TWO-DIMENSIONAL BARCODES FOR HARDCOPY DOCUMENT INTEGRITY VERIFICATION

TEOH CHIN YEW

UNIVERSITI TEKNOLOGI MALAYSIA

ABSTRACT

Although we live in an electronic age, it cannot be denied that hardcopy documents are still needed in our life, such as land titles, application forms and contracts. However, document forgery is still happening. Forgery makes a document lose its integrity. Integrity means the assurance that the information is authentic and complete. This study proposes a simple, fast and efficient system that provides integrity protection to hardcopy documents. The Integrity Verification System uses two-dimensional barcodes to store integrity information of a document that improves integrity protection of the document. Two-dimensional barcodes have high data capacity and error correction capability. The integrity information stored in the barcode consists of the text content of the document, a timestamp and a tracking number. The timestamp is used to prove that a page is created at a certain period, whereas tracking number is used as an identification number of a document that is created by a user. However, before all these values are encoded into the barcode, they are hashed to enhance the security of the document. In addition, data compression is used in order to efficiently use the storage space of the barcode. The recipient of the document is able to verify the integrity of the document by first scanning the document. The scanned image is then converted to text using an optical character recognition (OCR) module. The text content is then compared with the integrity information encoded in the barcode. If there is any disparity, it means that the document has been tempered. Besides that, data compression techniques and two-dimensional barcodes are also analyzed. The system design is also explained to facilitate the understanding of the system operation. All the techniques in the system are organized into functional modules that build up as the system architecture. This modular structure is useful when there is any upgrading in the future. The particular module can be replaced to improve the system. The Integrity Verification System is an Add-In for Microsoft Office Word 2007 developed using Microsoft Visual Studio 2008 Professional Edition and Microsoft .NET Framework Version 3.5. The system is tested on a permanent test bed to ensure consistent results. The verification testing tests whether the system is capable of verifying a document's integrity, whereas the modification detection tests whether the system is capable of detecting forgery. The test results are then analyzed and discussed to clearly outline the system's capability in providing integrity verification services. Results has shown that grayscale images cause less OCR recognition error compared to black/white images and the barcode is high resistant to damage. This study has shown that the integrity verification of a hardcopy document can be performed with the use of two-dimensional barcode. Further studies on better data compression techniques, error correction capabilities of the barcode and other barcode standards can be conducted. The system can also be enhanced with password protection through software or administrator control.

ABSTRAK

Walaupun kini merupakan zaman elektronik, ia tidak boleh dinafikan bahawa dokumen salinan keras seperti geran tanah, borang permohonan dan kontrak masih digunakan. Akan tetapi, kes-kes pemalsuan dokumen masih berlaku. Pemalsuan menyebabkan integriti dokumen terjejas. Integriti bermaksud suatu maklumat mestilah diyakinkan tulen dan sempurna. Projek ini mencadangkan satu sistem yang mudah dan cepat digunakan, serta efisien dalam memberi perlindungan integriti kepada dokumen salinan keras. Sistem ini menggunakan barkod dua-dimensi untuk menyimpan maklumat integriti suatu dokumen. Barkod dua-dimensi mempunyai kapasiti data yang tinggi dan kebolehan pembetulan ralat. Maklumat integriti terdiri daripada kandungan teks dokumen, tanda masa dan nombor kawalan. Tanda masa digunakan untuk membuktikan bahawa suatu muka surat dicetak pada tarikh dan masa tertentu, manakala nombor kawalan digunakan untuk mengenalpasti sesiapa yang mencetak dokumen itu. Walau bagaimanapun, sebelum nilai-nilai itu dikodkan ke dalam barkod, ia dicincang terlebih dahulu untuk meningkatkan sekuriti dokumen. Selain daripada itu, data dimampatkan untuk menjimatkan muatan storan barkod. Penerima dokumen boleh memeriksa integriti dokumen dengan mengimbas dokumen ke dalam komputer. Imej dokumen yang diimbas kemudiannya ditukarkan kepada teks dengan menggunakan modul pengecaman optikal aksara. Selepas itu, teks yang diperoleh dibandingkan dengan maklumat integriti yang dikodkan ke dalam barkod. Jika terdapat sebarang perbezaan, ia bermaksud bahawa dokumen itu diubahsuai tanpa kebenaran. Selain itu, teknik-teknik pemampatan data dan pelbagai barkod dua-dimensi dianalisa. Rekabentuk sistem juga dijana untuk memberi gambaran tentang pengoperasian sistem. Kesemua teknik-teknik dibangunkan di dalam bentuk bermodul yang kemudiannya membentuk senibina sistem. Modul-modul ini menyenangkan kerja-kerja peningkatan fungsi. Suatu modul boleh ditukarkan dengan senang tanpa pengubahsuaian besar kepada sistem. Sistem ini merupakan satu *Add-In* untuk Microsoft Office Word 2007 yang dibangunkan dengan Microsoft Visual Studio 2008 Professional Edition dan Microsoft .NET Framework Version 3.5. Sistem ini diuji dengan menggunakan tapak pengujian yang tetap untuk memastikan keputusan yang konsisten. Pengujian pengesahan dijalankan untuk mengetahui kebolehan sistem dalam mengesahkan integriti dokumen. Manakala pengesanan modifikasi adalah untuk menguji kebolehan sistem dalam mengesan pemalsuan dokumen. Keputusan pengujian menunjukan bahawa imej hitam/putih menyebabkan lebih banyak ralat pengecaman optikal aksara berbanding dengan imej skala kelabu dan barkod dua-dimensi adalah lebih tahan terhadap kerosakan. Kajian ini menunjukkan bahawa pengesahan integriti dokumen salinan keras berjaya dilakukan dengan menggunakan barkod dua-dimensi. Kajian lebih lanjut mengenai teknik pemampatan data yang lebih baik, kebolehan pembetulan ralat barkod dan piawai barkod yang lain boleh dijalankan. Untuk meningkatkan keselamatan sistem, fungsi katalaluan menggunakan kawalan perisian ataupun pentadbir boleh ditambah.

