International Journal of Security and Its Applications
Vol.9, No.8 (2015), pp.347-366
http://dx.doi.org/10.14257/ijsia.2015.9.8.31

Printed Document Authentication Using Two-Dimensional (2D) Barcodes and Image Processing Techniques

Mohammed AL-Gawda[1], Zou Beiji[2] and Nurudeen Mohammed[3]

School of Information Science and Engineering, Mobile Health-Ministry of Education-china-Mobile Joint Laboratory, Central South University
1. algawda@csu.edu.cn 2.bjzou@vip.163.com 3. nurudeen_saeed@yahoo.com

Abstract

Advancements in information and communication technologies have made it easier to forge printed documents these days. Forgery of printed document could have serious repercussions including financial losses, so maintaining the reliability of valuable documents is one that is timely and necessary. The document authentication approach in this study is one that involves tracing of the origin of a document through its contents. The approach explores viability of embedding certain information on printed documents for authentication purposes. Such information could be extracted from the content of documents, which can be divided into two: the context, which includes the text, figures and shapes and the visual attributes of the document, such as its color and design. In short, this study presents a practical and secure method to prevent the forgery of important documents. The proposed method makes use of the public key infrastructure (PKI) and Quick Response (QR) code and the cryptographic hash algorithm shows robustness to printing or scanning noise. The method also makes use of the unique tracking numbers such as device serial numbers, and timestamps enhance the security. To test the feasibility of the method several experiments were perform and the experimental results showed significant improvement in printed document authentication.

Keywords: *document authentication, feature extraction, digital signature, intrinsic and extrinsic signatures, Printed-and-scanned documents, visual attributes*

1. Introduction

1.1. Motivation

Despite the numerous merit of modern technology, which now makes it possible for many transactions occur with the click of a button it has also made forgery of printed documents much easier. A wide variety of legal documents that includes national identity cards, social security cards, birth certificates, driving licenses, passports, insurance documents, educational documents, wills, power of attorney, and land titles are still needed in printed format for daily life [1]. The cheap and easy availability of sophisticated scanning and printing technologies is one of the core escalators of document forgery; it is relatively easy to produce very high-quality imitations of documents at a very low cost, which demonstrates why the authenticity of documents is at great risk. [2-3].Along with cheap technological hardware, easy-to-use digital imaging software also helps the average person to counterfeit any legal document at the convenience of their homes, using a personal computer [4].

Currently, two of the most common methods used to prevent forgery rely on intrinsic and extrinsic signatures. Intrinsic signatures are responsible for securing a document with device-specific information, such as a printer model number or a device serial number. This method does not involve any security measures regarding the document content itself. The success of this approach is based on the assumption that the forged document

ISSN: 1738-9976 IJSIA

will always be processed on a device other than the originating device. Thus, a different device number or timestamp that is quite obviously different from the original is generated. The extrinsic signature approach is responsible for adding security features based on the importance of a document. Security features might include watermarks, special printing inks, security threads or holograms. Each of these security features is associated with a different mechanism for detecting forgery. Currency notes are the most common examples of secure items that use watermarks. Various legal documents are also typically equipped with additional security features as well, for example, special inks that change colors of documents when they react with other chemicals are used to print important documents [5]. Another approach is to use special types of fabric in the paper manufacturing process that, when tested, proves the authenticity of a document. Overall, security features are a crucial way to verify the authenticity of a document. However the aforementioned solutions are not cheap, usually the implementation of such security features maybe more expensive than the actual cost of the document that is being protected.

1.2. Related Work

Document verification is a highly valued research area, and several approaches have been studied to improve the authentication process and to enhance forged documentation detection methods. An intrinsic security approach proposed by Amidror *et al.* [6] integrates print-based authentication and a security strategy for the protection of valuable documents. This technique is based on more intensity profiles and mathematical transformations of the microstructure of a document to generate an attractive, dynamic visual effect. Another study by Mikkilineni *et al.* [7]was based on securing printed documents by extracting the banding signals from documents and then measuring the corresponding signals. These banding signals are used as an extrinsic feature. This approach makes use of intrinsic information, such as printer serial numbers and dates of printing. This technique requires a modification to the printing device used to reduce the ease of being hacked. Holograms and micro text are new approaches to extrinsic security that have been introduced by multiple researchers, including Wang *et al.* [8-9]. This approach is concerned with protecting a particular part of a document and is not focused on document's content. Garg *et al.* [10] proposed a method for protecting pictures in documents by generating photo signatures based on the cosine transform generated from the documents to validate their authenticity.

A framework for automatic authenticity verification was introduced by Garain et al. [11]. This approach focuses only on protecting a particular area of a document without considering other document contents. Ibrahim *et al.* [12] proposed a digital watermarking technology for protecting the copyrights of printed documents such as certificate and identification cards. This approach detects noise introduced in a document by printing or scanning processes. Noise makes the detection of an embedded watermark more difficult, thus making a forged document easier to detect. The use of digital signatures and QR codes was introduced by Warasart *et al.* [3]. One of the advantages of this approach is that it does not require any databank information for the verification of a document. On the other hand, their method does not offer multi-lingual support, and the accuracy does not approach 100%. This approach is also semi-automated due to printing or scanning distortion. Eldefrawy *et al.* [13] discussed the challenges of secure document transfer. The authors proposed an algorithm that is only responsible for protecting the issuing number and the timestamp of a document. This algorithm lacks the protection of the content of a document. Furthermore, this approach uses hash values in the verification process, which typically vary. The approach introduced by Gebhardt *et al.* [14] only detects the differences between document features produced using different printing technologies. Mikkilineni *et al.* [15] studied a way to trace documents generated on low-cost consumer printers, such as inkjet and electro photographic (laser) printers, through the use of

