Jurnal Teknologi

Full paper

Printed Document Integrity Verification Using Barcode

Affandi Husain*, Majid Bakhtiari, Anazida Zainal

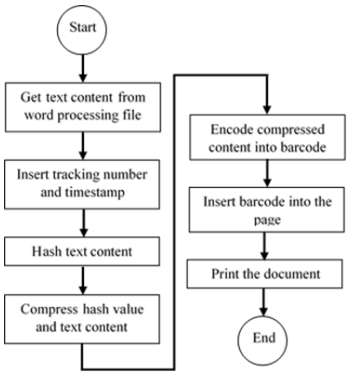
Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: affandi03@gmail.com

Article history

Received :13 April 2014
Received in revised form :
5 June 2014
Accepted :15 August 2014

Graphical abstract



Abstract

Printed documents are still relevant in our daily life and information in it must be protected from threats and attacks such as forgery, falsification or unauthorized modification. Such threats make the document lose its integrity and authenticity. There are several techniques that have been proposed and used to ensure authenticity and originality of printed documents. But some of the techniques are not suitable for public use due to its complexity, hard to obtain special materials to secure the document and expensive. This paper discuss several techniques for printed document security such as watermarking and barcode as well as the usability of two dimensional barcode in document authentication and data compression with the barcode. A conceptual solution that are simple and efficient to secure the integrity and document sender's authenticity is proposed that uses two dimensional barcode to carry integrity and authenticity information in the document. The information stored in the barcode contains digital signature that provides sender's authenticity and hash value that can ensure the integrity of the printed document.

Keywords: Barcode; digital signature; printed documents; forgery; falsification; authenticity; integrity

Abstrak

Dokumen bercetak diperlukan dalam kehidupan seharian dan maklumat di dalam dokumen tersebut perlu dilindungi daripada sebarang ancaman dan serangan seperti pemalsuan, atau pengubahsuaian yang tidak dibenarkan. Ancaman tersebut boleh menyebabkan integriti atau kesahihan dokumen tersebut dipertikaikan. Terdapat beberapa kaedah yang telah dicadangkan dan digunakan bagi melindungi kesahihan dokumen bercetak. Namun begitu, kebanyakan kaedah tersebut kurang sesuai digunakan secara umum disebabkan oleh proses yang amat rumit, kesukaran untuk mendapatkan bahan-bahan khas untuk menjamin kesahihan dokumen dan memerlukan kos yang tinggi. Kertas kajian ini telah mengkaji beberapa teknik yang digunakan di dalam keselamatan dokumen bercetak seperti teknik *watermarking* dan kod bar serta kesesuaian penggunaan kod bar dua dimensi bersama-sama dengan teknik pemampatan data dalam pengesahan dokumen. Satu konsep yang mudah dan cekap telah dicadangkan untuk menjamin integriti dan kesahihan penghantar dokumen menggunakan kod bar dua dimensi. Maklumat yang disimpan dalam kod bar mengandungi tandatangan digital yang memastikan kesahihan penghantar dokumen dan nilai cincang (*hash*) yang dapat memastikan integriti dokumen yang dicetak.

Kata kunci: Kod bar; tandatangan digital; dokumen bercetak; pemalsuan; kesahihan; integriti

1.0 INTRODUCTION

Data is a collection of facts and statistical that can be used for reference, analysis and as knowledge [1]. Data has been preserved in digital and analog forms. In analog, paper documents such as newspaper, books, printed records and legal documents are used. But the world is becoming paperless and data preservation in digital form has been widely used. However, printed documents to represent information such as birth certificates, educational transcript and certificates, land titles, official letters, wills and contracts are still relevant.

Data or information must be secured from attacks or unauthorized access. There are many techniques that have been used to protect digital data and one of the available techniques that is widely used is cryptography. There are also steganography techniques which are used for data hiding into objects such as images, audio, video and within text. Data in analog forms are susceptible to attacks. Threats such as counterfeiting and forgeries are common threats to printed document security [2][3]. Printed documents are often being forged, altered or faked to deceive intended parties who thought that the documents are real in order to gain benefits from it such as cases happened in US where two criminals mistakenly released due to forged court orders [2] and illegal immigrants from India able to seek employment in Malaysia using counterfeit social visit passes [3].

This study is organized as follows; Section 2.0 describes current security issues in printed documents, Section 3.0 focuses on related works in printed document security, Section 4.0 discusses the proposed conceptual solution and Section 5.0 concludes the study and highlights the future works.

■2.0 CURRENT SECURITY ISSUES IN PRINTED DOCUMENT

Printed documents also need security protection. The information or data inside printed documents are subjected to threats such as forgery and counterfeiting. Despite controls taken to circumvent document forgery and falsification, attacks to printed documents still succeeded. This is due to the lack of practical generic authentication techniques that can be use generally on all kinds of printed document. Document forgery cases are reported in many parts of the world. In United States, document forgery is one of the common problems faced by the authorities and private companies especially on health insurance where The National Health Care Anti-Fraud Association (NHCAA) estimated that United States of America lost 3% to 10% of total healthcare cost to fraud [4]. One of the healthcare insurance fraud case related to healthcare document forgery happened where former medical record director of Health Care Solutions Networks Inc. (HCSN) found guilty of conspiracy on 25 April 2013 that resulted a total loss of USD63 million on Medicare and Medicaid Savings Programs in Florida [5]. Another document forgery case reported on 18 October 2013 in Orlando, Florida where two inmates that were life sentenced without parole for murdering had mistakenly released by the guards at Franklin Correctional Institutes, Panhandle Community of Carrabelle, Florida after being shown with forged documents. The falsified documents forged with fake signature of Orlando-state attorney or the assistant state attorney and the court orders were filed with Orange County Clerk's office looks legitimate with the fake county's seal and forged signature of US famous judge. As the result, the two killers were released on 27 September and 8 October respectively [2].

