2016-12-07

Leveraging Biometrics for Insider Misuse Identification

Alruban, A

10.22619/ijcsa.2016.100107
International Journal on Cyber Situational Awareness
Centre for Multidisciplinary Research, Innovation and Collaboration (C-MRiC)

Leveraging Biometrics for Insider Misuse Identification

Abdulrahman Alruban[1, 2], Nathan Clarke[1, 3], Fudong Li[1] and Steven Furnell[1, 3, 4]

[1]*Centre for Security, Communications and Network Research, Plymouth University, Plymouth, UK*
[2]*Computer Sciences and Information Technology College, Majmaah University, Majmaah, Saudi Arabia*
[3]*Security Research Institute, Edith Cowan University Perth, Western Australia*
[4]*Centre for Research in Information and Cyber Security, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa*

ABSTRACT

Insider misuse has become a real threat to many enterprises in the last decade. A major source of such threats originates from those individuals who have inside knowledge about the organization's resources. Therefore, preventing or responding to such incidents has become a challenging task. Digital forensics has grown into a de-facto standard in the examination of electronic evidence, which provides a basis for investigating incidents. A key barrier however is often being able to associate an individual to the stolen data—especially when stolen credentials and the Trojan defense are two commonly cited arguments. This paper proposes an approach that can more inextricably link the use of information (e.g. images, documents and emails) to the individual users who use and access them through the use of transparent biometric imprinting. The use of transparent biometrics enables the covert capture of a user's biometric information—avoiding the potential for forgery. A series of experiments are presented to evaluate the capability of retrieving the biometric information through a variety of file modification attacks. The preliminary feasibility study has shown that it is possible to correlate an individual's biometric information with a digital object (images) and still be able to recover the biometric signal even with significant file modification.

***Keywords**: Digital forensics; biometrics; grille cipher; insider misuse; data leakage; guilty identification.*

1 INTRODUCTION

Insider threats to enterprises have become widespread in the last decade (PwC, 2015). Therefore, it has been considered an important security issue by many recent research studies (Collins et al., 2013; Colwill, 2009; Huth et al., 2013; George Magklaras et al., 2010; Roy Sarkar, 2010; Shabtai et al., 2012; Stamati-Koromina et al., 2012). In addition, insiders who have legitimate access to the organization's internal systems and databases, have advantage of accessing all kind of information including those classified as confidential. A study found that more than 300,000 internal security breaches took place in the United Kingdom in 2013 (IS Decisions, 2014). These breaches lead to substantial damage to the exploited organisation by losing or disclosing its sensitive and confidential intellectual property. In particular, when the exposure originates from an authorized person (e.g. employee, contactor, etc.) who misuses the advantage of privileged and legitimate access to the firm's internal resources, this intensely increases the scope of the devastation. This is because insiders are more likely to bypass security controls compared with outsiders who supposedly have a limited knowledge about the internal infrastructure. As a result, insiders pose significantly greater threats to organisations than outsiders do. According to a survey by Gartner Inc. (2016), revealed the fact that less than five percent of organisations were actually tracking and reviewing privileged activities; while the remainder were, at best, controlling access and logging when, where and by whom privileged access takes place – but not what is actually done. Unlike those who monitor and evaluate privileged activity, they are at risk being blindsided by insider threats, malicious users or errors that cause significant threats.

Digital forensics aims to produce and test a hypothesis about who did what, where, when, why and how in relation to the incident under an investigation. Indeed, existing methods and tools used by investigators to conduct examinations of a digital crime significantly help in collecting, examining and presenting the digital evidence and have become the de-facto standard in analysing incidents (Brown, 2015; Carbone, 2014; SANS Institute, 2016; Shavers, 2013; Vincze, 2016; Widup, 2014). However, during this process, it is often difficult for digital forensic professionals to establish that a particular person has used the specific identity of a digital subject at a certain time (Brown, 2015; Shavers, 2013; Vincze, 2016). In many scenarios, criminals have effectively argued and denied the charge by claiming that someone else used their computer (often through the use of

stolen credentials) or their computer was infected by malware or Trojans (Bowles et al., 2015).

This paper has focused upon the use of biometrics that could provide such a link through transparently capturing the user's biometrics and instantly generating a *biometric imprint* that correlates the user interaction with the digital object providing the ability for investigators to answer the "who?" question (Widup, 2014). To this end, the paper introduces a proactive framework that uses transparent biometrics to aid digital forensic investigators in their analysis of electronic evidence. Also, it examines the feasibility of linking a subject (i.e. computer user) with an object of interest such as photographs, documents, or emails. To validate the approach, a set of experiments that employ a grille cipher to link embed the transparent biometric sample are conducted. Unlike most existing methods, such as digital watermarking or null ciphers, the integrity of the object is modified (Charbonneau et al., 2014; Nelson et al., 2014), a grille cipher simply employs a template that is used to cover the carrier message; the words that appear in the openings of the template are the hidden message. Furthermore, the proposed approach only "imprints" user's biometric feature vector. Therefore, the employed imprinting process can be described as a correlation of the feature vector with the object.

The rest of the paper is organised as follows: the second section provides the background information about the role of biometric technology in digital forensics. It also discusses some of the related work in the area of insider misuse identification. The third section introduces the proposed approach, including the core processes, followed by the forth section which explains the experimental methodology of different possible types of attack. Section five presents the experimental results. Section six presents a discussion of the approach and identifies a number of challenges. Finally, the paper concludes in section seven with the future work.

2 BACKGROUND

The science of digital forensics has existed for a long time, aiding organisations in investigating cyber-crimes. Digital forensics can be described as the process and science of extracting information and data from electronic devices to serve as electronic evidence for proving and legally prosecuting digital crime (Casey, 2009). This includes but is not limited to extracting relevant information from computers, smartphones, network devices, databases, and storage media. In addition, with the involvement of biometric technologies, forensic capabilities could be significantly increased; which in turn may answer crucial questions that investigators trying to figure out.