intrinsic and extrinsic features. This approach uses a technique for comparing the printing characteristics of different printing devices. However, this approach fails to identify the authenticity of a document if a similar printer make or model is used to forge the document. Embedded feature problem (extrinsic features) to verify the authenticity of any paper-based documents, in addition to a specialist, sophisticated equipment such as ultraviolet light, a magnifying glass, or infrared detectors are also needed. This approach requires skilled scrutiny, and the results of this approach are based on the skills of the specialist inspecting the document. This technique does not guarantee 100% detection of forged documents because the advanced software and the machines used today can easily simulate very exact features of original documents that make it very difficult to detect forgeries. The most interesting aspect of this approach is that the security features introduced into a document to verify its authenticity are not themselves secure. To reproduce these security features is fairly easy in the modern world; reproduction is also not too expensive. In short, a mechanism that embeds different features within these existing security features that are unique and cannot be easily reproduced is needed. Device feature problems (intrinsic features) in different parts of the world, many organizations are dedicated to forensic science and are heavily equipped with the latest technology facilities for detecting forged documents. However, advancements in technology and detection techniques require the facilities to be up to date with the modern approaches for detecting different security features built into documents themselves. Most verification procedures are tedious and time-consuming because they often involve many legal entities, based on pertinent local laws, which vary between countries [3]. Previous studies have focused on the loosely coupled relationship between a document's content and the document itself. It is quite possible to forge documents in this context because the core focus either lies on a document or a particular content area instead of the tightly coupled relationship between a document and its content. A modification of even a single character could have catastrophic effects on the authenticity of a document, which is quite possible to achieve. To tackle the above mentioned issues, an automated approach that can provide users the ability to process large amounts of documents conveniently and within a short period is needed. This paper proposes a solution to the above-described issues by proposing an automatic approach for detecting counterfeit or forged documents during processing. This approach also guarantees not only the authenticity of the document but also the protection of the security features incorporated into it. The technique considers the intimate relationship between a document's content and the document itself, which includes various security features. This approach has been applied to academic certificates to demonstrate its reliability.

The proposed process can be divided into two basic steps:

(1) Computationally extract the visual attributes and the integration of intrinsic and extrinsic features from a document image.
(2) Process the extracted information using an automated algorithm that identifies and localizes the tempered areas and the difference levels between the original and the forged document.

This approach relies on the visual attributes of an image that are represented by common attributes called features. These features include the shape, the color histogram and the texture of a document image.

2. Materials and Methods

This section discusses the key aspects of the security features used in this study.

2.1. Security Features

Specific security features are normally embedded when preparing any important official document. These features should not be modifiable even if modern technologies

are used. The nature of the security features varies according to the importance of the document; specifically, the importance of a document determines the number of security features required in the design for the document's protection.

2.1.1. Color Ink Features: Color features refer to the distribution of colors within an entire document [16]. These features play a key role in human perception; therefore, special colors can be used to secure documents. Due to these characteristics, color features are considered visual features that can be used in document design for the purposes of protection and verification. According to the literature, color histogram, color moment, and color coherence are several methods that utilize color to verify the authenticity of a document image. However, the most effective method is the color histogram [17].

2.1.2. Paper Security Features: Key security features of a document are often printed on the document itself. Examples include pulp of a particular color, a distinctive fiber length, invisible fibers, the smoothness and roughness of a document's surface, individual opacity, and the ability of a document to maintain its strength during folding. Each document reacts differently to different chemicals and solvents.

2.1.3. Texture features: Texture features refer to the low-level designs and textures within a document, such as graininess or smoothness. These features also refer to the grainy details of surface design patterns on documents, making them important for validating a scanned document. Different texture features can be obtained from an image by utilizing a co-occurrence matrix. Texture parameters include but are not limited to entropy, contrast, energy and homogeneity. These parameters can be utilized to enhance scanned document authentication. Gabor filters are efficient and effective in texture analysis based on wavelet transforms filter methods [18].

2.1.4. Shape Features: Shape features are a major area of interest within the field of forgery recognition and detection; these features provide powerful clues to object identity. Shape features refer to the shapes that appear in a document, as determined by utilizing region-based methods. Additionally, this approach is capable of extracting object boundary and region features. This method is generally founded upon moment- based theory, which involves the use of Hu, Legendre and Zernike moments. These moments afford valuable information for characterizing an object in a document image designed for shape feature extraction [18].

3. Feature Extraction

As mentioned in section 2, the visually recognizable protection features in documents are represented by four essential attributes:

i. Color features
ii. Background artwork (texture features)
iii. Shape features
iv. Paper quality

This study only focuses on the first three attributes. First, five different traits are considered for color- associated security. Second, two further features are calculated from the texture pattern; finally, one trait for shape-related security features is used. The reason for choosing these nine features is to distinguish a genuine document from a counterfeit document, as discussed below in section 3. The computation of these features, when embedded in a QR code, is completed and protected by storing the encrypted feature values when a document is being prepared. These values can be used for document protection purposes.

3.1. Color Features

3.1.1. Average Document Image Hue (F_h): A comparison between hue values in two document images provides an important clue to determining whether color qualities are identical. In fact, dominant wavelength is identical to the perceptual attribute, i.e., the hue. Thus, the proposed system considers a RGB image (of documents, e.g., academic certificates) as input. The hue for a single pixel (p), ^{h_p} is calculated from its RGB values ^{r_p}, ^{g_p} and ^{b_p} as follows [11, 19].

$$h_p = \{\theta \ \ if \ \ b_p \le g_p \ , 360-\theta \ \ if \ \ b_p > g_p \tag{1}$$

Where θ is the angle measured with below formula:

$$\theta = \cos^{-1}\left[\frac{((R-G)+(R-G))/2}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right] \tag{2}$$

For each pixel, h_p is obtained and then an average image hue is computed for the entire image. Let F_h indicate this average document image hue.

3.1.2. Gray Level Variation (f_{gv}): Gray level variation can be considered a feature used to calculate the standard deviation of a gray level distribution of pixel values in a document image. This variation is denoted $^{f_{gv}}$. For more details on how to compute these values, see [11, 20].

$$f_{gv} = \sqrt{\frac{\sum\left(g_p - \bar{g}\right)^2}{N-1}} \tag{3}$$

Where g_p the gray value of pixel is ^{p}, $\bar{g}$ is mean gray value of the document image and *N* total number of pixels. Note that convert the color document image into its corresponding a gray image according to the following formula: (i.e. Gray = 0.3 Red + 0.59 Green + 0.11 Blue).

3.1.3. Binary Correlation (f_{bc}): The calculation of binary correlation supposes previous knowledge regarding originality. A given document is examined (f_{kr},f_{kg},f_{kb}) and matched with the original document [11]. To compute the correlation between two binary document images, the Otsus thresholding method is used [21] to convert gray images into binary images. The correlation coefficient calculates the similarity between the scanned documents and the genuine documents, which are known as the reference documents (^{d_R}). The reference image is already in the QR code, and the other image is produced by scanning the document to verify the similarity between the QR code (reference) and the scanned document (target). The correlation coefficient, r between (^{d_R}) and the (^{d_T}) is measured as follows:

$$r(d_R, d_t) = \frac{1}{2} - \frac{s_{10}s_{01} - s_{00}s_{11}}{\sqrt{(s_{11}+s_{10})(s_{01}-s_{00})(s_{11}+s_{01})(s_{10}+s_{00})}} \tag{4}$$

Where s_{00}, s_{11}, s_{01} *and* s_{10} represents the number of: zero matches, one matches, zero mismatches, and one mismatches respectively. The value of this correlation coefficient gives the value for the feature f_{bc} .