TABLE OF CONTENTS

CHAPTER 1

PROJECT OVERVIEW

1.1 Introduction

Since the birth of the Internet, there has been an increase in distribution of digital documents. The Internet has revolutionized information exchanges by a click of a button. Protection of digital document has become more prevalent since the inception of digital security due to this widespread information distribution through the Internet. Many costs have been invested into research work for the design of methods that technically support the security of digital documents. Although we live in an electronic age, the usage of hardcopy documents still cannot be disregarded. We still use hardcopy documents for certificates, academic transcripts, wills, contracts, land titles and so on.

However, as we know, there have been many forgery cases of hardcopy documents over the years (Ensor, 2003 and Sim, 2007). Fake and altered documents are created to deceive anyone who is not aware of the authenticity of the documents. Currently, most methods of hardcopy document integrity protection are expensive due to cumbersome process and hard to come by materials needed. Some of these methods are not only expensive but slow as well. Hence, although it is effective, it is not an efficient way of providing integrity protection to hardcopy documents.

Whilst these weaknesses still exist in some of the current technologies, this study intends to address them by using two-dimensional barcodes for every page in a

document. The barcodes are used to store text information in a page. They can be easily generated just before a document is being printed out by any common laser printer. They are particularly useful when a person would like to verify a page's integrity. As we know, forgery does cause a significant damage in terms of trust and authenticity. Therefore, it is essential to note that integrity of important documents should be maintained to avoid any implication of the document being jeopardized by criminals.

1.2 Problem Background

Forgery cases are still occurring despite the availability of different kind of techniques to prevent or detect them. All cases of forgery are able to cause harm to anyone who are framed by the criminals. For example, it was an embarrassment to the United States of America when they did not detect fake Iraq documents that they received to prove that Iraq was developing nuclear weapons (Ensor, 2003). Another recent case was when two policemen was charged for 14 counts of forging witness statements and using forged documents at the Criminal Investigation Department (CID) office in Kluang, Johor, Malaysia (Sim, 2007).

Thus, important documents no matter in any form, digital or hardcopy requires security protection. The requirements for a complete security protection are privacy, authentication, integrity and non-repudiation. Digital document security provides integrity protection through the usage of hashing algorithms such as Secure Hash Algorithm and Message Digest 5. However, in order to maintain the integrity of a hardcopy document, a lot of cumbersome physical process is needed. Thus, this directly affects the cost and time of producing the document.

For example, active measures to detect forgery can be implemented by featuring a "smart" chip in a card or passport and producing documents using special papers and inks. There are also several integrity verification techniques that are

registered with the United States Patent and Trademark Office as described in Table 1.1.