2.1 The Role of Biometric Technology in Forensics

Biometric technology has various important applications in forensic science. For example, it has been used for identifying missing individuals following natural disasters or accidents, such as fires (Kolude et al., 2011). Moreover, biometrics has helped law enforcement agencies in identifying attackers who are involved in terrorist crimes (Federal Bureau of Investigation, n.d.). Since the use of facial-recognition technologies has significantly speeded up and automated the process of matching the questioned individual by comparing fingerprints or facial photograph against the database (Spaun, 2007).

Despite promising performance of automatic face recognition algorithms in a controlled setting (Kemelmacher-Shlizerman et al., 2015), many applications require accurate identification at planetary scale, i.e., finding the best matching face in a database of millions of people. For instance, face recognition algorithms failed to identify criminals in the Boston marathon bombing (Klontz et al., 2013). All these applications and challenges applied to reactive forensic activities, where the whole investigation takes place post the incident. In recent years, there has been an increasing interest to overcome such challenges by proactively record, gather and analyse intelligence prior to tackling an incident. This includes capturing biometric data in unobtrusive manner to provide incident response teams with the information that could help in accelerating the investigation process time (Alruban, et al., 2016).

2.2 Transparent Biometric

In recent years, there has been an increasing interest in transparent biometric authentication (Clarke et al., 2006; Frank et al., 2013; Martinho-Corbishley et al., 2016; Prakash, 2014; Reid et al., 2014). Typically, a transparent authentication is performed by any means that is able to acquire an individual biometric sample required for the verification non-intrusively (Clarke, 2011). In addition, the concept of transparent could be used in monitoring and profiling computer users. Nevertheless, achieving that is a challenging task, especially when it is meant to be unobtrusive, since such an approach provides more flexibility in terms of environment under which the capture takes place, which can impact the quality of the sample. For instance, in the case of facial recognition, external factors such as illumination, the subject's distance from the camera and facial orientation significantly affect the recognition accuracy of the technique. As such, not all biometric techniques can be adapted to operate in a transparent manner and those that are tend (but not exclusively) to be behavioural rather than physiological (Li et al., 2009).

Despite all the aforementioned challenges and obstacles that the transparent authentication introduces, it has a high potential benefit over conventional intrusive biometric in this research due to covert nature of the biometric sample capture minimising opportunities for sample forgery. Given the nature of insider misuse and the desire to point the attention away from the guilty party, providing a robust approach such a continuous and covert capture of biometric signals will increase the complexity of attack.

2.3 Misuse Identification and Detection

Previous studies have primarily concentrated on monitoring systems that operate proactively for forensic and audit purposes (Cohen et al., 2011; G Magklaras et al., 2011; Rafique et al., 2013; Shields et al., 2011). In (G Magklaras et al., 2011), the authors have developed an audit engine for actively logging user actions in Relational Database Management System. The proposed system could be used to aid an incident investigation by post-case forensic examiners. It stores actions include accessed files information (name, type, location), times stamp, process execution, network endpoint and hardware device information in addition to other related information. Furthermore, the engine employs a linguistic analysis of users' correspondence as a monitoring technique, thus as to proactively detect potential insider threat risks in the organization. In addition, it facilitates the use of Structured Query Language, which enables instance selection and completion. Such function allows investigators to enumerate that database and performs a variety of enquires. The system was tested on a variety of simulated insider misuse scenarios. Although the evaluation results are promising in terms of logging different types of user's actions along with useful information, it still does not correlate those actions to the individual who performed them.

Similarly, (Shields et al., 2011) proposed a system that proactively and continuously collects evidence by creating and storing file signatures that are deleted, edited, or copied within computers on the local network. The system uses a centralized database to store the generated objects' signatures, which provide significant information, such as user identifier, object timestamp, and type of the event. Thus, investigators could use such information as a lead when conducting a forensic activity. The generated fingerprints are equal to ~1.06 percent of the original file size, which is a huge reduction in terms of storage space. Furthermore, the system supports several file types, such as Microsoft Word documents and Portable Document Formats (PDF). For the deployment, the system requires patching the system kernel in order to intercept system calls. Unfortunately, such low-level kernel hardcoding is typically limited to only open source operating systems. In addition, the login details can be shared, stolen, and

compromised. Hence, the user identifier can be unreliable. In contrast, our proposed approach does not require any modification on the kernel level.

Furthermore, a variety of studies have examined the possibility of identifying the person that leaked data (Chavan et al., 2013; Jadhav, 2012; Kale et al., 2012; Papadimitriou et al., 2011). (Papadimitriou et al., 2011) investigated the feasibility of inserting fake objects into data of interest before distributing these data to third party agents. However, adding these fake objects is not always possible. For example, in the case of medical records, manipulating the data or injecting invalid information could lead to huge risk and consequences on the patients' life. The examination of the feasibility of their method found that it is better in identifying the source of the data leakage compared to the simple data allocation algorithms. Moreover, 95% of confidence was obtained via their experiments in identifying the leakage source (suspect agent).

Subsequent practical implementations of the guilt model resulted in the development of several prototype models (Chavan et al., 2013; Jadhav, 2012; Kale et al., 2012). All of these models use the same concept introduced in (Papadimitriou et al., 2011) by inserting unique fake objects or digital watermarks to the data prior to the distribution. In general, the data creator (in this case the distributer) is responsible for generating and embedding the fake objects. However, in many cases the data can be created by an insider who leaks the sensitive data by himself. In additional, the fake object creation process could be a complicated task.

3 THE PROPOSED APPROACH

The proposed framework acts as a proactive biometric-based forensic system that can inextricably link the use of information (evidence) to the individual users who access it (which is the subject of a patent by the authors (Alruban & Clarke, 2016). The framework mainly consists of two engines, these are: biometric capture and processing and an imprinting engine. Each of which performs multiple tasks through its built-in functions as illustrated in Figure 1.