3.1.4. Kurtosis of Image Colors (f_{kr}, f_{kg}, f_{kb}): Whereas gray level variation is used to detect whether variations are caused by rare deviations, the kurtosis is measured separately for the R, G, and B channels. Suppose that f_{kb} represents the convexity of the blue channel; the value is computed as described in [11, 22].

For instance, let $M \times N$ represent the kurtosis of blue channel and it is measured as

$$f_{kb} = \left\{ \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum \left[\frac{b_p - \bar{b}}{\sigma_r} \right]^4 \right\} - \frac{3(N-1)^2}{(N-2)(N-3)} \quad (5)$$

Where b_p is the blue pixel value p and $\bar{b}$ and σ_b are the mean and standard deviation of the blue pixel value, respectively. Similarly, f_{kg} and f_{kr} are subsequently calculated to represent the kurtoses of the green and red channels.

3.1.5. Histogram for Visual Color Features: A histogram provides a compact summary of the distribution of data in an image [23]. The histogram method is used for visual feature representation. The image is represented by its histogram. The histogram method is used for visual feature representation. The color histogram helps to find the images which contain similar color distribution. It is achieved by comparing histograms signatures of two document images and matching the color content of the first document image with the other and measuring the similarities through computing distance between two histograms. If the attempt to falsify the content of the document will show that there is a difference in data values between the two document images. To design a descriptor of an image, a histogram of the image is generated as follow.

$$\bar{h}_p = \sum_{i=1}^{M} \sum_{j=1}^{N} \delta_b(i,j), \forall_b = 0,1,2,..... \quad (6)$$

Where $M \times N$ are a 2D mapping for a digital image correspond to y-axis and x-axis respectively, $b(i, j) = 1$ if the v at pixel location $[i , j]$ falls in b, and $\delta_b(i,j) = 0$ otherwise. Similarities between different histograms h_p and $\bar{h}_p$ can be calculated using different methods such as Euclidean distance and histogram intersection as a similarity measure [18]. It is known that the Euclidean distance is one of the most effective methods for similarity measure. As a result, Euclidean distance was used in the proposed approach. The main drawback of histograms is the dependence on the color of the object being studied, which ignores the object's shape and texture.

3.2. Background Artwork

This section reviews two features that are used to examine alterations that may occur in background artwork as a result of forgery. This artwork mainly involves texture design. Furthermore, if fraud is carefully executed, it will be difficult to determine whether alterations have been made by examining a document using only the naked eye [24].

3.2.1. Texture Features (Gabor): Texture features represent the second major factor used in the proposed method and are important for document authentication purposes.

These features are extracted by utilizing the Gabor wavelet algorithm; see [25]. Generally, the Gabor filter captures energy, which represents information pertaining to an image at a particular scale and orientation [26]. The Gabor filter in the spatial domain is given by

$$g,\lambda,\theta,\psi,\sigma,\gamma(x,y)=\exp\left(\frac{x'^2+\gamma^2\lambda'^2}{2\sigma^2}\right)\cos\left[2\pi\frac{x'}{\lambda}+\psi\right] \tag{7}$$

Where $x'=x\cos\theta+y\sin\theta$ 젠nd 젠$y'=-x\sin\theta+y\cos\theta$

In this equation λ represents the wavelength of the cosine factor, θ represents the orientation of the normal to parallel stripes of a Gabor function in degrees, ψ is the phase offset in degree γ is the spatial aspect ratio and it specifies the elliptically of the support of the Gabor function and σ is the standard deviation of the Gaussian envelope that determines the linear size of the receptive field. When an image is processed by Gabor filter, the output is the convolution of the image $I(x,y)$ with Gabor function $g(x,y)$.

$$r(x,y)=I(x,y)*g(x,y) \tag{8}$$

Where * denotes the two dimensional convolution. The process can be performed at various orientations and scale. After applying Gabor filters on the image by orientation and scale, we are able to obtain an array of magnitudes [27].

$$E(m,n)=\sum\sum\left|G_{mn}(x,y)\right| 젠젠젠m= ,1,...,M-1;n=0,1,...,N-1 \tag{9}$$

The magnitudes symbolize the energy content at different scales and orientations of the image. The main purpose of texture feature is to find images or regions with a similar texture. Therefore, the mean μ_{mn} and standard deviation σ_{mn} of the magnitude of the transformed coefficients are used to represent the texture feature of the region [28].

$$\mu_{mn}=E(m,n)\Big/p\times q \tag{10}$$

$$\sigma_{mn}=\sqrt{\sum_x\sum_y(\left|G_{mn}(x,y)\right|-\mu_{mn}^2)\Big/p\times q} \tag{11}$$

Where M represents the scale and N represents the orientation. The mean μ_{mn} and standard deviation σ_{mn} are used to create a feature vector that represents texture features; specifically, texture features can be saved as two feature vectors. These two vectors are combined into a single feature vector that is considered an image texture feature [29].

3.2.2. Fourier Power Spectrum (f_{ps}): Fourier analysis of an image provides clues to detecting unexpected changes in the style of an artwork. The gray level of a document image is considered in examining the image's centered Fourier spectrum. This attribute is captured by the feature f_{ps}. For more details on the captured centered Fourier spectrum feature is given by

$$f_{ps}=\log(1+FS) \tag{12}$$

Where FS is the centered Fourier spectrum $FS = \sqrt{R^2 + l^2}$. R and l are the real and imaginary parts of the Fourier Transform [11].

3.3. Shape Features

3.3.1. Shape Features: Shape features are the third major component of the proposed approach, in which a Hu moment invariant algorithm is used to represent shape features. It is known that the Hu moment invariant algorithm is one of the most effective methods for extracting shape features. As a result, the moment invariant algorithm used in the proposed approach was carefully chosen. The moment invariant algorithm is invariant to location, orientation and size. This algorithm provides descriptions of shape features that are independent of location, size and orientation. This algorithm provides descriptions of shape features that are independent of location, orientation and size [18]. In order to calculate seven invariant moments .The 2D moment of order (p + q) of a digital image f (x; y) is defined as

$$m_{p,q} = \sum_x \sum_y x^p y^q f(x,y) \tag{13}$$

$$\mu_{p,q} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \tag{14}$$

Where $\bar{x} = m_{10} / m_{00}$ and $\bar{y} = m_{01} / m_{00}$, with $\bar{x}$ and $\bar{y}$ denoting the center of the region.