There are also document forgery cases reported in Malaysia. One of the mainstream newspaper reported that a statement given by the Congress of Unions of Employees in the Public and Civil Services (CUEPACS) stated that more than 45,000 or 3% of 1.5 million government's staff in Malaysia forged medical certificate as a reason to absent from work to do part-time jobs [6]. Another recent fraud case related to document forging is when Johor Immigration Department successfully busted a syndicate counterfeiting the social visit passes using special ink obtained from India on 05 November 2013. The fake document produced looks like a genuine social visit pass, but Immigration's trained officers are able to detect the defects and deficiencies on the fake social pass [3].

There are few initiatives taken to strengthen the document security. In digital data security, there are few hash algorithm used for integrity protection such as Secure Hash Algorithm (SHA), Message Digest 5 (MD5) and RACE Integrity Primitives Evaluation Message Digest (RIPEMD). While authentication that verifies the originality of the digital data can be achieved by applying digital signatures and challenge-response authentication. But for the integrity and authentication of printed documents are something that is still consider a challenge in information security. Security printing is the field in printing industry that prevents forgery, tampering or counterfeiting of printed documents or items such as banknotes, certificates, passports and others. Security printing uses few technical techniques such as the use of special paper like synthetic fibers, special watermarks that looks lighter or darker when viewed with lights from behind, intaglio printing that makes images in the printed document raised, micro printing of endless text, optically variable color-changing inks, holograms, security threads, fluorescent ink that can be seen using ultraviolet light and rainbow coloring which subtly blends colors together to create gradual color change effect [7]. There are other techniques that have been developed by researchers to overcome this integrity verification problem such as optical watermarking technique on printed document that formed by the superposition of multiple two dimensional binary image with different carrier structural patterns that embeds secret information to verify the integrity of the documents [8]. A similar study with has been done where two dimensional barcode that carried hash value for printed document integrity verification has been implemented which is good and inexpensive [9]. But the drawback of this technique is no verification of whether the printed document comes from the real sender, thus creating authentication loophole where forger can recreate the printed document with undetected fake information because the hash value can also be recreated. Most of these techniques require deep knowledge, the implementation is costly and only suitable for certain documents such as birth certificates and bank notes but cannot be implemented for printed documents in general such as official letter, medical certificate and academic transcript.

■3.0 RELATED WORKS

Few techniques have been used to prevent document forgery and falsification. Security printing is one of them but it is costly and hard to implement because it needs special machines or materials and also require deep knowledge in using the printing techniques. This section will discuss two common techniques that are widely used to verify the integrity of printed documents which are watermarking technique and verification using barcode technique. Several watermarking techniques and barcode techniques that have been used to ensure integrity and authenticity of printed documents will be discussed.

A. Integrity Verification Using Watermarking Techniques

Watermarking technique have been widely used in preserving document copyright and authenticity. Watermarking can be applied in both digital format and printed form. Like steganography, watermarking employs the concept of data hiding into media such as image, audio and text where watermarking emphasize more on robustness instead of the ability of not being perceived by human visual and auditory. Watermarking technique generally can be divided into digital watermarking and printed watermarking. Digital watermarking involves embedding watermark containing key information such as copyright ownership or intellectual property into the original information [10]. The original information will maintain its states after embedded with the watermark. In digital watermarking, there are four categories of watermarking which are image watermarking, audio watermarking, video watermarking and text watermarking. The principle within the first three non-text watermarking techniques are the same where the redundant spaces in the image, audio or video have been used to carry the secret information, but it is different for text watermarking. The lack of redundant spaces in text media makes it harder to implement compared to image or audio [11]. Printed watermarking is also known as physical watermarking. It inserts the watermark into physical form of document, mostly through printing. Printed watermark is recognizable image or pattern in paper that becomes visible when viewed with light behind the paper due to variations of thickness or density in the paper [12]. [13] mentioned that watermark is where a hidden image is embedded in the form of dot arrays or halftones which is invisible to normal human sight but can be decoded or viewed by using means of periodic phenomena such as absorptive grating, lenticular screen or sampling of copying system. There are only two watermarking categories that are applicable to printed media which are image watermarking and text watermarking.

Image watermarking uses cover image to hide watermark image or text as the secret information used to verify the copyright or authenticity of the document as its main objective. There are two main processing techniques used to hide the watermark which are spatial-domain technique such as least significant bit (LSB) and transform-domain technique based on human visual system (HVS) which consists most commonly used transforms such as discrete cosine transform (DCT). However, only several image watermarking techniques that can be applied in printed document. Spatial-domain technique uses gray level of cover image pixels to embed the watermark bits directly into the image which is simpler to implement and able to carry high amount of watermark bit. The watermark is less robust against noise and attacks. Transform-domain technique converts the cover image using defined transforms algorithm where the transformed image or the coefficients of the transformation are used to hide the watermark bits. This technique is more robust towards noise and print-and-scan operations which makes it more suitable for printed document [14]. The transform domain which uses DCT-based technique to hide the watermark bits were introduced by [15] and [17]. Other HVS-based techniques have been proposed for color printed document watermarking insertion that uses color modulation and whitening filter to embed the watermark. The retrieval of the watermark is done using correlation operator based on 2D fast Fourier transform [16]. From the researches that have done, it shows print-and-scan image watermarking technique is able to protect the copyright of printed documents, however the design and implementation of the algorithms are complicated and there are issues related to imperceptibility and robustness against attacks.