3.1 Biometric engine

The biometric engine transparently and continuously captures and extracts a user's biometric features. The engine is designed to be a multimodal biometric tool, which monitors the user's interaction processes with the computer and instantly captures various biometric samples. For instance, while the user is editing/writing a document, different biometric modalities could be used to profile the subject, including facial recognition, keystrokes

dynamics, and mouse dynamics. Facial samples can be captured by a fitted camera while the user is looking at the computer screen.

While the keystrokes dynamics identifies an individual by the way in which they type, the user's keystrokes behaviour could be triggered and profiled during the observation process. Similarly, mouse dynamics involves a signature that is based on selected mouse movement characteristics, which are computed to generate a unique individual biometric feature. However, the performance of these modality may vary significantly, since the amount of interactions that users perform—using keyboard and/or mouse—vary among them and also changes from one session to another (Traore et al., 2012). Therefore, employing such a multimodal biometrics system increases the confidence rate of profiling the user and generates a reliable reference template to be used in the imprinting process by the imprinting engine. In addition, it is envisaged that the use of multimodal approaches will further prevent forgery-based attacks. Once the biometric sample is obtained, the engine temporarily stores it in a local database on the user's machine. Thereby, the database would continually have the most updated biometric data that can be used by the imprinting engine for generating the imprints.

3.2 Imprinting engine

The imprinting engine retrieves the object's metadata along with its hex representations and requests the latest user's biometric feature vector from the biometric engine to be used in the imprinting process. Its main function is mapping and linking the biometric sample with the interacted object to produce the imprint. Finally, these generated imprints are stored in a centralized database for later use. Upon the detection of data leakage, the object (whether it is posted on a public website or captured by the network) can be analysed for the biometric imprint. The sample is extracted and then processed by a biometric system in order to determine the last user who interacted with the object as presented in Figure 2.

3.3 Imprinting process

The generation process of the imprints is inspired by the benefits of employing the grille cipher technique. Grille ciphers has been used in the past (prior to the modern null ciphers) as a means for transferring/exchanging secret messages between two parties. It was originally used to extract hidden messages from plain text by mapping the text throughout a pierced sheet or a cardboard. Therefore, the embedded secret message can be retrieved by mapping specific locations. Hence, applying the same technique to imprint the biometric feature vector to an object file is possible, where the object can be an image file, document, video, or any digital file types. The key advantage in utilising a grill rather

than null cipher is the ability not to modify any information contained within the digital object (thereby preserving the integrity of the original data). In order to adapt the grill cipher technique to the proposed approach, it involves several consecutive steps, as follows:

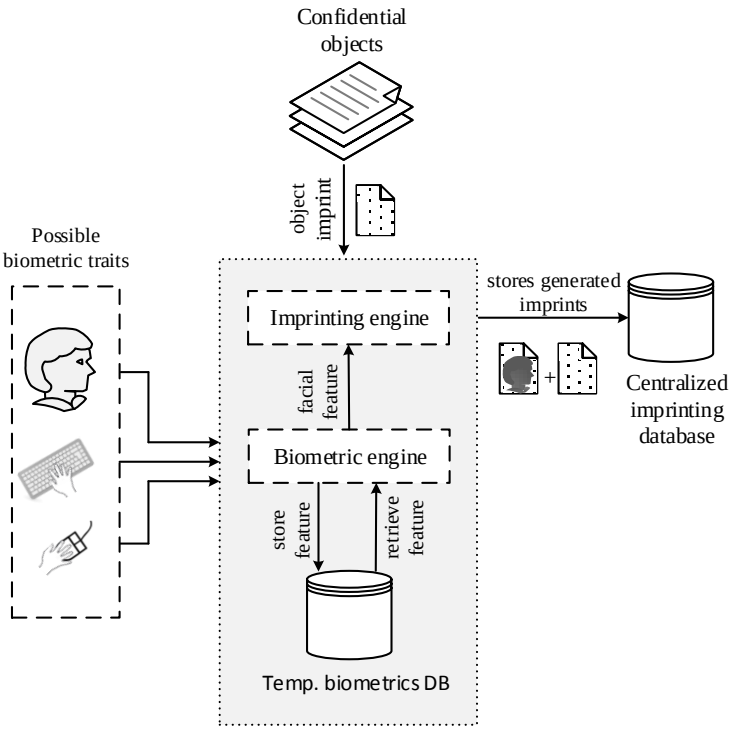

Figure 1. The Proposed framework architecture

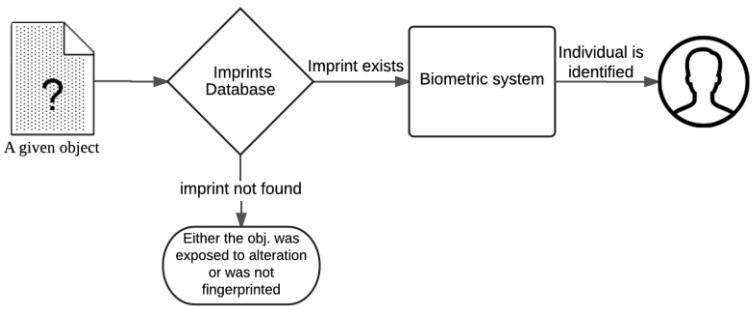

Figure 2. The process of identifying an individual

3.3.1 Preparation of Feature Vector and Object

The preparation step converts both feature vector and object into its Hex representations for the mapping purpose. In addition, the index of each character is preserved during this conversion, which begins with '0' for the first character and ascendingly continues until the last one. Furthermore, the process of conversion is not necessarily achieved by transforming each character, since reading the whole object in binary mode allows for low-level representations of both Hex and Binary. However, still character-by-character (or byte-by-byte) indexing is required in order to generate the object index list.