Table 1.1 : Patented integrity verification techniques

Inventor	Title	Description
Vigano (1998)	Method for protecting against forgery sheet-like printed documents	A microcapsule layer is applied on the printed document. This layer comprises of a chemical product adapted to release an indelible mark that reveals tampering on surfaces portions of a document subjected to a pressing or rubbing tampering force.
Prakash (2004)	Document alteration indicating system and method	The invention evaluates markings in half-tone fields of the document. The markings will form an image that will help in determining whether the document has been altered. An alteration is indicated where the first dot count and second dot count are different.
Lunt *et al.* (2006)	Systems and methods for policy based printing	The document is printed using forgery detection and deterrence technologies, such as fragile and robust watermarks, glyphs, and digital signatures that are appropriate to the level of protection determined by the policy. A plurality of printers is managed by a print management system. The policy determines the protection technologies for the document to be printed and the print management system routes the print job to a printer that can apply the appropriate protections and sets the appropriate parameters in the printer. A document can be verified as an original or a forgery by inspecting the copy evidence and/or tracing information in the watermark.

Some of the methods discussed are complex and it is not easy for an end-user to just simply implement it in order to protect a document from forgery. Complexity does not only cause problems but it increases cost and time spent as well. Hence, in order to detect document forgery with a lower cost and simpler method, this study proposes the usage of symbols (two-dimensional barcodes) in every page of a document. Two-dimensional barcodes have the capability of storing more data compared to one-dimension barcodes (Hee *et al.*, 2002 and Premaratne *et al.*, 2007). Several barcodes can be logically linked in sequence to allow even more storage of data. They can also withstand errors in subsequent scans through its error correction features.

1.3 Problem Statement

Since forgery of important documents causes serious implications, several methods have been proposed to prevent or detect document forgery. Document forgery can be prevented using lamination (Databac Group Limited, 2003) or detected through the usage of chemical layers (Vigano, 1998). Whereas, some methods use print management or analyze hardcopy printouts to verify documents (Prakash, 2004, Lunt *et al.*, 2006, Lampert *et al.*, 2006 and Gupta *et al.*, 2007). However, these methods require special equipments and skills to implement. Printable digital signatures are also proposed in Korea for an online hardcopy document integrity verification application (Lee *et al.*, 2002). Besides that, digital watermarking is used to embed a watermark into an image (Song *et al.*, 2001) and a photo and fingerprint of a driver into the driver's license barcode (Noore *et al.*, 2004).

From all these studies, it can be found that efficient techniques must not be only effective but affordable and simple to implement. By such, anyone who needs integrity protection can just use the techniques without hurdling over technical complexities. This is because high technology is not value added unless it is user friendly. Thus, this study aims to provide an effective and yet simple and fast

method of document integrity verification through the usage of two-dimensional barcodes. The hypothesis of this study can be stated as:

Two-dimensional barcodes can improve hardcopy document security requirements.

1.4 Objectives

Several objectives have been set to ensure successfulness of this project. They are listed as below.

i. To design and implement a hardcopy document integrity verification program using two-dimensional barcodes.
ii. To generate or encode correct two-dimensional barcodes for each page in a document.
iii. To detect two-dimensional barcodes in a page and correctly decode them.
iv. To detect differences between the scanned page and the two-dimensional barcode content.

1.5 Theoretical Framework

There are two main processes in this application system. Figure 1.1 illustrates two flow charts for the processes. The first process is Document Preparation. This process creates a printed document that contains a symbol (two-dimensional barcode) for integrity verification. It starts by inserting a timestamp and tracking number into the document specified by the user. Then, a symbol is generated, with it containing the text content of the document and hashed value of the text content, timestamp and tracking number. The symbol is then inserted back into the document for printing.

The second process is Document Verification. In this process, the printed document is verified for forgery that might occur during document transportation. First, the document is scanned into the computer. Then, the document is feed into the OCR (Optical Character Recognition) module for recognition. Once the document has been processed, the symbol is extracted and decoded. After that, the decoded text content is compared with the document content. Finally, the hash values are compared as well. Any irregularities results as a suspected forgery.