Text is the most important multimedia object because most of the information in documents are written in text. Text watermarking uses text content in the document to hide secret information that will be used to verify the authenticity or copyright of the document. As discussed by [18] and [19], there are four approaches used in text watermarking which are image-based approach, syntactic approach, semantic approach and structural approach. One of the advantage of text watermarking is that text watermarking can be used in both digital and printed form provided that blurry effect and geometric distortion on printed document are being handle effectively. Image-based approach in text watermarking uses the image of text content in the document itself as the source for embedding watermark. The first technique was proposed by [20] where the authors created three types of image-based non-blind text watermarking techniques which are line-shift coding which use text line to insert watermark by shifting it vertically, word-shift coding which use spaces between words in the text line by shifting the words horizontally and feature coding by modifying the feature of chosen characters in the text to insert the watermark bits. Similar work done by [21] by using sine wave(s) while [22] and [23] uses feature coding in Arabic / Persian characters to embed the watermark bits. Syntactic approach in text watermarking focuses on the syntactic structure of sentences in the text document to embed the watermark. It uses natural language processing (NLP) algorithm to examine the syntactic structure of the sentences where syntactic sentence tree is built and transformation of the sentence syntax is applied for watermark embedding process without altering the meaning of the sentence itself. Syntactic transformation paraphrased sentences without changing its initial meaning such as active-passive transformation, sentences' complexity transformation and extra-position. The technique was theoretically described by [24] which embeds watermark into natural language text. The proposed technique transforms the original natural language text, T to insert watermark information, W and generates watermarked natural language, T' which has the similar meaning with the original text. The watermark insertion depends on the value of quadratic residue modulo p, the secret key. Other works that have used this approach are [25] and [26] that have introduced natural language watermarking based on Chinese text syntax and also [27] which has implemented morpho-syntactic alteration method for English and Turkish language for watermark bits insertion.

The third approach in text watermarking is also based on NLP of the text document. Semantic approach uses meaning of words instead of sentences to embed the watermark. Sentences component such as acronyms, verbs, nouns, grammar rules, word spelling, prepositions and others are manipulated to embed watermark bits without changing the meaning represented in the text document. The first technique based on this approach was also developed by [28] using semantic tree representation which modifies text meaning representation (TMR) and syntax of the sentences in the text document. TMR is used to map each word in the sentence to the semantic tree using ontology knowledge base. This approach has been researched further by [29] which manipulated presupposition in the sentences and transformation of text syntax to embed the watermark bits. [30] and [31] uses synonym substitution in English text language and morpheme segmentation of Korean text language respectively to insert the watermark into the original text document. The fourth approach in text watermarking is structural approach. It is based on frequency of double letters occurrences in words of the text document. [32] mentioned that in English language text there are almost 7% to 11% words in text content contain double letters. There have not been many researches done in this field yet. From the reviews conducted, it is known that the first published work was proposed by [19] where the watermark is generated by manipulating the existence of double letters and a watermark image which contains text in the image. The image indicated the ownership of the document and converted into watermark alphabet using min-max normalization technique. A maximum occurring first letter (MOFL) list which was used to construct watermark key and stored by Certification Authority (CA) as the trusted third party for verification purpose later on.

Four approaches in text watermarking had been discussed in this subsection. The approaches have its own strength and disadvantages. Image-based approach is a simple approach that does not require complex computation and algorithm to be applied. The watermark embedded using these techniques are also hard to perceive by human visual system to decode the embedded watermark. However, in most cases the watermarks embedded by using these techniques could be destroyed when converting the image back into text using optical character recognition (OCR), or retyping attack. Most of the approaches need to have linguistic tools for watermarking processes and deep knowledge on the language that are being studied, need to use complex algorithm to be implemented and there are limitation on the amount of watermark bits that can be embedded in the text document. Three of the approaches are also heavily depended on specific language on embedding the watermark bits. There are also robustness issue that need to be taken care so that the watermark resilience to the attacks that can destroy the watermark embedded in the text document. The next subsection will discuss another printed document verification technique that has been recently researched as an alternative to watermarking technique.

B. Integrity Verification Using Barcode Technique

Several research have explored the applicability of barcode in data representation, especially two-dimensional barcode that can hold more data than one-dimensional barcode. Cryptography

techniques have been used with the barcode to strengthen the security. Few integrity verification techniques using barcode will be discussed and they are Identity Document Authentication Based on Visual Sharing Secret and QR Code [33], Printed Document Authentication Based on Public Key Encryption and 2D Barcodes [34] and also Two-Dimensional Printed Document Integrity Verification [9].

[33] studied authentication technique based on two-dimensional (2D) barcode and visual sharing secret (VSS). The work was done based on several case study for identity document (ID) card authentication such as using timestamp comparison, scanned photograph comparison and authentication using password. The authors introduced a concept of visual secret sharing and QR Code where by using (2,2) threshold VSS scheme, a secret binary image can be represented as a characteristic of the card holder such as initial name of the card holder. The secret binary image encrypted into two shares and one of the shares is encoded into QR Code which has been printed on the ID card. The other share encoded into another QR Code and stored in a database or a mobile device. In the experiment for authentication process, a smartphone have been used to decode the QR Code on the ID card and then the other QR Code in the database server is retrieved and decoded. The two shares are stacked and the OR operation is performed. If it reveals the secret binary image, then the card is considered as authentic. Otherwise, it is considered as a fake ID card. This technique can be used to verify the authenticity of ID card, however it still can be fabricated. When the smartphone acquires a share, unauthorized person can obtain a copy of the share and use it to forge new fake ID card.