3.3.2 Mapping the Feature Vector with the Object

After obtaining the Hex representations of the feature vector and object, each Hex value in the feature vector is mapped with its equivalent positions in the object's Hexes to retrieve the possible positions where both are match. Accordingly, the mapping process returns lists of indexes for those matched Hexes.

3.3.3 Generating the Feature Vector Imprints

By retrieving the positions of each character of the feature vector with the object, now it is possible to generate the imprints based on the list of indexes, which means that multi- imprints of the whole feature vector can be generated by combining those positions.

The pseudocode of the imprinting process starting from the preparation is illustrated below in Algorithm 1.

Algorithm 1: imprinting algorithm:

Input: Feature Vector (*FV*), Object (*O*)
Output: Imprints
1: **function** PREP (*FV, O*)
2: **for** *each value in FV & O:*
3: Convert *FV, O* into its HEX representations
4: Retrieve the index of each value
5: **Return** $FV_{HEX,\,index,}\ O_{HEX,\,index}$
6: **function** MAPPING ($FV_{HEX,\,index,}\ O_{HEX,\,index}$)
7: **for** each value in $FV_{HEX,}\ O_{HEX:}$
8: $index\,(O_{index}) \leftarrow FV_{HEX} \cap O_{HEX}$
9: **Return** $index\,(O_{index})$
10: **function** IMPRINTING (*indexes*)
11: *imprint* ← Combine unique indexes from the
12: retrieved index list
13: **Return** *imprints*

The next section investigates the feasibility of imprinting biometric feature vectors with images and later recovering them (even after object modification).

4 EXPERIMENTAL METHODOLOGY

The main goal of the experiment is to assess the feasibility of the proposed hypothesis where the subject's feature vector can be forensically linked and retrieved from an object of interest. Therefore, it is critical to evaluate its performance in a complex, subject-related manner. In total, four experiments were conducted as follows:

- The first experiment retrieves the feature vector from the original imprinted image.
- The second experiment examines the situation where the image is modified in one area with an increasing proportion of modification.
- The third experiment verifies the case where the image is modified in several areas.
- The final experiment investigates when only parts of the original image are available, while the rest is missing.

In these experiments, the used feature vector presents a real facial feature vector sample with a length of 57 numeric characters, as illustrated in Figure 3. The length of the vector relies upon the used feature extraction algorithm to compute the feature vector. In this study, Fisherfaces algorithm is used to compute the feature vector for the captured users' faces images (Belhumeur et al., 1997). In addition, the algorithm performs a Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for dimensionality reduction (Yu et al., 2001).

Regarding the used objects in the performed experiments, the UCID image dataset version 2 is used (Schaefer et al., 2003). It contains a total of 1,300 images with two sizes, either (1,234 x 1,858) or (1,858 x 1,234) width, height in pixels respectively. For the purpose of this study, the first 100 images are used from this dataset, since it is assumed that this number is enough for the purpose of evaluation. The implementation of the proposed algorithm was developed in Python due to its flexibility in terms of list comprehension and image processing. Moreover, Python's built-in library has several useful functions, such as *map* and *zip* which facilitate many relevant operations (Python.org, n.d.). As regards the deployment, these tests were conducted on a machine with Microsoft Windows 7, Intel Core i5 2.70GHz and RAM 4.00 GB.

[1679.2235398,-1555.40390834,-1140.07728186,-1999.85500108]

Figure 3. Facial feature vector

4.1 Retrieving the Feature Vector from the Original Imprinted Image

The aim of this experiment is to imprint the feature vector as many times as possible with each image in the dataset. The first experiment examines the possibility of generating the imprints between the feature vector and the digital object in use by the user. Since there is a high probability that the subject or other party (for intentional or unintentional reasons) somehow will modify the questioned object after it is imprinted, the subsequent experiments investigates the accuracy of retrieving the feature vector from the object under several modification scenarios.

4.2 Modification in One Area

Experiment two evaluates the imprinting mechanism after the image is modified by a different percentage. The simulation of this is performed by randomly choosing a section of the image as a rectangle box at a growing size to reflect an increasing proportion of modification. In addition, equation 1 is used to determine the size and the random position of the modified section. The equation takes three variables, which are:

- *w*: image width,
- *l*: image height,
- *s*: the desired modification percentage.

The equation gives four values; x and y are random values between (0, image width) and (0, image height) respectively. These set the top left pixel position of the modified rectangle (as presented in red colour in Figure 4). The third and fourth values are for the right down corner of the rectangle (as presented in blue colour).

$$P_{(w,l,s)} = \sum_{x=0}^{w-1} \sum_{y=0}^{l-1} (x, y, x + \frac{w}{10 \cdot \sqrt{s}}, y + \frac{l}{10 \cdot \sqrt{s}}) \quad (1)$$

In this experiment, the imprinted images have been modified by 5% increments, which means that the first alteration rate is 5% then 10%, 15% and so forth, until reaching 100%. Figure 5 demonstrates some samples of an image modified in different rates. The upper left image is modified by 5% of its original size, where the rest are modified at rates of 35%, 65%, and 95% respectively.

4.3 Modification in Multi Areas

The third experiment is similar to the previous one, except that the modifications occur in several parts of the image instead of an increasing proportion of one area. This type of attack is more influential since various and random parts of the image are affected by such alterations. In order to simulate such modifications, the dataset images are altered using multiple rectangle boxes, each of which is equal to 1% of the total image size. Therefore, simulating 5% randomly locations alteration, it would need five of these boxes among an image. In addition, this experiment assesses the proposed technique with an alteration size on the objects by 5% increments of its original size. Figure 6 illustrates four sample images modified by 5%, 35%, 65%, and 95% respectively.

4.4 Partial Image

Further investigation was needed to better understand the effects of different attack vectors on retrieving the imprinted feature vector. Therefore, the last experiment in this study is interesting as it simulates the scenario where only part of the imprinted image is available and the rest is missing. For instance, the imprinted image could be resized or cropped. To simulate such alterations, a random section of the images in the dataset was cropped in different sizes, starting from 5% of the original size, and then in each subsequent test, again a random section was cropped with an increment of 5%. Figure 7 illustrates some of these cropped samples.