Hence the central moments of the third order can be computed as:

$$\left.\begin{aligned}
&\mu_{0,0} = m_{0,0} \\
&\mu_{1,0} = 0 \\
&\mu_{0,1} = 0 \\
&\mu_{1,1} = m_{1,1} - \bar{y} m_{1,0} \\
&\mu_{2,0} = m_{2,0} - \bar{x}^2 m_{1,0} \\
&\mu_{0,2} = m_{0,2} - \bar{y} m_{0,1} \\
&\mu_{3,0} = m_{3,0} - 3\bar{x} m_{2,0} + 2m_{1,0}\bar{x}^2 \\
&\mu_{2,1} = m_{2,1} - 2\bar{x} m_{1,1} - \bar{y} m_{2,0} + 2\bar{x}^2 m_{0,1} \\
&\mu_{1,2} = m_{1,2} - 2\bar{y} m_{1,1} - \bar{x} m_{0,2} + 2\bar{y}^2 m_{1,0} \\
&\mu_{0,3} = m_{0,3} - 3\bar{y} m_{0,2} + 2\bar{y}^2 m_{0,1}
\end{aligned}\right\} \tag{15}$$

The central moment can be normalized as:

$$\eta_{p,q} = \eta_{p,q} / \eta_{0,0}^{\gamma} \tag{16}$$

For $\gamma = [(p+q)/2] + 1$ where $p,\ q\ =\ 0,\ 1,\ 2,\ \ldots$, and $p + q = 2, 3 \ldots$, a set of seven 2D moment invariants can be derived from second and third central moments. The feature vectors can be calculated using these seven moments :$\phi 1 - \phi 7$. Moments $\phi 1 - \phi 6$ are

invariant to scaling, rotation and translation and moment $\phi 7$ is invariant to skewness. The moments $\phi 1 – \phi 7$ can be written as follows:

$$\left.\begin{aligned}
\phi_1 &= \mu_{20} + \mu_{02} \\
\phi_2 &= (\mu_{20} + \mu_{02})^2 + (4\mu_{11})^2 \\
\phi_3 &= (\mu_{30} + 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \\
\phi_4 &= (\mu_{30} + \mu_{12})^2 + (\mu_{21} - \mu_{03})^2 \\
\phi_5 &= (\mu_{30} + 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} - \mu_{03})^2] + (3\mu_{21} + \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} - \mu_{03})^2] \\
\phi_6 &= (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4\mu_{11}(3\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \\
\phi_7 &= (3\mu_{21} - \mu_{03})(\mu_{30} - \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] - (\mu_{30} - \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]
\end{aligned}\right\} \quad (17)$$

This set of normalized central moment is invariant features on changes in a document image translation, scaling and rotation [30-31].

4. System Design and Architecture

The QR barcode is used as a symbol in a system because it exhibits many sophisticated features, such as readability along any direction and high efficiency in storing valuable page content in a document. Automatic recognition, encoding, and decoding are the technologies used in 2D barcode applications and in barcode printing. The automatic recognition of barcodes is the most important among these technologies. Data integrity is an important component of detecting or preventing the forgery of any document, and hashing can be used to verify whether data tampering has occurred. In the hashing process, there are several sources for errors like unintentional modifications that might occur during the document's life cycle [32]. The main sources of error are printing and scanning operations. It was determined that the best method to adopt for this investigation was to match the hash values based on the similarity of their contents rather than exactly matching the data bits. This approach was helpful in detecting any change in content while tolerating the noise and distortion caused by printing, scanning and copying. To achieve this, we used a variant of perceptual hash function (PHF) similar to [33].The generation of the barcode and the encoding thereof into the document is illustrated in Figure 1. Table 1 illustrates the protocol for barcode formation.

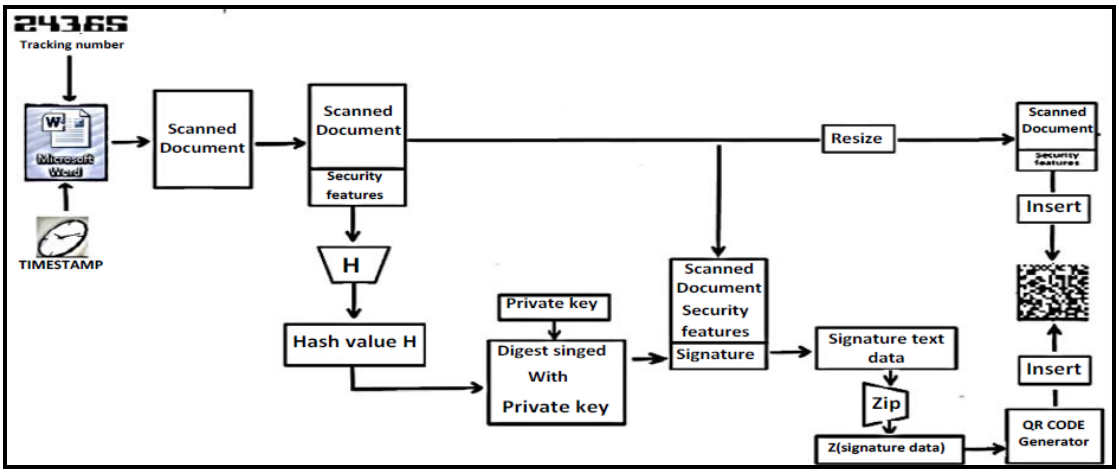


Figure 1. Barcode Generation Process

Integrating visual attributes, a timestamp, and a tracking number on each page may enhance security feature protection. The visual attributes of a scanned document image include the color histogram, the shape and the texture of the image. The time at which a secure document was created (timestamp) is embedded as the unique time value. Therefore, it helps to reduce any repetition of the document. The tracking number is used as a serial number of the unique information about the machine that produced the

Table 1. Barcode Generation Protocol

S	Description	Symbolic representation
1	Page content(M)	M
2	Append Time stamp(T) to M	T||M
3	Append Tracking Number (N) to M	N||T||M
4	Convert the document to a scanned image	JPG(N|| T || M)
5	Append extracted visual attributes of a scanned document image (SF)	SF||JPG(N|| T ||M)
6	Hash (H) the page content	H(SF|| JPG(N|| T || M))
7	Encrypt hash value by using the signer's private key to producing the signature	S(H(SF|| JPG(N|| T || M))
8	ZIP Signature, data	Z(S(H(SF|| JPG(N|| T || M)))
9	Encode (E) the page content	E(Z(SIGNATURE DATA))

