory or
- a certain group of files in directory/disk drive(s) or
- a single file.

Two potential applications of the DESK file system integrity checker are as follows:

- When a computer is suspected to have been hacked, DESK can be used to check whether any files have been modified.
- Use DESK to build a database for standard commercial software, and use this to check whether some of the software's standard files (e.g. 'Freecell.exe') have been modified to store illegal information.

Another important application of the hash function is to compute the hash value of the file system (e.g. a hard disk) of the subject machine. Suppose a subject machine is seized for further investigation, it is necessary to ensure that the files in the subject machine cannot be modified by the law enforcement agency. DESK provides the functionality of computing the hash value of the entire hard disk. If a single byte on the hard disk is changed then the hard disk hash value will change too. Therefore, by recording this hard disk hash value properly, the law enforcement agency can easily prove that the content of the hard disk has not been modified, simply by re-computing the hash value and by comparing it with the recorded hard disk hash value.

Since preserving the integrity of the subject machine is extremely important, it is advisable to keep its use in forensic analysis to a minimum in order to reduce the probability of causing accidental damage. Consequently, it is often the case that exact copies of disks (also called clone images) of the subject machine are made and then used in subsequent analysis. Hashing can be used to ensure that the copies are exact replicas of the original disks (both having identical hash values) and the results obtained from analyzing these copies should be identical to those performed on the actual disks.

Note that the DESK software also provides a 'lock' facility that disables the subject machine's interrupts to avoid accidental modification of the contents of the subject machine. This function greatly enhances the reliability of the evidence.

5. Case management

DESK provides a unique way for managing each forensic case. A special hardware token is issued to each forensic agent who is responsible for managing all of the information collected for a specific case. The subject machine is enabled *only* when a hardware token is attached to it. When the token is removed, the subject machine will be locked and no one will be able to access the data except for the responsible agent.

The case management function also provides facilities for a forensic agent to organize and navigate around all of the information that has been generated relating to a particular case. This information includes multiple clone images, remarks about clone images, and other ‘physical’ evidence such as photos taken at the crime scene. All operations performed on this information are logged for audit purposes. A special case transfer function authenticated by digital signature is provided to ensure that the electronic evidence will not be tampered with when the case is transferred from one agent to another.

6. Look for differences

Another function unique to DESK has been added recently for identifying differences between a clone image and a physical disk. The operation is also applicable between a pair of physical disks and between a pair of clone images. The starting sector in both the source and the destination of the comparison can be selected to allow for *locating* a smaller source in a larger destination (in case ‘diff’ reports that there is no difference between the two).

One potential application of this difference operation is to cover the case of not being able to match the hash value of a clone image with the hash value of the original hard disk even though both have not been tempered with. This could happen if one or more sectors in the hard disk turned bad when, for example, it was restarted to compute the hash value of that device for the purpose of comparison with that of the clone image. The difference function will come in handy to identify the locations that give rise to the differences. If further examination reveals that the differences are all bad sectors, it could be used to explain the failure to match the hash values.

7. Summary

A typical usage of the DESK software begins with the preparation of the pattern file, which contains search keywords, and the hash value database. Next, the DESK machine is connected to the subject machine; the proper hardware token is then attached to the subject machine. The subject machine is then started up using the DESK floppy diskette. The inspection process will usually consist of repeated applications of the three search methods (physical search, logical search, and deleted file search) and the matching of the hash values. The outcome of these operations will be one of the following:

- The subject machine contains some information that is considered to be of value for further investigation. In this case, the subject machine will be seized. Reports containing the hard disk hash value will be generated and processed in accordance with standard procedures.
- The subject machine contains no information worthy of further investigation. The subject machine can be released.

In short, the main value of DESK is to assist the law enforcement agency to quickly examine a subject machine, and to make a quick decision of whether a full-scale and time-consuming investigation of the subject machine should be carried out.

Without the DESK software, or similar systems, law enforcement agents will have no choice but to seize all subject machines for lengthy inspections. The whole investigation process of cyber crimes will become very slow and expensive. Therefore software systems like DESK are essential to criminal investigation in a modern society.

Acknowledgement

The project is supported by a grant from the Technology Transfer Seed Fund of Versitech Limited (a wholly owned subsidiary of the University of Hong Kong).

References

[1] Marcella, Albert J. and Robert S. Greenfield (Eds.), *Cyber Forensics A Field Manual for Collecting, Examining, and Preserving Evidence of Computer Crimes*, Auerbach, 2002.

[2] Evidence Ordinance, Chapter 8, Hong Kong Ordinances.

[3] Digital Evidence Search Kit (DESK) User's Guide, Center for Information Security and Cryptography, The University of Hong Kong, 2001.

[4] Digital Evidence Search Kit (DESK) Quick Reference, Center for Information Security and Cryptography, The University of Hong Kong, 2002.

[5] Stallings, William, *Cryptography and Network Security: Principles and Practice*, 2nd Ed., Prentice-Hall, 1999.

[6] Rivest, R., *The MD5 Message-Digest Algorithm*, RFC-1321, IETF, 1992.