[34] implemented two-dimensional barcode as a part of their integrity verification technique. The authors proposed a new technique using public key cryptography with trusted authority and application of two dimensional barcode and claimed as is more efficient, affordable and simple. Related algorithms and elements that made up components of their approach are Integer Factorization Problem (IFP), Discrete Logarithm Problem (DLP), Chinese Remainder Theorem (CRT), ElGamal cryptosystem and security requirements for the printed document. IFP focuses on difficulty to factorize large integer into its prime divisors which is also the main theory used in cryptosystem such as RSA. Meanwhile, DLP refers to the difficulty of obtaining solution for x in formula $g^x = h$ where g and h are elements in finite cyclic group G. CRT focuses on the result of congruencies in number theoretic:

$$\begin{aligned} x &\equiv r_1 \bmod m_1 \\ x &\equiv r_2 \bmod m_2 \\ &\vdots \\ x &\equiv r_k \bmod m_k \end{aligned} \quad (1)$$

CRT is used to find x unique value by given a set of remainders, r_i by the division of x by a set of numbers, m_i. If the set of m_i are relatively prime to each other in the set, then $x = \sum_{i=1}^{k} r_i \; M_i^{-1} \; M_i$ where $M = \prod_{i=1}^{k} m_i$, $M_i = \frac{M}{m_i}$ and $M_i^{-1} M_i \equiv 1 \bmod m_i$. ElGamal Cryptosystem is one of the asymmetric cryptography that can be used to do public key encryption. In the approach, the usage of two dimensional barcode to secure the printed document is applied where the barcode contains set of required information, M which can be presented in the following equation :

$$M = \{T, N, S, H\} \quad (2)$$

T = Timestamp of the printed document issuing time and date
N = Document indexing number
S = Sender's manual signature converted into binary image
H = Hash value of the printed text content.

The trusted authority generates its own public and private key using RSA key generation technique to calculate user's identity, S_i and using CRT to calculate short term private and public key for each registered users. If two registered users want to communicate with each other, the sender uses its own user identity and receiver's public key to encrypt the required information, M by using ElGamal encryption technique and encodes the encrypted information into 2D barcode. The 2D barcode is printed on the document and sent to the receiver. The receiver uses its private key and sender's public key to decrypt the required information, M that has been used to do authentication and integrity verification. Sender's user identity ensures the authenticity of the printed document along with the sender's manual signature, timestamp and document indexing number while the hash value of the message ensures the integrity of the printed document. This technique also provided confidentiality where only authorized receiver who has the private key can read the required information. Therefore, the authenticity and integrity of printed document can be achieved. However, in their study, the types of barcode used and the technique of hashing the printed text content in order to verify the integrity of the printed document are not mentioned.

A similar research to verify integrity of printed document by using two dimensional barcode is proposed by [9]. The technique emphasizes on document integrity and it provides non-repudiation for document creation date by using tracking number of the page and timestamp of the created page. There are two processes involved, the document preparation and document verification processes. The author used several functional components which are two dimensional barcode, hashing mechanism, optical character recognition (OCR) and data compression. Two dimensional barcode, Data Matrix have been selected due to its small printout size compared to QR Code and higher error correction rate, Secure Hash Algorithm 256 bit (SHA-256) is used for hashing mechanism and also GZIP compression for efficient barcode size. In document preparation process, the page content, M with a set of required information are hashed and both values are encoded into Data Matrix barcode and printed on the document as shown in Figure 1.

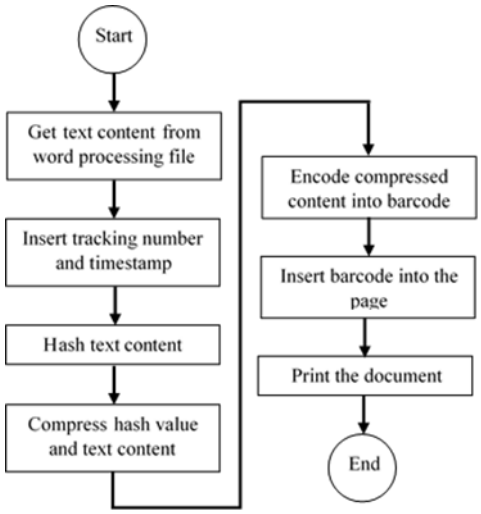


Figure 1 Document preparation process

During the verification process, receiver of the printed document reads the barcode by using scanner or any optical barcode reader to get the page content document and the hash value. The printed document is scanned and the page content retrieved from OCR process were compared to the page content decoded from the barcode. The next step is to compare the hash value from decoded barcode with the hash value of page content from the OCR process. Series of testing were done which include detection tests on text content modification and barcode modification. In text modification simulation, the page content, serial number and timestamp were modified and through the document verification process, all the modifications were detected. However, the result is different in barcode modification detection. Due to error detection and correction in Data Matrix barcode, the test failed because the mechanism corrected the altered pixel in the barcode and the barcode decoded successfully. The author concluded that the proposed technique is able to preserve the integrity with the timestamp to show the time of page created and serial number acted as identification number of the page which is created by the user. The technique was simple and effective against modification of the document, however man-in-the-middle can recreate the document, hashes the fake page content with the fake timestamp and serial number, re-encode and print the barcode into the fake document before sending it to the intended receiver. The receiver could counter check by contacting the printed document sender or creator to check the serial number with the timestamp, provided that the sender / creator of the printed document saved the document's information officially which is quite troublesome for the receiver. Another weakness is that the forger can somehow use the original timestamp and serial number information obtained through social engineering attacks by pretending as the authentic receiver to recreate printed document with fake text content. The sender or creator cannot deny it due to the real information used in the document.