Integrating visual attributes, a timestamp, and a tracking number on each page may enhance security feature protection. The visual attributes of a scanned document image include the color histogram, the shape and the texture of the image. The time at which a secure document was created (timestamp) is embedded as the unique time value. Therefore, it helps to reduce any repetition of the document. The tracking number is used as a serial number of the unique information about the machine that produced the document and thus deters any person from altering the original document. These attributes (intrinsic and extrinsic) are combined and hashed to produce the hash value of the document. The hash value *h*, which depends on the feature data for all of these visual attributes, is encrypted using the private key of the owner to produce a signature. Without encrypting the hash value, the timestamp and tracking number, anyone can change the date of the document to indicate that the document was signed on a different date. Therefore, the timestamp and tracking number are significant in restricting document date alteration .Table 1 illustrates the protocol for barcode formation. This framework provides a significant and crucial feature that may be used to help in lawsuits and insurance claims. This case reveals the need for further investigation in identifying by whom, when and where a document was created.

The QR code barcode acts as a carrier of encrypted security feature values and as a miniature version of a document. Thus, the proposed method does not need any of the following for further security: a built-in security chip for secure data storage, superior paper for printing, plating, a hologram or sophisticated printing device for integrity protection. Normal laser printers can print documents that contain this barcode. Figure 2 illustrates the printed document decoding and verification processes. To verify the integrity of the document, it is scanned and the security features are extracted from the scanned document. The timestamp, the tracking number and the visual attributes of the scanned document image are extracted. Thereafter, the data are hashed and matched with the hash value stored in the 2D barcode to verify the originality of the document. If there are any differences between the values and the differences in the values are somewhat greater than the threshold values, then the document can be suspected to be a forged copy. This conclusion can be efficiently arrived at using a fuzzy logic controller. As mentioned

previously, this system certainly offers an inexpensive and fast method for the verification of documents such as forensic reports, court orders, memos or circulation and confidential documents.

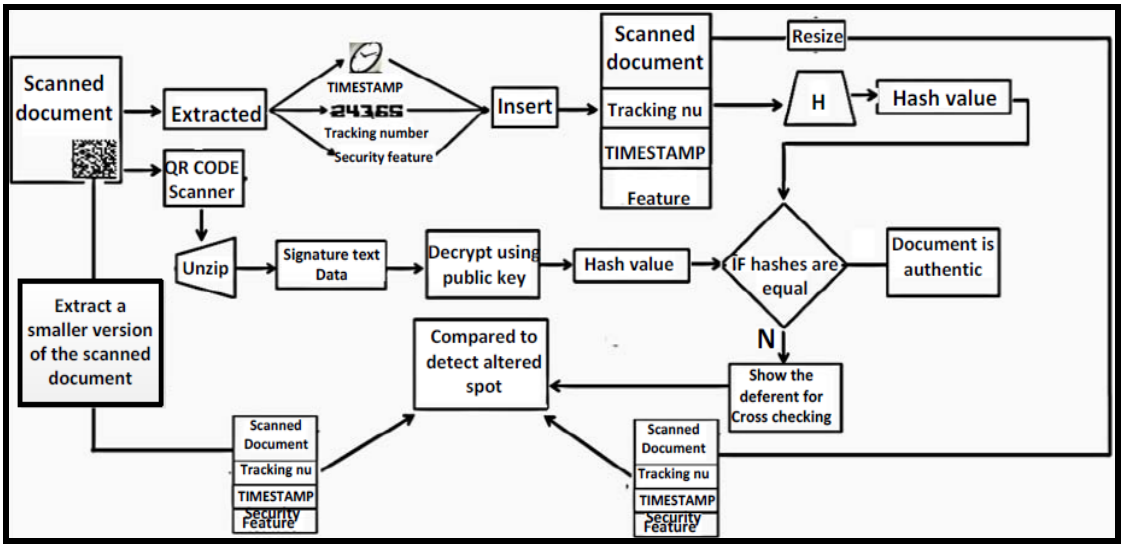


Figure 2. Barcode Decoding and Verification Process

Each functional unit is responsible for the development of specific actions and is separated from the rest of the system. If it is necessity to upgrade, this type of system architecture allows for an easy upgrade process because the system can simply connect the modules and begin. Any new technology or improvements to the current method may be implemented without much trouble.

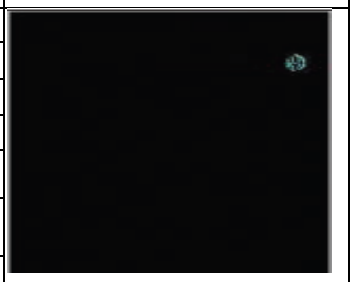
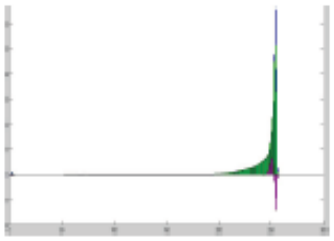
There are five major functional components: 1) the document image hash function, 2) digital signature generation and verification, 3) QR code encoding and decoding, 4) document compression and 5) extraction of visual attributes and security features. Furthermore, the size of the generated 2D barcode is dependent on the amount of the data in the document. Therefore, to have an efficient barcode size, the signature text data and a smaller version of the document are compressed before the barcode is generated.

5. System Analysis and Testing

Full automation and integration testing are performed to demonstrate the effectiveness of the proposed system and to illustrate that the service could be used for forensic science without the need for any additional equipment or human intervention to identify any printed document forgery. The system should be dependable overall and in each integrity test.

To demonstrate the potential of this approach and its suitability as a document authentication system, five main tests were conducted in this study. These modification detection tests used multilingual text content as well as image content, including a photograph, an image logo, a signature or seal, a texture or background content, the shape of the document, and the colors of the content in a document. To simulate the falsification process, the information of the authentic document was modified.

Table 2. The Difference between Forged Documents from its Original Counterparts

Features		Genuine	Forged	Result
Color	Average Hue	0.2612	0.2612	
	Gray level variation	43.82	43.69	
	Binary correlation	0.940354213920108		
	Kurtosis of colors	11.86	14.70	
	Histogram for visual color	0.0680	0.0561	
Texture	Gabor features	0.8210	0.7982	
	Fourier Power Spectrum	232.8	232.5	
Shape	Seven Invariant Moments	0.000756	0.000758	
		6.1371	6.1585	
		2.88	2.84	
		4.30	4.21	
		1.267	1.238	
		8.0682	7.8987	
		4.6323	4.4406	

This modification can be performed by a number of methods, such as adding or deleting different symbols in the text content, altering the security features, producing the document by using another device, cropping the hand signature or seal, or, finally, by changing the logo, photo, timestamp or tracking number. As shown in Table 2, demonstrates the ability of the system to track changes in a document, even if the change is in the smallest unit of the document content, such as a change in the smallest number in the document content .