[7] Eastlake, D. 3rd and P. Jones, *US Secure Hash Algorithm 1 (SHA1)*, RFC-3174, IETF, 1991.

0-7695-2478-8/05 $20.00 © 2005 IEEE

Figure 1 – DESK System Overview

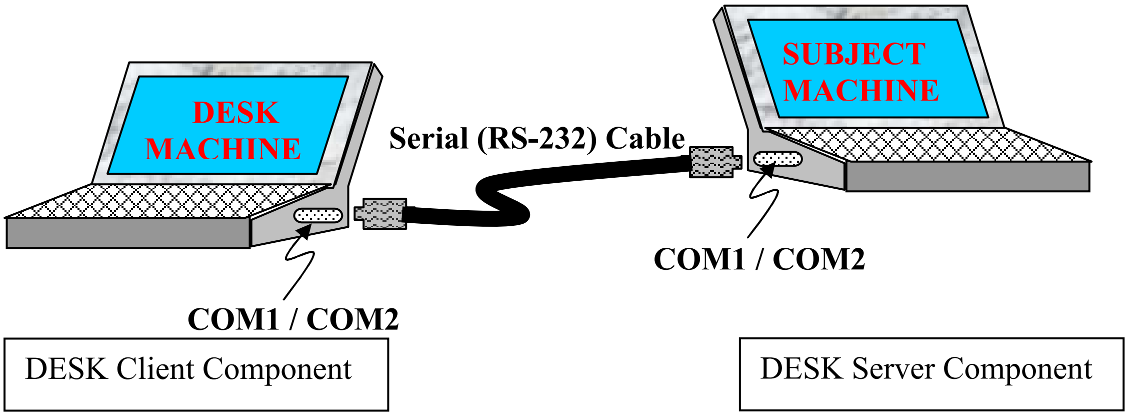
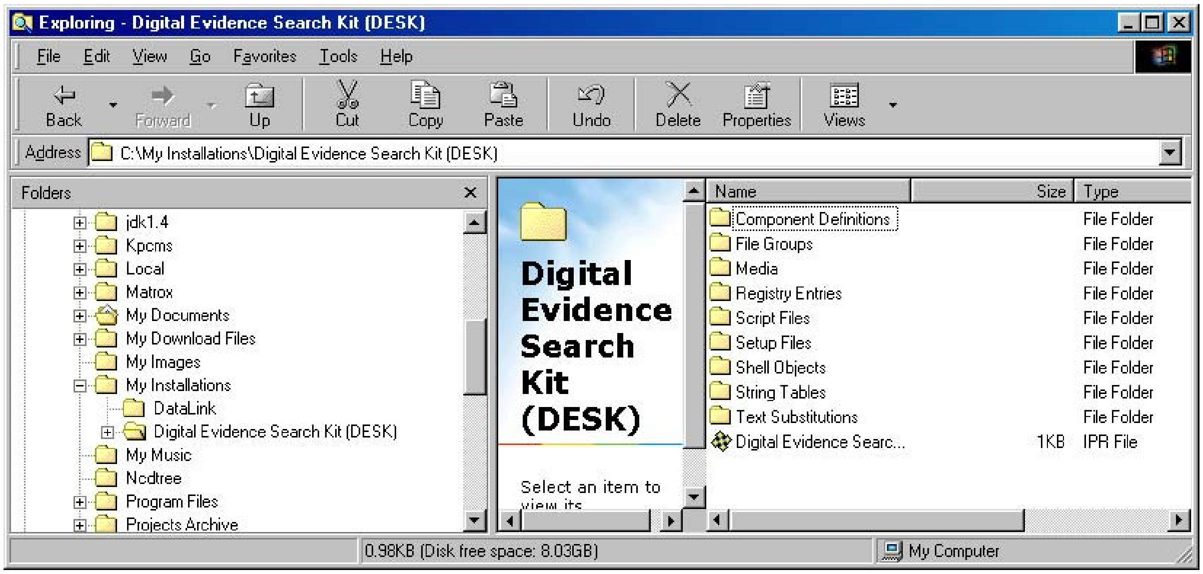


Figure 2 – Sample DESK Search Screen Dump


Exploring - Digital Evidence Search Kit (DESK)
File
Edit
View
Go
Favorites
Tools
Help
Back
Forward
Up
Cut
Copy
Paste
Undo
Delete
Properties
Views
Address
Folders
Digital Evidence Search Kit (DESK)
Select an item to view its

Address: C:\My Installations\Digital Evidence Search Kit (DESK)

Folders:

- jdk1.4
- Kpcms
- Local
- Matrox
- My Documents
- My Download Files
- My Images
- My Installations
  - DataLink
  - Digital Evidence Search Kit (DESK)
- My Music
- Ncdtree
- Program Files
- Projects Archive

Name	Size	Type
Component Definitions		File Folder
File Groups		File Folder
Media		File Folder
Registry Entries		File Folder
Script Files		File Folder
Setup Files		File Folder
Shell Objects		File Folder
String Tables		File Folder
Text Substitutions		File Folder
Digital Evidence Searc...	1KB	IPR File

0.98KB (Disk free space: 8.03GB) My Computer