This section has discussed and explored various watermarking and barcode techniques used in authenticating and verifying integrity of printed documents. Each of the techniques has its own strength and weaknesses. This study focuses more on protecting printed document integrity and authenticity by using two-dimensional barcode instead of watermarking technique. This is because most of the watermarking technique involves complex preparation, detection and verification processes and more suitable in protecting the digital based document instead of printed document. Moreover, most of the techniques do not cater for the authenticity of the document's sender. In the next section, a conceptual solution will be presented as an alternative to the work that have been done by [9] by integrating the concepts of two dimensional barcode usage, OCR, data compression, hashing and also digital signature to secure integrity and authenticity of printed documents. This proposed approach is more cost effective, practical and at the same time provides protection of document's integrity and authenticity of its sender.

4.0 CONCEPTUAL SOLUTION

This section will describe the conceptual proposed solution for hardcopy document authentication where there are five components involved in this technique. The system design which consists of two processes will also be discussed in this section. This design provides protection mechanism for integrity and sender's authenticity of the printed document, thus prevent document from being forge or falsified by unauthorized party. These components has been chosen to carry the integrity and authenticity information of the document so that the receiver can perform the verification in a simple manner without the need to go online or the hassle of contacting the sender for document's serial number or timestamp for the verification. This method is expected to solve the weakness found in the previous work proposed by [9] where forger is able to recreate the document with the timestamp and serial number obtained from the sender. The digital signature will provide sender's authenticity to prove that the document received is really originated from authentic sender. Forger will not be able to recreate the document as the forger need to have private key of the authentic sender.

A. Components of Proposed Solution

Barcode is a representation of data that can be read by optical scanner device. The environment where the printed documents are kept is not as bad as public or factory environment where code may easily get dirty or corrupted, so it is sufficient to use code with lower error correction rate such as Quick Response (QR) Code with Level M error correction level to have smaller size of barcode. Assuming that generic documents are often handled with care, QR Code is suitable to be used in storing data for printed document authentication process because it can hold more data with high speed readability and sufficient error detection and correction level.

Data compression can prove useful to store more information in limited space. It can optimize two dimensional barcode's data capacity usage and may reduce the size of the barcode. There are four classes of data compression technique which are lossless compression, lossy compression, predictive compression and transform compression [35]. In this proposed solution, lossless data compression where data cannot afford to be lost will be used. There are two types of compression techniques and they are statistical compression and dictionary based compression technique [36]. There are few algorithms developed in each techniques which bit per character (BPC) comparison of few considered algorithms in statistical technique to measure text compression performance are performed. In the statistical compression technique comparison, Arithmetic Coding algorithm is the best text compression performance compared to other algorithms [36]. Although Arithmetic Coding has the best compression ratio, regularly used compression software such as BZIP2 and GZIP use composite compression algorithm, which combines more than one compression algorithm to adapt to the regularities of data in real world usage [35]. In this study, BZIP2

compression software that uses Burrows-Wheeler transform and Huffman coding will be used to do data compression.

Optical Character Recognition (OCR) is one of the components in this proposed solution. Hardcopy documents contains valuable information which integrity need to be checked by scanning and extracting the information from scanned printed form into digitized form using OCR. OCR recognizes visible characters in the scanned image as text and convert it to editable text to be stored into computer. OCR uses complex algorithm in character recognition and heavy image processing steps that require a lot of computational resources. OCR can be viewed as integration of few subsystems that have its own functionality. There subsystems involved in process order are image capturing, image pre-processing, image segmentation, feature extraction and character recognition [37]. OCR is useful in converting information in the printed document into digital text for integrity verification process using hash algorithm and as an input in digital signature scheme that will be described in next sub topic.

Hash function is one way cryptosystem that provides authenticity and reduction of arbitrary protected information length into a smaller fixed length and also optimization for digital signature scheme [38]. In this proposed solution, hash function will be used in integrity verification and also as an input to the digital signature scheme to verify the authenticity of the printed document. There are few widely used hashing algorithms and Secure Hash Algorithm 256 Bits (SHA-256) will be selected as this algorithm is resistant towards collision, differential and linear attack [39]. SHA-256 is an iterated cryptographic hash function with compression function developed based on Merkle-Damgard scheme and Davis-Meyer mode [40]. SHA-256 digests 512 bit message block which are divided into 32 bit words and appended with padding block to make the total padded message length plus 128 bit of unsigned integer length field block to be a multiple of 512 bit message block size. Each message block will be digested into 256 bit message digest.

Digital signature is the final component in this solution. Besides integrity, other security goals that need to be achieved in this solution is the printed document sender's authenticity. To prove that a hardcopy document is originated from the correct sender, conventional signature is included in the hardcopy document to be compared with the previous signature on authenticated document. Digital signature has the same concept but for digitally signed document, the digital signature is not included in the document, it will be sent as a separate document [41]. Digital signature uses the concept of asymmetric key encryption technique where the sender signs the document with the private key and receiver will verify it with the sender's public key. This proves that the document is really originated from the sender because no one has the private key besides the sender. However, asymmetric key encryption technique uses a lot of computational resources and inefficient when dealing with large amount of data. Instead of original message, only the short message digest is signed to make the digital signature scheme more effective. This is when hashing algorithm is introduced. Verification still can be done by verifying the message digest in the digital signature with the message digest of received message. There are several digital signature schemes available, and Digital Signature Standard (DSS) scheme will be used. DSS needs both private and public keys for signing and verifying processes. Public key of DSS contains four variables which are e_1, e_2, p and q, whereas private key of DSS is d. In the signing process, DSS performs two functions to create two signatures where one of the signature will be compared to another in verifying process.