Data stored in a barcode that is appended at the end of an original document is divided into two sections. The first section contains the encrypted hash values and the encrypted visual feature values that are extracted from the document during construction. The second section contains a smaller version of the original document. The main challenge facing the system is the ability to ensure the integrity of the document and to conduct results analysis automatically without human intervention. The proposed system verifies the integrity of a document in two phases. The first phase decrypts the original hash value using the digital signature and the public key. Furthermore, a new hash value is generated by checking document, and the value is compared with that of the original document using a similarity measure. If the two values are very similar, the expected results will be obtained from the document. In the case of a mismatch, the proposed method shifts to the second phase. The analysis automatically provides detailed results pertaining to amendments that give rise to a mismatch between the two values based on the visual attribute similarity measure and the alignment information regarding the genuine document.

Table 3. All Cases of Modification

1-,text,,sim:0.998021	DIS: 0.0020
2- LOGO,,sim:0.996847	DIS:0.0032
3-SEAL,,sim:0.997892	DIS:0.0021
4-SIGNATURE,,sim:0.998108	DIS:0.0019
5-TEXTURE,,sim:0.999581	DIS:4.1900e-004
6-shape,,sim:0.999116	DIS:8.8400e-004
7-color,,sim:0.744099	DIS:0.2559
8-authentic doc ,sim:1.000000	DIS:0.0000

This paper addresses the following questions: What is modified in the content of a document, and how does an investigator specify the special location of the modified document? It was observed that the answers to these questions can be obtained by comparing the thumbnail of a document stored in a barcode with a thumbnail image of the scanned document. Table 3 demonstrates the ability of the system to track changes in a document, all cases of modification of the document content. There is a significant difference between the reference document and forged document. Testing demonstrated that the proposed model is suitable for detecting all cases of modification. In the system, results suggesting fraudulent behavior are expected. The system allows legal and judicial authorities to identify what data have been modified as well as the spatial location in a document of that data, with consideration of all types of data contained in the document.

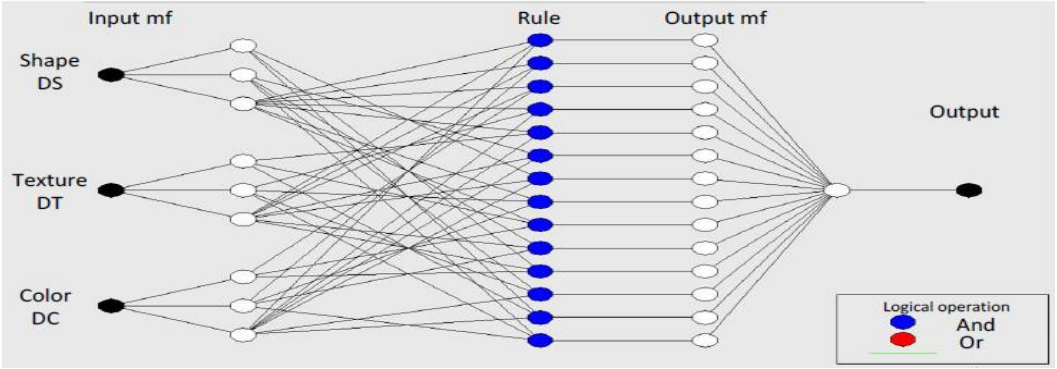


Figure 3. Fuzzy Similarity Measures

5.1. Visual Attributes Similarity Measure

The present work compares the visual attribute vectors between genuine and forged documents that reflect the similarity of the documents. To measure the similarity, three types of visual features are used, namely color, texture and shape. These features can be used to determine the overall document similarity, which can be achieved by calculating the Euclidean distances of each feature. These distances are defined as the shape distance D_S , the color distance D_C and the texture distance D_T . The distances are extracted from a document by using moment invariants, a histogram of the scanned document image (red, green, blue and intensity) and the Gabor wavelet. The distance between the corresponding elements of the two feature vectors is defined as follows:

$$ED(X^K, Y^t) = \sqrt{\sum_{i=1}^{n} (X_i^K, Y_i^t)^2} \tag{17}$$

X^K and Y^t are the genuine and forged document, respectively; i is a feature range. A small distance represents a high similarity between the scanned documents. Suppose the vector from the visual feature of the genuine document is G and the visual feature vector from the forged document is F . The shape distance $D_{SHAPE}(S_G, S_F)$, the color distance $D_{COLOR}(C_G, C_F)$ and the texture distance $D_{TEXTURE}(T_G, T_F)$ between the scanned document G and F are three inputs of the fuzzy similarity measure, as shown in Figure 3.

After obtaining the visual feature vector, fuzzy heuristics are used to measure similarity, taking into account the priority of each attribute in finding the percentage of similarity for each vector D_{SHAPE} , D_{COLOR} and $D_{TEXTURE}$. The common denominators between the two documents are measured using the following criteria. D_{SHAPE} , D_{COLOR} and $D_{TEXTURE}$ are used to calculate the distance of the visual attributes between the genuine and the forged documents. S is the value of the visual attribute similarity measure. To process the three Euclidean distances, a set of fuzzy rules is implemented as follows:

Step 1: Determine the number of feature vector inputs. In this case, three entries are used, for instance color, texture and the shape distance between the visual attributes in the genuine and the forged documents.

Step 2: For these three inputs, the membership functions are assigned, i.e., great similarity, medium similarity and weak similarity, as three different types of fuzzy sets.

Step 3: Declare output fuzzy sets value ranges such as great similarity between [0 , 0.90], medium similarity between [0.91 , 0.96] and weak similarity between [0.97 , 1].

Step 4: Create the fuzzy rules and apply logical operators such as IF and THEN conditional statements.

Step 5: Fuzzify the crisp inputs by the Sugeno-style fuzzy inference method to regulate the most suitable fuzzy sets for each input.

Step 6: Utilize an AND operator to obtain the output as a single value.

Step 7: The output of an aggregation process is one fuzzy set for each output variable, which is obtained by the unification of all of the output rules of the last step.

Step 8: Finally, via defuzzification, the output fuzzy set that is gathered must be transmuted into a single crisp number.