5 RESULTS

In this study, the aim is to critically assess the hypothesis of linking a subject's biometric feature vector to an object of interest using the grille cipher technique. On average, it takes only ~3 milliseconds to generate an imprint with size average of those imprints is less than ~472 bytes per imprint. The result of experiment one shows that the average number of the generated imprints are 854 per image. While the minimum number of imprints in a single image was 244, and the maximum is 1,815. This means that the mapped feature vector could be retrieved and reconstructed from any of these imprints. This achieved number of imprints is not surprising,

Figure 4. Sample of a modified area

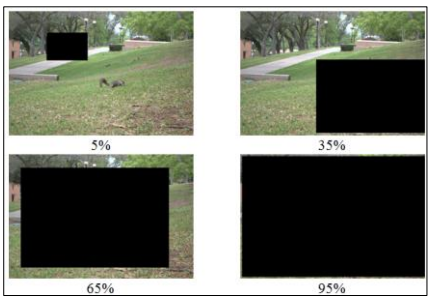

Figure 5. Sample of a modified part of an image

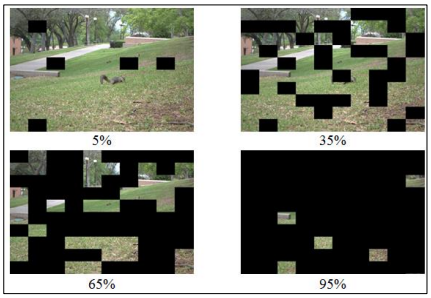

Figure 6. Sample of a modified multiple parts of an image

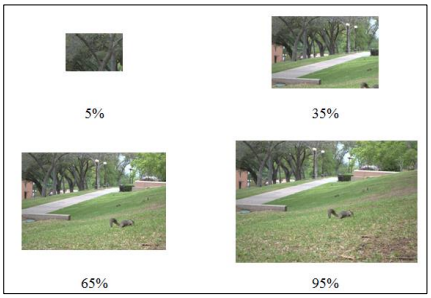

Figure 7. Samples of a cropped image in certain percentage

since the feature vector always contains numerical values (0-9). Therefore, there are many matches between the feature vector and those images' Hex values. In addition, a reconstruction of the feature vector from these unmodified images were possible by using those generated imprints with an accuracy of 100%. This was achieved easily by reversing the imprinting processes.

In the second experiment, it was found that this imprinting technique is very effective, since the imprinted feature vector is successfully retrieved from an average of 97 out of 100 images even when the modification percentage is 80%, as Figure 8 illustrates. However, after a modification of 80% on the images, the number of valid retrieved feature vectors significantly drops due to the loss of most of the imprints values across those images. This decline occurred for the reason that critical set of mapped indexes values are changed after such high modification rate. Yet, it is clearly illustrated that it is feasible to reconstruct the feature vector from the imprinted objects even though the huge destruction to its original values.

In the third experiment where the modification took a place in multiple areas, the result shows that the imprinted feature vector are successfully retrieved, even when the images are altered in more sophisticated way than the one area modification attack (experiment two). Figure 9 exhibits the percentage of images where the feature vector was successfully retrieved. Since changing certain pixels' values-by printing those black boxes- after the imprinting process with the feature vector consequently affects the mapped indexes' values. Therefore, many of the imprints became useless after such attack. Despite massive destruction on the image visualisation with the increased rate of the modification, it is possible to recapture the feature vector from some of those images, even under enormous alteration such as when the object is changed by 95%. At the same time, this attack caused a major loss of the mapped indexes values comparing to the modification in one area experiment. Where the latter is less vandalism than the former in terms of impacting the interested pixels.

Finally, in the last experiment the most striking finding to emerge from the results is that among all these tests in this experiment, the feature vector is retrieved and reassembled 100% among all the tested images. This means that by giving only part of the original imprinted image, it is possible to restore the feature vector to its original values. Figure 10 shows that the average, maximum, and minimum numbers of a retrieved feature vector cross on all examined images (i.e. 100 images). However, these results were obtained by assuming that the preserved indexes of the hexes of interest are not changed after the cropping process. This means that all of the imprints in the database are correlated with the questioned samples as a part of the original images. In practice this is not always possible since the original object might not be accessible or available after the imprinting process took a place. Therefore, more research is needed to find a link between such parts of an object and the original.

6 DISCUSSION

The nature of the imprinting process reveals no information about where to locate the imprinted object–thereby making it particularly challenging to recover or modify as illustrated in the experiments. In addition, the results have evidently shown that by mapping the Hex representations of a feature vector with the Hex representations of an image of interested, it is feasible to generate one or more imprints of this feature vector. The first conducted experiment results revealed an 'expected' outcome by imprinting the feature vector from the original imprinted image. Since 100% of the imprinted feature vector is retrieved using only the generated imprints that contains is retrieved using only the generated imprints that contains the indexes of the

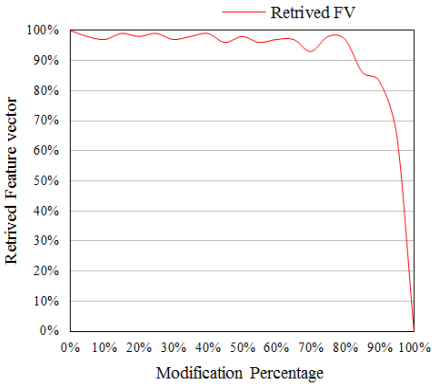

Figure 8. Percentage of images with successful retrieved feature vectors under one area modification attack