B. System Design

The proposed document authentication technique consists of two core processes; the document creation process and document verification process. QR Code is used to store two signatures that will be produced using DSS and the hash value of the information in the printed document using SHA-256. The public key management will use current approach and method that have been developed such as controlled trusted center approach, Certification Authority (CA) using X.509 certification format and public key infrastructure model. DSS security scheme is sufficient to prove that the printed document is originated from the correct sender. While hash value will provide integrity verification of the printed document to prove that there are no modification or alteration of the document. With only two signatures and one hash value coupled with data compression, the QR Code printout can be small and will not take a lot of space in the document.

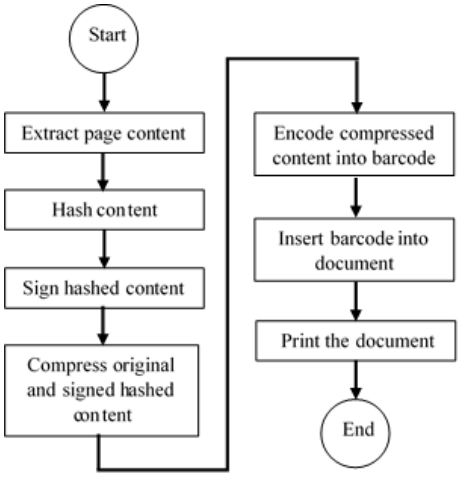


Figure 2 Document creation process

In document creation process, the sender will first create the private and public key that will be used in this process and also document verification process later on. The text content of the printed document will be hashed using SHA-256 algorithm and this hash value will also be used as an input in the signing process. The sender then creates digital signature for the document using its own private key and the hash value obtained earlier. The two signatures, S_1 and S_2 and also the hash value, H will be compressed using BZIP2 compression software and encoded into QR Code. The generated QR Code will be printed on the document by using any laser printer before the printed documents sent to the intended receiver. The QR Code will be the information carrier in this authentication process. Figure 2 shows the document creation process.

In document verification process, the receiver need to scan the document using any scanner with density not less than 600 dpi. After that, the text content in the scanned image will be extracted by using OCR software and have to be verified manually by the receiver as the OCR software only have 98% accuracy in character recognition and there might be unwanted character that may represent images or other component in the printed document. Once all text content is retrieved, the text content will be hashed using SHA-256 to get the hash value, H'. Next the QR Code needs to be read by using barcode scanner or normal image scanner which is included in the document to ensure the integrity and sender's authenticity. The information from the QR Code then need to be uncompressed back to its

original state that contains the signatures and the hash value. By using the public key published by the sender, the receiver need to calculate V from the second signature, S_2 to compare it with the first signature S_1 as stated in the digital signature scheme to ensure the sender's authenticity. The hash value, H then will be compared with the obtained hash value form the printed document, H' to verify the integrity of the text content in the document. If the one of the comparison failed, it means that the document is not the original one or has been altered. Figure 3 illustrates the process of the document verification.

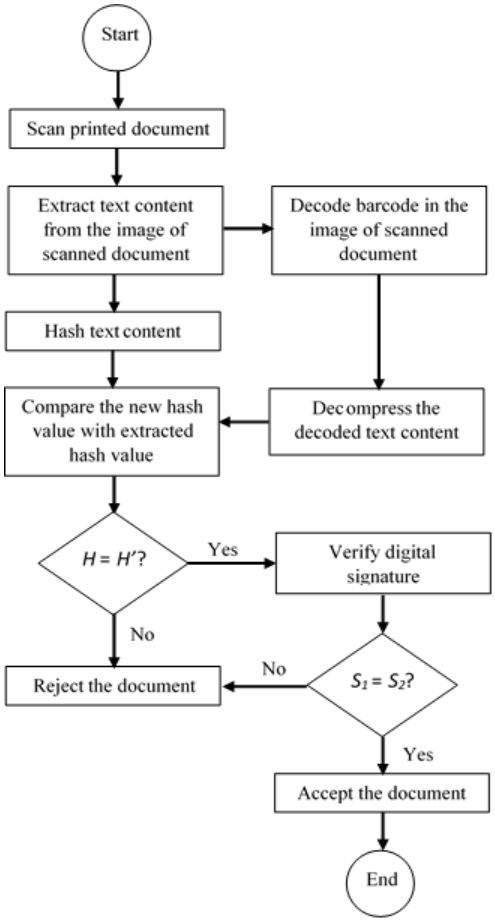


Figure 3 Document verification process

The evaluation of the conceptual solution will use dataset derived from any available text document with minimum amount of image to develop a test case that will consists 10 samples of text document. The dataset will be separated into 3 groups based on the size of the text document which are Small Size Text (SST), Medium Size Text (MST) and Large Size Text (LST). SST contains about 20-500 words, MST contains 501-1,500 words while LST contains about 1,501-5,000 words. About 7 out of 10 samples will be randomly chosen for unauthorized modification attacks and the other three, one from each group will remain as original document. Based on the studies done by [42] and [43], there are few ways to induce document forgery on printed document as mentioned below:

- Print, Copy and Paste (PPC) forgery. The forger will print a new set of characters or words that are needed on a piece of paper, cut it and paste in on the original document to replace the unwanted characters or words. Finally the forger will scan the document using a color copier machine to reproduce the text document. This type of forgery are often used by non-technical person.
- Imitation forgery. The forger simply try to insert or modify the current character or word by finding the same document font properties or font that almost similar to the original text document font. Sometimes the changes are unnoticeable using normal eyesight especially for small font or text document with a lot of words.
- Reversed Engineered Imitation (REI) forgery. This type of forgery is also using the imitation concept. The forger will need to obtain the genuine document and scan so that it can be used as template for creating new fake document by retyping all text, inserting the logo and others. REI is suitable for form-based text document such as invoices, receipts, application forms and other form-based documents.
- Scan, Edit and Print (SEP) forgery. This is one of the common way for computer-literate person. The forger will scan the genuine document, edit the text document using image manipulation software or any other tools to insert, delete or modify certain information in the document before printing it back to create forged text document.