The priority for each attribute is used to determine the degree of similarity when documents are compared, with a small distance reflecting a close match. As a result, the first priority is specified to be shape features because the shape of a document is not easily affected by external factors; the shape remains unchanged when a specified transformation is applied, such as rotation, translation or orientation. However, the second priority is specified to be the color features because these features are only invariant to

rotation and translation. The third priority is specified to be texture features. The priority for each attribute is shown in Figure 4.

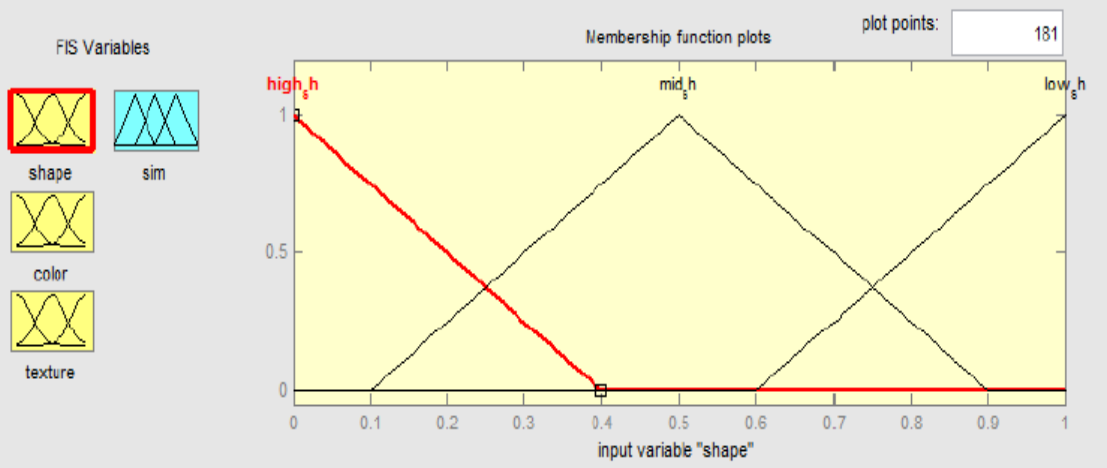


Figure 4. Fuzzy Set that Identified each Input as Low, Medium and High

5.2. Experimental Results

5.2. 1. Building Data Set: As explained in section 1, academic certificates were used as a case study in this study to evaluate the results of the experiments. An academic certificate was designed with the utmost precision, virtually replicating the original design with 100% accuracy. All security protection standards and the same sophisticated devices that were used to protect and obtain the real certificate were used in this study. Concurrently, the equipment utilized to create a forged document was exactly the same as that used to create the authentic document, and basic features were exactly matched. For this study, an expert on forged documents in the forensic sciences was contacted to determine the details of the forgery process. All types of modern counterfeit devices, for instance high-end scanners, high-resolution offset printers, recent image editing application programs and even experts, are involved in producing forged academic certificates. Furthermore, the exact same brand and quality of paper used in printing genuine academic certificates were used in this study. In total, Table 4 shows 3200 sets of document samples were produced by utilizing different printers and scanners to introduce different noise levels as a result of the printing and scanning processes.

Table . Capturing Samples of Documents by Utilizing Different Printers and Scanners

DATA	Captured from	Scanner1		Scanner2		Scanner3		Scaner4	
		Genuine	Forged	Genuine	Forged	Genuine	Forged	Genuine	Forged
Set1	Printer1	100	100	100	100	100	100	100	100
Set2	Printer2	100	100	100	100	100	100	100	100
Set3	Printer3	100	100	100	100	100	100	100	100
Set4	Printer4	100	100	100	100	100	100	100	100

5.2. 2. Classification Result: Our method is compared with two other existing methods, namely: trained forensic document examiners and employees of ratification. We select randomly fifty documents for each set for the experiments. Table 5 shows the test results. According to the test results forensic experts shows high accuracy even though it takes relatively high processing time to finish the work as it use optoelectronic devices for the process. On the other hand employees in the ratification use less time for the process but achieve lower accuracy as it depend upon the employee's personal experience. But our proposed system able to achieve 100% accuracy with less time as it does not use any extra equipment as the trained forensic document examiners.

Table 4. The Performance Comparison with Forensic Experts and Employees of Ratifications

Testing method	Samples	Test result	Accuracy %	Time(min)
Forensic expert	Set1(F:50,G:50)	Set1(F:50,G:50)	100	240
	Set2(F:50,G:50)	Set2(F:50,G:50)	100	135
	Set3(F:50,G:50)	Set3(F:50,G:50)	100	76
	Set4(F:50,G:50)	Set4(F:50,G:50)	100	110
Employees of ratification	Set1(F:50,G:50)	Set1(F:42,G:58)	84	40
	Set2(F:50,G:50)	Set2(F:48,G:52)	96	42
	Set3(F:50,G:50)	Set3(F:53,G:47)	94	45
	Set4(F:50,G:50)	Set4(F:46,G:54)	92	48
Our system	Set1(F:50,G:50)	Set1(F:50,G:50)	100	3
	Set2(F:50,G:50)	Set2(F:50,G:50)	100	3
	Set3(F:50,G:50)	Set3(F:50,G:50)	100	3
	Set4(F:50,G:50)	Set4(F:50,G:50)	100	3

6. Evaluation and Results

The performance of the system was tested on individual feature categories and a combination of all the feature categories. This was done partly to justify our priority assumption and partly to select the optimal performance of the system. The results are shown in Table 6.

Table 5 . Classification on Individual and Combined Features

Classify Using	Data	Accuracy (%)	Error
Color Features	SET1	92	8
	SET2	91	9
	SET3	93	7
	SET4	89	11
Average		**91.25**	**8.75**
Texture Features	SET1	90	10
	SET2	85	15
	SET3	92	8
	SET4	87	13
Average		**88.50**	**11.50**
Shape Features	SET1	93	7
	SET2	95	5
	SET3	97	3
	SET4	96	4
Average		**95.25**	**4.75**
Combination of all features	SET1	100	0
	SET2	100	0
	SET3	100	0
	SET4	99.7	0.3
Average		**99.92**	**0.08**

From Table 6 it can be observed from the table that, among the individual feature types, shape features ranked highest, followed by color features and then texture features, which is consistent with our priority ranking. Besides, the best performance was realized with the combination of all the features categories, thus justifying our prioritization assumption and our deployment of fuzzy heuristics for computation of the final similarity value. Figure 5 is a graphical representation showing the classification accuracy of individual feature categories and a combination of all features.