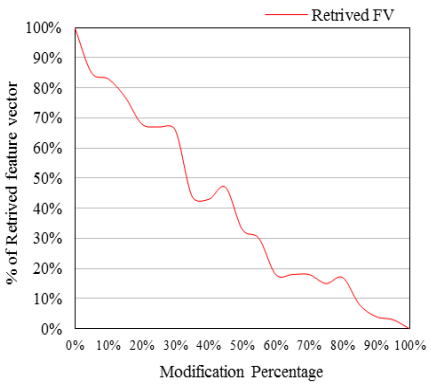

Figure 9. Percentage of images with successful retrieved feature vectors under multiple parts modification attack

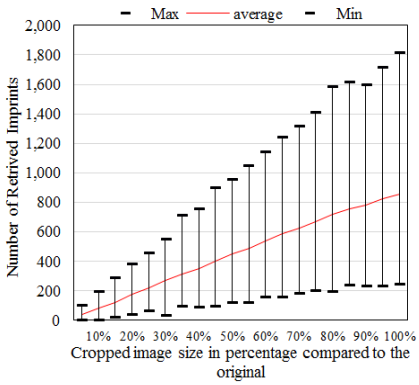

Figure 10. Performance under partial image attack

corresponding positions, it is expected because the mapped objects (images in this case) have not been exposed to any kind of alteration and, therefore, were tested based on their original status. The explanation of being able to score those high results is attributed to the nature of the examined object. Since images are a set of pixels that range from 0 to 255, changing one pixel's value does not affect other pixels' values, or their position. Thus, altering part of the image is not necessary, as it affects all imprinted indexes' values. Therefore, generating as many imprints as possible in various positions of the image, this in turn will increase the probability of successfully retrieving the imprinted values.

It is worth highlighting that the approach introduced in this paper can be applied to other types of objects such as Office Word documents and PDF files. However, the results do not necessarily reflect a robust success rate, since those types of objects are considered a binary format for storing a document. In addition, an initial experiment was carried out where a small set of Office Word and PDF files are examined using the same imprinting technique that was conducted on images. The results showed that unlike images, the dynamic nature of binary files results in changing a small value in the document/file content requires recompiling of the whole binary file, which consequently leads to adjustment of the whole sequence of indexes. For that reason, many attacks would considerably impact the accuracy of retrieving the imprinted feature vector from such objects. Therefore, further work needs to be undertaken to ensure the biometric capturing, processing and imprinting systems need to be hardened against attack and modification in order for the approach to remain valid.

Whilst the proposed approach has the foundations for identifying individuals who are misusing systems and information, several concerns and issues need to be considered and solved before ubiquitous adoption and effective operation of this system. So far, the following aspects have been identified:

- Performance of biometric techniques: the use of transparent biometrics to monitor and acquire subject's traits introduces several challenges that need to be considered when developing such a system. For instance, individual's face pose and illumination, as well as expression and occlusions, may become disturbing factors. Even with extensive research being undertaken in this field, such issues cannot be overcome very easily (De Marsico et al., 2013). Besides much complications would be introduced with the use of transparent biometric where the sample is captured unobtrusively and unsurprisingly.
- Privacy: since the proposed framework incorporates biometric recognition technologies, this involves the use of an individual's characteristics. Hence, those data are considered as personal and sensitive; therefore, issues related to biometric security and privacy have been raised. For that reason, storing and transferring subject's biometrics must be achieved in a manner that minimises the threat to interception and misuse of the information.
- Scalability: developing a system that continuously generates data and stores and transfers them to a central data server, introduces several challenges. This includes traffic management and synchronisation where optimisation is needed.
- Storage: capturing, generating, and storing the data raises a number of technical and conceptual challenges. Particularly, as the proposed

system will generate and collect larger amounts of data, this requires more investigation for enhancing the system to adapt the storage issue.

- Attack vectors: several threats are introduced during the experiment, which includes object manipulations (as examined in this paper). It is therefore necessary to proceed to investigate and identify attack vectors and security concerns and subsequently consider mitigation techniques in order to counter any potentially significant threats.

7 CONCLUSIONS AND FUTURE WORK

The proposed approach is a novel, proactive digital forensic method that enables investigators to inextricably link the use of information (e.g., images, documents, and emails) to the individual users who use and access them. The framework establishes this correlation by matching the biometrics binary representation to equal binary representation within the object to be marked and records these matches as reference in database. Also the system utilises the available computer hardware camera, keyboard, and mouse to transparently and continuously monitor individual's interactions by incorporating different biometric techniques.

Further work is required where not all object types are adapted to behave in the same manner with the proposed approach. For instance, in the case of documents, instead of mapping the feature vector with the object at a Hex level, a higher level of representation could be used. Also, mapping the feature vector with static representations of the document's text, possibly will become less vulnerable to such alteration attack on the object, especially when the generated imprints preserve more static values related to the object. Furthermore, to help prevent some such attack vectors, an investigation into error correction code will also be undertaken, to examine whether it helps in making the proposed approach more robust against certain threats vectors.

8 REFERENCES

Alruban, A., & Clarke, N. (2016). Method Of Associating A Person With A Digital Object. UK: UK Intellectual Property Office.

Alruban, A., Clarke, N., Li, F., & Furnell, S. (2016). Proactive Biometric-Enabled Forensic Imprinting. In *The International Conference On Cyber Incident Response, Coordination, Containment & Control (Cyber Incident 2016)*. London, UK.

Belhumeur, P. N., Hespanha, J. P., & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720. https://doi.org/10.1109/34.598228

Bowles, S., & Hernandez-Castro, J. (2015). The first 10 years of the Trojan Horse defence. *Computer Fraud & Security*, *2015*(1), 5–13. https://doi.org/10.1016/S1361-3723(15)70005-9

Brown, C. S. D. (2015). Investigating and prosecuting cyber crime: Forensic dependencies and barriers to justice. *International Journal of Cyber Criminology*, *9*(1), 55–119. https://doi.org/10.5281/zenodo.22387

Carbone, F. (2014). *Computer Forensics with FTK* (1st ed.). Birmingham: Packt Publishing Ltd.