In this study, the samples will be forged randomly using PPC forgery and SEP forgery as these two types are commonly used in document forgery. These forged samples then will be used as test case to analyse and measure the performance of the conceptual solution.

5.0 CONCLUSION

Printed documents are still relevant and need to be protected against corruption, unauthorized modification or damaged by threats. Despite all the efforts taken to strengthen security mechanism of printed documents, document forgery and falsification cases still happen. Securities printing is not suitable for all types of document as the techniques are expensive and need special materials and complex procedure thus making it impractical for implementation. Some of the techniques from previous researches are applicable to printed documents, but they suffer from easiness of use and usability in order for it to be adopted by public. Watermarking technique and integrity verification using barcode technique are among security techniques that have been discussed in this study. This study provides printed document security mechanism by using two dimensional barcode and digital signature scheme. The usage of two dimensional barcode in document authentication and integrity verification has been explored. Other important components such as hash algorithm, data compression technique, digital signature scheme and optical character recognition (OCR) have been discussed as well. The system design integrates all the components in creating and verifying the hardcopy document. However, there are still improvements that can be done to this proposed conceptual. In future work, this proposed solution will be implemented and tested for its feasibility in protecting the integrity and authenticity of printed document.

References

[1] Oxford Dictionaries. *Data: Definition of Data in Oxford Dictionary (British and World English)*. Retrieved on 26 January 2014 from http://www.oxforddictionaries.com/definition/english/data.

[2] Netto, J. and Carter, C. J. 2013, October 18. *Official: Forged Documents Used In Prison Break From Fla. Prison. Channel News Network (CNN)*. Retrieved on November 20, 2013 from http://edition.cnn.com/2013/10/16/us/florida-inmates-mistakenly-freed/v.

[3] Kim, C. B. 2013, November 05. *Johor Immigration Busts Syndicate Producing Fake Social Visit Passes*. New Straits Times. Retrieved on November 22, 2013 from http://www.nst.com.my/latest/johor-immigration-busts-syndicate-producing-fake-social-visit-passes-1.392796.

[4] Simborg, D. W. 2008. Healthcare Fraud: Whose Problem is it Anyway? *Journal of the American Medical Informatics Association*. 15: 278–280.

[5] U.S. Department of Justice. 2013, July 8. *Supervisor of $63 Million Health Care Fraud Scheme Sentenced in Florida to 10 Years in Prison. Official Press Release*. Retrieved on November 20, 2013 from http://www.justice.gov/opa/ pr/2013/ July/13-crm-763.html.

[6] Berita Harian. 2010, May 25. *45,000 Kakitangan Awam Tipu Sijil Cuti Sakit.* Retrieved on November 20, 2013 from http://www.bharian.com.my/articles/45_000kakitanganawamtipusijil cutisakit/Article.

[7] Consilium. 2009. *The Council Of The EU Glossary Of Security Documents, Security Features And Other Related Technical Terms*. Retrieved on 29 January 2014 from http://prado.consilium.europa.eu/en/glossarypopup.html.

[8] Huang, S. and Wu, J. K. 2007. Optical Watermarking for Printed Document Authentication. *IEEE Transactions on Information Forensics and Security*. 2(2): 164–173.

[9] Yew, T. C. 2008. *Two-Dimensional Barcode For Printed Document Integrity Verification*. Master, Universiti Teknologi Malaysia, Skudai.

[10] Oliveira, A. L. 2001. Techniques For The Creation Of Digital Watermarks In Sequential Circuit Designs. *Computer-Aided IEEE Transactions on Design of Integrated Circuits and Systems.* 20(9): 1101–1117.

[11] Yawai,W. and Hiransakolwong, N. 2013. Grid-Line Watermarking: A Novel Method For Creating A High-Performance Text-Image Watermark. *Journal of Science Asia*. 39(4): 423–435.

[12] Biermann, C. 1996. *Paper and Its Properties*. Handbook of Pulping and Papermaking London : Elsevier. 158–189.

[13] Renesse, R. L. V. 2002. Hidden And Scrambled Images: A Review. Proceedings of the 2002 SPIE Optical Security and Counterfeit Deterrence Techniques IV. 19 January 2002. San Jose, CA : SPIE, 333-348.

[14] Daraee, F. and Mozaffari, S. 2014. Watermarking In Binary Document Images Using Fractal Codes. *Journal of Pattern Recognition Letters*. 35: 120–129. Elsevier.

[15] Cox, I. J. *et al.* 1996. Secure Spread Spectrum Watermarking For Images, Audio And Video. International Conference on Image Processing. 16-19 September. Lausanne, Switzerland : IEEE, 243–246.

[16] Mayer, J. and Simske, S. J. 2012. Modulation in the HVS Domain for Hardcopy Watermarking of Color Documents. Eighth International Conference on Signal Image Technology and Internet Based Systems. 25-29 November. Naples, Italy : IEEE, 188–194.

[17] Agani, N. *et al.* 2013. Document Authentication Using Print-Scan Image Watermarking Based on DCT (Discrete Cosine Transform) Algorithm. Information Systems International Conference (ISICO). 2-4 December. Bali, Indonesia, 465–470.