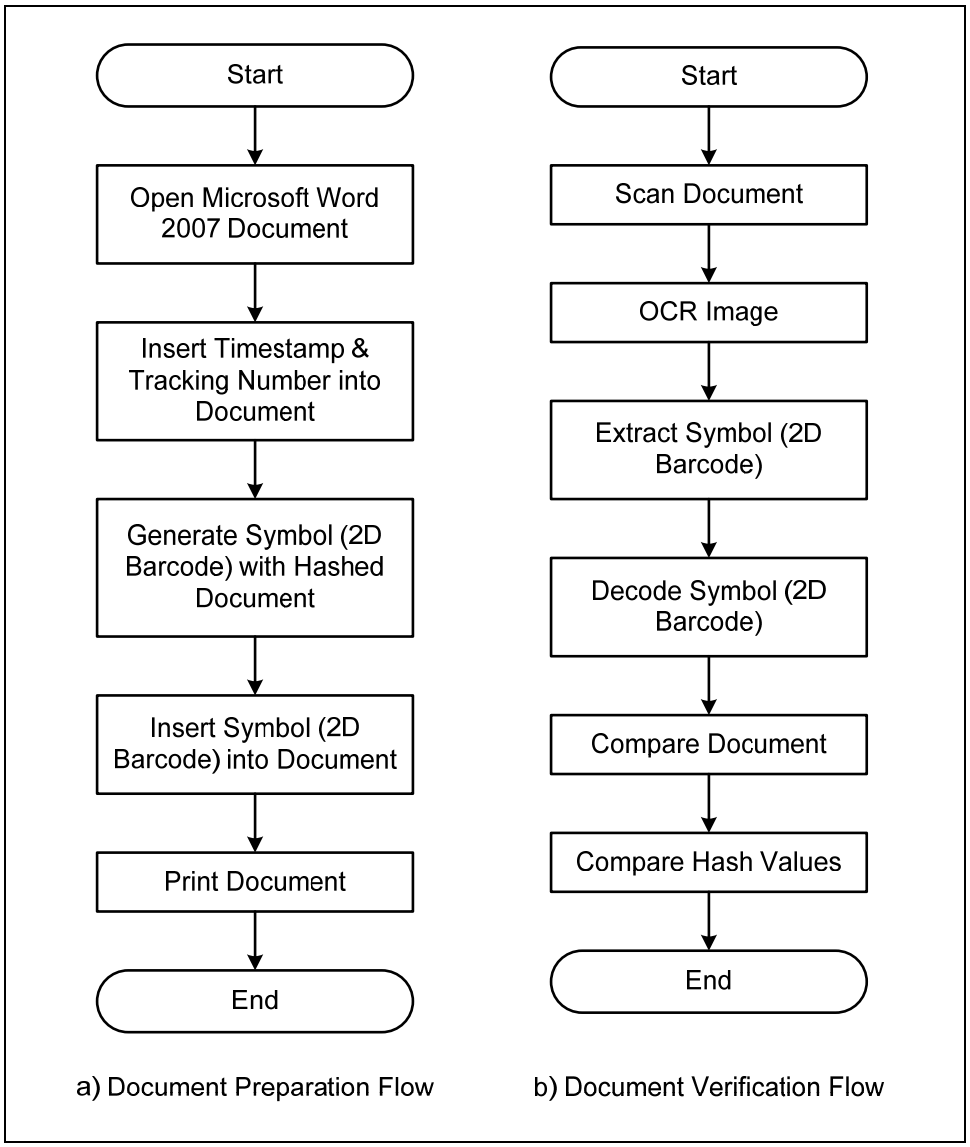

Figure 1.1 System Flow Chart

1.6 Project Scope

The execution of this project is bound to several scopes as listed below.

i. The program is an Add-In for Microsoft Office Word 2007 developed using Microsoft Visual Studio Tools for the Microsoft Office system Version 3.0, Microsoft Visual Studio 2008 Professional Edition and Microsoft .NET Framework Version 3.5.
ii. Each page in a document has at least one barcode representing the text content of the page. Only text content of a page is kept in the barcode. Hence, only text integrity is protected.
iii. Each page should be printed with at least 600 dots per inch resolution using a laser printer and scanned with at least 150 dots per inch resolution in grayscale mode into an image file of Tagged Image File Format (TIFF) format.
iv. The barcode standard that is supported is Data Matrix format.
v. The hashing technique used is SHA-256.
vi. The data compression technique used is GZIP (GNU zip).
vii. The Libdmtx barcode encoding and decoding library is used to generate and recognize the barcode.
viii. Microsoft Office Document Imaging Library (MODI) is used for Optical Character Recognition (OCR). Slight errors induced by the OCR library during the recognition process are ignored. System users are required to manually correct recognition errors.

1.7 Significance of the Project

This study discusses regarding two-dimensional barcode application in the area of hardcopy document integrity verification. The application is useful for certificates, patenting, contracts and confidential documents that are used in various institutions and organizations. The usage of this application system is to deviate people from tempering or forging important documents.