Casey, E. (2009). *Handbook of Digital Forensics and Investigation*. Academic Press.

Charbonneau, S. R. D. J., & Simon, E. J. (2014). Method and system for generating trusted security labels for electronic documents. U.S.

Chavan, J., & Desai, P. (2013). Relational Data Leakage Detection using Fake Object and Allocation Strategies. *International Journal of Computer Applications*, *80*(16), 15–21. https://doi.org/10.1.1.403.2895

Clarke, N. (2011). *Transparent user authentication: biometrics, RFID and behavioural profiling*. Springer Science & Business Media.

Clarke, N. L., & Furnell, S. M. (2006). Authenticating mobile phone users using keystroke analysis. *International Journal of Information Security*, *6*(1), 1–14. https://doi.org/10.1007/s10207-006-0006-6

Cohen, M. I., Bilby, D., & Caronni, G. (2011). Distributed forensics and incident response in the enterprise. *Digital Investigation*, *8*, S101–S110. https://doi.org/10.1016/j.diin.2011.05.012

Collins, M. L., Spooner, D., Cappelli, D., Moore, A. P., & Trzeciak, R. F. (2013). *Spotlight On : Insider Theft of Intellectual Property inside the U . S . Involving Foreign Governments or Organizations*. *Intellectual Property*.

Colwill, C. (2009). Human factors in information security: The insider threat – Who can you trust these days? *Information Security Technical Report*, *14*(4), 186–196. https://doi.org/10.1016/j.istr.2010.04.004

De Marsico, M., Nappi, M., Riccio, D., & Wechsler, H. (2013). Robust Face Recognition for Uncontrolled Pose and Illumination Changes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *43*(1), 149–163. https://doi.org/10.1109/TSMCA.2012.2192427

Federal Bureau of Investigation. (n.d.). Fingerprints & Other Biometrics. Retrieved June 26, 2016, from https://www.fbi.gov/about-us/cjis/fingerprints_biometrics

Frank, M., Biedert, R., Ma, E., Martinovic, I., & Song, D. (2013). Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. *IEEE Transactions on Information Forensics and Security*, *8*(1), 136–148. https://doi.org/10.1109/TIFS.2012.2225048

Gartner Inc. (2016). Gartner Says By 2018 25 Percent of Organizations Will Review Privileged Activity and Reduce Data Leakage Incidents By 33 Percent. Retrieved March 18, 2016, from http://www.gartner.com/newsroom/id/3207217

Huth, C. L., Chadwick, D. W., Claycomb, W. R., & You, I. (2013). Guest editorial: A brief overview of data leakage and insider threats. *Information Systems Frontiers*, *15*(1), 1–4. https://doi.org/10.1007/s10796-013-9419-8

IS Decisions. (2014). The Insider Threat Security Manifesto: Beating the threat within. Retrieved April 24, 2016, from http://www.isdecisions.com/resources/pdf/insiderthreatmanifesto.pdf

Jadhav, R. (2012). Data leakage detection. *International Journal of Computer Science & Communication Networks*, *3*(1), 37–45.

Kale, S. A., & S.V.Kulkarni. (2012). Data Leakage Detection. *International Journal of Advanced Research in Computer and Communication Engineering*, *1*(9), 668–678.

Kemelmacher-Shlizerman, I., Seitz, S., Miller, D., & Brossard, E. (2015). The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. Retrieved from http://arxiv.org/abs/1512.00596

Klontz, J. C., & Jain, A. K. (2013). A Case Study of Automated Face Recognition:

The Boston Marathon Bombings Suspects. *Computer*, *46*(11), 91–94. https://doi.org/10.1109/MC.2013.377

Kolude, B., Adeyemi, B., Taiwo, J., Sigbeku, O., & Eze, U. (2011). The role of forensic dentist following mass disaster. *Annals of Ibadan Postgraduate Medicine*, *8*(2), 111–117. https://doi.org/10.4314/aipm.v8i2.71826

Li, S. Z., Schouten, B., & Tistarelli, M. (2009). Biometrics at a Distance: Issues, Challenges, and Prospects (Vol. 6256, pp. 3–21). https://doi.org/10.1007/978-1-84882-385-3_1

Magklaras, G., & Furnell, S. (2010). Insider Threat Specification as a Threat Mitigation Technique. In *Advances in Information Security* (Vol. 49, pp. 219–244). https://doi.org/10.1007/978-1-4419-7133-3_10

Magklaras, G., Furnell, S., & Papadaki, M. (2011). LUARM – An Audit Engine for Insider Misuse Detection. *International Journal of Digital Crime and Forensics*, *3*(3), 37–49. https://doi.org/10.4018/jdcf.2011070103

Martinho-Corbishley, D., Nixon, M. S., & Carter, J. N. (2016). Soft biometric retrieval to describe and identify surveillance images. In *2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISBA.2016.7477240

Nelson, M. D., & Xie, M. (2014). DATA LEAK PROTECTION. U.S.

Papadimitriou, P., & Garcia-Molina, H. (2011). Data Leakage Detection. *IEEE Transactions on Knowledge and Data Engineering*, *23*(1), 51–63. https://doi.org/10.1109/TKDE.2010.100

Prakash, A. (2014). A biometric approach for continuous user authentication by fusing hard and soft traits. *International Journal of Network Security*, *16*(1), 65–70.

PwC. (2015). *2015 Information Security Breaches Survey*. Retrieved from https://www.pwc.co.uk/assets/pdf/2015-isbs-technical-report-blue-03.pdf

Python.org. (n.d.). 2. Built-in Functions — Python 2.7.11 documentation. Retrieved February 6, 2016, from https://docs.python.org/2/library/functions.html

Rafique, M., & Khan, M. N. A. (2013). Exploring Static and Live Digital Forensics: Methods, Practices and Tools. *International Journal of Scientific & Engineering Research*, *4*(10), 1048–1056.