[18] Bharati, P. D. and Nitin, P. N. 2012. Text Watermarking Algorithm Using Structural Approach. World Congress on Information and Communication Technologies (WICT). 30 October-2 November. Trivandrum, India : IEEE, 629–633.

[19] Jalil, Z. *et al.* 2010. A Zero-Watermarking Algorithm for Text Documents Based on Structural Components. International Conference on Information and Emerging Technologies (ICIET). 14-16 June. Karachi, Pakistan : IEEE, 1–5.

[20] Brassil, J. T. *et al.* 1995. Electronic Marking and Identification Techniques to Discourage Document Copying. *IEEE Journal On Selected Areas In Communications*. 13(8): 1495–1504. IEEE.

[21] Huang, D. and Yan, H. 2001. Inter-word Distance Changes Represented by Sine Waves for Watermarking Text Images. *IEEE Transactions On Circuits And Systems For Video Technology*. 11(12): 1237–1245. IEEE.

[22] Shirali-Shahreza, M. H. and Shirali-Shahreza, M. 2006. A New Approach to Persian/Arabic Text Steganography. Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR). 10-12 July. Honolulu, Hawaii : IEEE, 310–315.

[23] Davarzani, R. and Yaghmaie, K. 2009. Farsi Text Watermarking Based on Character Coding. International Conference on Signal Processing Systems. 15-17 May. Singapore: IEEE, 152–156.

[24] Atallah, M. J. *et al*. 2000. Natural Language Processing for Information Assurance and Security : An Overview and Implementations. 9th ACM/SIGSAC New Security Paradigms Workshop. September, 2000. Cork, Ireland : ACM Press, 51–56.

[25] Liu, Y. *et al.* 2005. A Natural Language Watermarking Based on Chinese Syntax. In Wang, L., Chen, K. and Ong, Y. S. (Eds.). Advances in Natural Computation Berlin-Heidelberg: Springer. 958–961.

[26] Wang, H. *et al*. 2008. Natural Language Watermarking Using Chinese Syntactic Transformation. *Journal of Information Technology*. 7(6): 904–910.

[27] Meral, H. M. *et al.* 2008. Natural Language Watermarking Via Morphosyntactic Alterations. *Journal of Computer Speech & Language.* 23(1): 107–125. Elsevier.

[28] Atallah, M. J. *et al.* 2002. Natural Language Watermarking and Tamperproofing. In Petitcolas, F. A. P. (Ed.). Information Hiding Berlin-Heidelberg: Springer. 196–212.

[29] Vybornova, O. and Macq, B. 2007. Natural Language Watermarking and Robust Hashing Based on Presuppositional Analysis. IEEE International Conference on Information Reuse and Integration (IRI). 13-15 August. Las Vegas, United States : IEEE, 177–182.

[30] Yu, Z. and Liu, X. 2009. A New Digital Watermarking Scheme Based on Text. International Conference on Multimedia Information Networking and Security (MINES). 18 - 20 November. Hubei, China: IEEE, 138–140.

[31] Kim, M. 2009. Natural Language Watermarking by Morpheme Segmentation. First Asian Conference on Intelligent Information and Database Systems (ACIIDS). 1-3 April. Dong Hoi, Vietnam: IEEE, 144–149.

[32] Bhambri, P. and Kaur, P. 2014. A Novel Approach of Zero Watermarking for Text Documents. *International Journal of Ethics in Engineering & Management Education (IJEEE).* 1(1): 34–38.

[33] Miyatake, M. N. *et al.* 2012. Identity Document Authentication Based on VSS and QR Codes. The 2012 Iberoamerican Conference on Electronics Engineering and Computer Science. 16-18 May. Guadalajara, Mexico: 241–250.

[34] Eldefrawy, M. H., Alghathbar, K. and Khan, M. K. 2012. Printed Document Authentication Based on Public Key Encryption and 2D Barcodes. 2012 International Symposium on Biometrics and Security Technologies. 26-29 March. Taipei, Taiwan: IEEE, 77–81.

[35] Jain, P. *et al.* 2013. Data Compression Techniques: A Comparative Study. *International Journal of Applied Research & Studies (IJARS).* 2(2): 1–8.

[36] Shanmugasundaram, S. and Lourdusamy, R. 2011. A Comparative Study Of Text Compression Algorithms. *International Journal of Wisdom Based Computing*. 1(3): 68–76.

[37] Laine, M. and Nevalainen, O. S. 2006. A Standalone OCR System For Mobile Cameraphones. The 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications. 11-14 September. Helsinki, Finland: IEEE, 1–5.

[38] Preneel, B. 2003. Analysis and Design of Cryptographic Hash Functions. Doctor of Philosophy. Katholieke Universiteit Leuven, Belgium.

[39] Gilbert, H. and Handschuh, H. 2004. Security Analysis of SHA-256 and Sisters.10th Annual International Workshop, SAC 2003. 14-15 August. Ottawa, Canada, 175–193.

[40] Yoshida, H. and Biryukov, A. 2006. Analysis of a SHA-256 Variant. In Preneel, B. and Tavares, S. (Ed.) Selected Areas in Cryptography Berlin-Heidelberg: Springer. 245–260.

[41] Forouzan, B. A. 2008. Cryptography and Network Security (International Edition).

[42] Bertrand, R. *et al.* 2013. A System Based On Intrinsic Features for Fraudulent Document Detection. 12th International Conference on Document Analysis and Recognition (ICDAR). 25-28 August. Washington DC, United States : IEEE, 106–110.

[43] Beusekom, J. v. *et al.* 2012. Text-Line Examination For Document Forgery Detection. *International Journal on Document Analysis and Recognition (IJDAR)*. 16(2): 189–207. Springer-Verlag.