The timestamp can be used to prove that a page is created at a certain time. Whereas the tracking number can be used as an identification number of a document that is created by a user. There is no need for any public key infrastructure support because it is an offline application. The application is not only user friendly but also fast and cheap to use since only a common laser printer and scanner is needed to generate and verify the documents. This study is also useful for future studies where it can be further enhanced with more security protections.

1.8 Organization of the Report

This study is divided into six chapters. Chapter 1 describes briefly about the overview of the project and understanding of the project's problem background. It also includes the project's scopes and objectives. Whereas Chapter 2 discusses about information security, two-dimensional barcodes, data compression and hashing. The methodology of this project and its execution of tasks are explained in detail in Chapter 3. It comprises of methods, techniques and phases that are systematically arranged whereby when executed, the objectives are successfully achieved.

Chapter 4 contains explanation regarding instrumentation and software components, design and implementation of this study. Comparison of data compression techniques are made to determine the best technique that offers the best compression ratio. Barcodes are also analyzed in the sense of footprint size and data capacity. These analyses assist in making correct decisions during system designing and development. The design of the system is also explained to give a deeper understanding on how the system is built.

Chapter 5 explains results of tests conducted on the system. These results are analyzed to fine-tune the system further and to determine the capability of the system in providing integrity verification services to users. Discussions are also done to conclude the system's functions and features. Finally, Chapter 6 reviews and summarizes the whole project findings regarding two-dimensional barcodes for

hardcopy document integrity verification. Future works are also recommended to further improve the system.

1.9 Summary

So far, this study has been introduced and explained briefly. Several patents and researches have been discussed as well. Some are effective but cumbersome. Some are efficient but requires public key infrastructure, and this sure comes together with key management issues. Besides that, the objectives and scope of this study has been clearly stated as guideline for this project. Theoretical framework has been attached as well for a better understanding of this study. In order to gain more understanding about this study, more information regarding concepts, methods and techniques is discussed in Chapter 2.

CHAPTER 1 PROJECT OVERVIEW1.1 IntroductionSince the birth of the Internet, there has been an increase in distribution of digital documents. The Internet has revolutionized information exchanges by a click of a button. Protection of digital document has become more prevalent since the inception of digital security due to this widespread information distribution through the Internet. Many costs have been invested into research work for the design of methods that technically support the security of digital documents. Although we live in an electronic age, the usage of hardcopy documents still cannot be disregarded. We still use hardcopy documents for certificates, academic transcripts, wills, contracts, land titles and so on.However, as we know, there have been many forgery cases of hardcopy documents over the years (Ensor, 2003; Sim, 2007). Fake and altered documents were created to deceive anyone who is not aware of the authenticity of the documents. Currently, most methods of hardcopy document integrity protection are expensive due to cumbersome process and hard to come by materials needed. Some of these methods are not only expensive but slow as well. Hence, although it is effective, it is not an efficient way of providing integrity protection to hardcopy documents.Whilst these weaknesses still exist in some of the current technologies, this study intends to address them by using two-dimensional barcodes for every page in a document. The barcodes are used to store text information in a page. They can be easily generated just before a document is being printed out by any common laser printer. They are particularly useful when a person would like to verify a page's integrity. The printed page can be scanned and its text content can be obtained using an Optical Character Recognition (OCR) software and its barcodes can be decoded using a barcode software. Once the barcode is decoded, the barcode content will be compared with the text content. Any irregularities will result as a suspected forgery.1.2Problem BackgroundForgery cases are still occurring despite the availability of different kind of techniques to prevent or detect them. All cases of forgery are able to cause harm to anyone who are framed by the criminals. For example, it was an embarrassment to the United States of America when they did not detect fake Iraq documents that they received to prove that Iraq was developing nuclear weapons(Ensor, 2003). Another recent case was when two policemen was charged for 14 counts of forging witness statements and using forged documents at the Criminal Investigation Department (CID) office in Kluang, Johor, Malaysia(Sim, 2007). Thus, important documents no matter in any form, digital or hardcopy requires security protection. The requirements for a complete security protection are privacy, authentication, integrity and non-repudiation. Digital document security provides integrity protection through the usage of hashing algorithms such as Secure Hash Algorithm and Message Digest 5. However, in order to maintain the integrity of a hardcopy document, a lot of cumbersome physical process is needed. Thus, this will directly affect the cost and time of producing the document.For example, active measures to detect forgery can be implemented by featuring a "smart" chip in a card or passport and producing documents using special papers and inks. There are also several integrity verification techniques that are registered with the United States Patent and Trademark Office as described in Table 1.1. Some of the methods discussed are complex and it is difficult for an end-user to just simply implement it in order to protect a document from forgery. Complexity does not only cause problems but it increases cost and time spent as well. Hence, in order to detect document forgery with a lower cost and simpler method, this study proposes the usage of two-dimensional barcodes in every page of a document. Two-dimensional barcodes are chosen due to their capability of storing more data compared to one-dimension barcodes (Hee, et al., 2002;