Reid, D. A., Nixon, M. S., & Stevenage, S. V. (2014). Soft Biometrics; Human Identification Using Comparative Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(6), 1216–1228. https://doi.org/10.1109/TPAMI.2013.219

Roy Sarkar, K. (2010). Assessing insider threats to information security using technical, behavioural and organisational measures. *Information Security Technical Report*, *15*(3), 112–133. https://doi.org/10.1016/j.istr.2010.11.002

SANS Institute. (2016). SANS Investigative Forensics Toolkit Documentation. Retrieved February 12, 2016, from https://media.readthedocs.org/pdf/sift/latest/sift.pdf

Schaefer, G., & Stich, M. (2003). UCID: an uncompressed color image database. In M. M. Yeung, R. W. Lienhart, & C.-S. Li (Eds.), *SPIE 5307, Storage and Retrieval Methods and Applications for Multimedia 2004* (pp. 472–480). https://doi.org/10.1117/12.525375

Shabtai, A., Elovici, Y., & Rokach, L. (2012). *A Survey of Data Leakage Detection and Prevention Solutions*. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4614-2053-8

Shavers, B. (2013). *Placing the suspect behind the keyboard: using digital forensics and investigative techniques to identify cybercrime suspects*. Newnes.

Shields, C., Frieder, O., & Maloof, M. (2011). A system for the proactive, continuous, and efficient collection of digital forensic evidence. *Digital Investigation*,

8(SUPPL.), S3–S13. https://doi.org/10.1016/j.diin.2011.05.002

Spaun, N. A. (2007). Forensic Biometrics from Images and Video at the Federal Bureau of Investigation. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems* (pp. 1–3). IEEE. https://doi.org/10.1109/BTAS.2007.4401932

Stamati-Koromina, V., Ilioudis, C., Overill, R., Georgiadis, C. K., & Stamatis, D. (2012). Insider threats in corporate environments. In *Proceedings of the Fifth Balkan Conference in Informatics on - BCI '12* (p. 271). New York, New York, USA: ACM Press. https://doi.org/10.1145/2371316.2371374

Traore, I., Woungang, I., Obaidat, M. S., Nakkabi, Y., & Lai, I. (2012). Combining Mouse and Keystroke Dynamics Biometrics for Risk-Based Authentication in Web Environments. In *2012 Fourth International Conference on Digital Home* (pp. 138–145). IEEE. https://doi.org/10.1109/ICDH.2012.59

Vincze, E. A. (2016). Challenges in digital forensics. *Police Practice and Research*, *4263*(February), 1–12. https://doi.org/10.1080/15614263.2015.1128163

Widup, S. (2014). *Computer Forensics and Digital Investigation with EnCase Forensic v7* (1st ed.). McGraw-Hill Osborne.

Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional data — with application to face recognition. *Pattern Recognition*, *34*(10), 2067–2070. https://doi.org/10.1016/S0031-3203(00)00162-X

BIOGRAPHICAL NOTES

Abdulrahman Alruban is a PhD researcher in the Centre for Security, Communications and Network Research (CSCAN) at Plymouth University (UK), and a lecturer in the Computer Sciences and Information Technology College at Majmaah University (KSA). He obtained his BSc in Information Technology form Southampton Solent University and MSc in Computer Systems Security from Glamorgan University (UK) in 2009 and 2010 respectively. Abdulrahman's research interests include Cybercrime, digital forensic, biometrics and privacy. He has published a number of peer-reviewed publications at international conferences and is a member of the Centre of Excellence in Information Assurance (KSA).

Prof. Nathan Clarke is a Professor in Cyber Security and Digital Forensics at Plymouth University. His research interests reside in the area of information security, biometrics, forensics and cloud security. Prof. Clarke has over 160 outputs consisting of journal papers, conference papers, books, edited books, book chapters and patents. He is the Chair of the IFIP TC11.12 Working Group on the Human Aspects of Information Security & Assurance. Prof. Clarke is a chartered engineer, a fellow of the British Computing Society (BCS) and a senior member of the IEEE. He is the author of Transparent Authentication: Biometrics, RFID and Behavioural Profiling published by Springer. Further information can be found at www.cscan.org/nclarke.

Dr. Fudong Li is a Research Fellow in Cyber Security and Digital Forensics within the Centre for Security, Communications and Network Research (CSCAN) at the Plymouth University, where he previously completed a BSc(Hons.) degree in Computer System and Networks, an MRes degree on the subject of Network Systems Engineering and a PhD degree in Behaviour profiling for mobile devices. His research interests are behaviour profiling, user authentication / intrusion detection techniques for mobile devices, biometrics and digital forensics.

Prof. Steven Furnell is the head of the Centre for Security, Communications and Network Research (CSCAN) at Plymouth University (UK), an Adjunct Professor with Edith Cowan University (Western Australia) and an Honorary Professor with Nelson Mandela Metropolitan University (South Africa). His interests include cybercrime, mobile device security, user authentication, and security usability. Prof. Furnell is the author of over 270 papers in refereed international journals and conference proceedings, as well as books including Cybercrime: Vandalizing the Information Society (2001) and Computer Insecurity: Risking the System (2005). He is also the editor-in-chief of Information & Computer Security, and the co-chair of the Human Aspects of Information Security & Assurance (HAISA) symposium

(www.haisa.org). Steve is active in a variety of professional bodies, and is a Fellow of the BCS, a Senior Member of the IEEE, and a Board Member of the IISP. Further details can be found at www.plymouth.ac.uk/cscan, with various security podcasts available via www.cscan.org/ podcasts.

Reference to this paper should be made as follows: Abdulrahman Alruban, Nathan Clarke, Fudong Li and Steven Furnell (2016). Leveraging Biometrics for Insider Misuse Identification. *International Journal on Cyber Situational Awareness*, Vol. 1, No. 1, pp130-151.