Premaratne, et al., 2007). Several barcodes can be logically linked in sequence to allow even more storage of data. They can also withstand errors in subsequent scans through its error correction features.1.3Problem StatementAs forgery cases still occur around us (Ensor, 2003; Sim, 2007), techniques that are more efficient should be researched in order to prevent such cases from happening. Previously, several methods have been proposed using lamination and chemical layers (Databac Group Limited, 2003; Vigano, 1998). However, these methods require special equipments and skills to implement. Whereas, some methods use print management or analyze hardcopy printouts to verify documents (Prakash, 2004; Lunt, et al., 2006; Lampert, et al., 2006; Gupta, et al., 2007). Printable digital signatures were also proposed in Korea for an online hardcopy document integrity verification application (Lee, et al., 2002). Besides that, digital watermarking was used to embed a watermark into an image (Song, et al., 2001) and a photo and fingerprint of a driver into the driver's license barcode (Noore, et al., 2004).From all these studies, it can be found that efficient techniques must not be only effective but affordable and simple to implement. By such, anyone who needs integrity protection can just use the techniques without hurdling over technical complexities. This is because high technology is not value added unless it is user friendly. Thus, this study aims to provide an effective and yet simple and fast method of document integrity verification through the usage of two-dimensional barcodes. The hypothesis of this study can be stated as:How does two-dimensional barcodes improve hardcopy document security requirements?1.4ObjectivesSeveral objectives have been set to ensure successfulness of this project. They are listed as below.i.To design and implement a hardcopy document integrity verification program using two-dimensional barcodes.ii.To generate or encode correct two-dimensional barcodes for each page in a document.iii.To detect two-dimensional barcodes in a page and correctly decode them.iv.To detect differences between the scanned page and the two-dimensional barcode content.1.5Theoretical FrameworkThis study uses two-dimensional barcodes to store valuable contents of a page in a document. The security feature is further enhanced by adding a timestamp and tracking number in every page before generating the barcodes. Timestamp is used since this is an offline application that does not use a public key infrastructure and digital signature. Whereas the tracking number is used as a serial number of the pages created by the user. Besides that, the timestamp and tracking number will be concatenated, hashed and inserted into the barcodes. This means that the barcodes acts as a carrier of the text content of the page and the hash value of the timestamp and tracking number.Thus, this method does not require any chip for security data storage, special paper for printing, special printing device and lamination or hologram for integrity protection. The document that is attached with the barcodes can be printed with any common laser printer. Please refer to Figure 1.1 for a better descriptive visualization. In order to verify the integrity of a page in a document, an end-user just needs to scan the pages using any common image scanner as depicted in Figure 1.2. Once the scanned image is obtained, the OCR software will extract text contents and the barcode software will decode the barcodes. When all data is retrieved, both text contents from the OCR and barcode software can be compared to verify the document's integrity. If there are no differences in between both of them, then a hash value will be calculated from the timestamp and tracking number printed on the page. This calculated hash value will then be compared with the extracted hash value from the barcode. If there are any differences in between both of them, then the document can be suspected as a forged copy. Thus, as described, this technique is

definitely a cheaper and faster method for verification of documents such as forensic report, court order, memo or circulation and confidential documents. This study discusses regarding two-dimensional barcode application in the area of hardcopy document integrity verification. The application developed is useful for certificates, patenting, contracts and confidential documents that are used in various institutions and organizations. The timestamp can be used to prove that a page was

8. The text content of the document will then be compared with the integrity verification information in the barcode. The “Compare Result” window such as in Figure F.5 will be displayed to simplify the comparison process.
9. The user must make the necessary changes to the OCR document (OCR.docx) based on the difference from the comparison with the barcode information (Barcode.docx). Some errors might have been introduced during the OCR process due to blurred printing, font styles, font sizes, document layout etc.
10. Once the OCR document have been corrected or updated, click the “Verify Hash Value” button in the Document Verification group in the Integrity System ribbon tab to verify the integrity of the document.