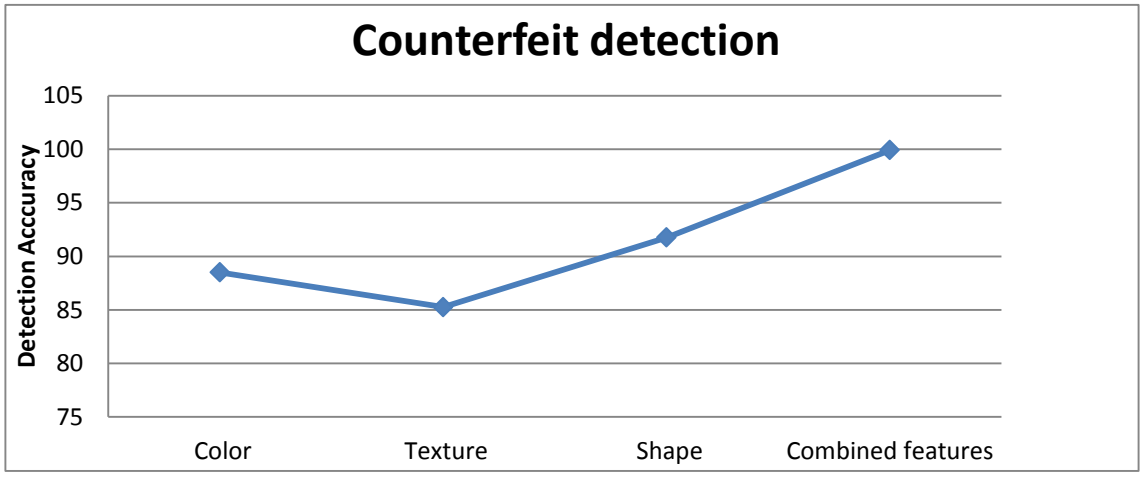


Figure 5. Classification Capability of Individual Features and Combination of All Features

7. Conclusions

Forging important official documents is a high- risk criminal activity that adversely affects society. The key strengths of this study were to design an efficient, low-cost solution to this problem. Accurately detecting forged documents and ensuring the full-scale deployment of such detection systems could be useful. This study provides an automatic authentication method for verifying the integrity of important, valuable documents that is based on visual features such as color, texture, shape. These features are directly related to the perceptual aspects of the document content, controlled by Euclidean measures and fuzzy heuristics. The main objective of this study was to extract distinct visual features and embed them with some distinct information in the printed document to prevent the document from being forged. By adding timestamps and tracking numbers with combined visual features, a document signature is constructed. The document signature is a distinct vector that ensures the proposed printed document produces results that are substantially related to the document's contents. In the proposed system, a QR barcode is used to store the valuable contents of the document and the corresponding hash value. Academic certification documents were used in experiments and for verification. The proposed approach ensures that the document is authentic, that it comes from a verified source, and that the contents have not been tampered with because it has been digitally signed. The desired forensic results require a more detailed investigation regarding the effects of fraud in generating documents.

Acknowledgements

This research is supported by Natural Science Foundation of China (61173122), Key Project of Natural Science Foundation of Hunan Province, China (12JJ2038), Natural Science Foundation of Hunan Province, China (09JJ6102).

References

[1] R. L. Van Renesse, "Paper based document security-a review", **(1997).**
[2] G. Gupta, S. Saha, S. Chakraborty and C. Mazumdar, "Document frauds: Identification and linking fake document to scanners and printers", Computing: Theory and Applications, 2007. ICCTA'07. International Conference on IEEE, **(2007**), pp. 497-501**.**
[3] M. Warasart and P. Kuacharoen, "Paper-based Document Authentication using Digital Signature and QR Code", p. 40.
[4] S. Juan, "4th International Conference on Computer Engineering and Technology", International Proceedings of Computer Science and Information Technology**, (2012**), pp. 94-98**.**

[5] M. Afrakhteh, S. Ibrahim and M. Salleh, "Printed Document Authentication Using Watermarking Technique", Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on IEEE, **(2010**), pp. 367-370**.**
[6] E. J. Delp III and P. W. Wong, "Security, Steganography, and Watermarking of Multimedia Contents VI", 5306 Security, Steganography, and Watermarking of Multimedia Contents VI **(2004**)**.**
[7] I. Amidror, "New print-based security strategy for the protection of valuable documents and products using moiré intensity profiles", Electronic Imaging 2002 International Society for Optics and Photonics, **(2002**), pp. 89-100**.**
[8] A. K. Mikkilineni, G. N. Ali, P.-J. Chiang, G. T. Chiu, J. P. Allebach and E. J. Delp, "Signature-embedding in printed documents for security and forensic applications", Electronic Imaging 2004 International Society for Optics and Photonics, **(2004**), pp. 455-466**.**
[9] S. Wang, S. Huang, X. Zhang and W. Wu, "Hologram-based watermarking capable of surviving print-scan process", Applied optics, vol. 49, no. 7, **(2010).**
[10] J. Li, X. Zhang, S. Liu and X. Ren, "Adaptive watermarking scheme using a gray-level computer generated hologram", Applied optics, vol. 48, no. 26, **(2009).**
[11] G. Garg, P. K. Sharma and S. Chaudhury, "Image based document authentication using DCT", Pattern Recognition Letters, vol. 22, no. 6, **(2001).**
[12] U. Garain and B. Halder, "On automatic authenticity verification of printed security documents", Computer Vision, Graphics & Image Processing, 2008. ICVGIP'08. Sixth Indian Conference on IEEE, **(2008**), pp. 706-713**.**
[13] S. Ibrahim, M. Afrakhteh and M. Salleh, "Adaptive watermarking for printed document authentication", Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on IEEE, **(2010**), pp. 611-614**.**
[14] M. H. Eldefrawy, K. Alghathbar and M. K. Khan, "Hardcopy Document Authentication Based on Public Key Encryption and 2D Barcodes", Biometrics and Security Technologies (ISBAST), 2012 International Symposium on IEEE, **(2012**), pp. 77-81**.**
[15] J. Gebhardt, M. Goldstein, F. Shafait and A. Dengel, "Document authentication using printing technique features and unsupervised anomaly detection", Document Analysis and Recognition (ICDAR), 2013 12th International Conference on IEEE, **(2013**), pp. 479-483**.**
[16] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T. Chiu, J. P. Allebach and E. J. Delp III, Printer identification based on graylevel co-occurrence features for security and forensic applications", Electronic Imaging 2005 International Society for Optics and Photonics, **(2005**), pp. 430-440**.**
[17] M. P. Khatkale and D. Lokhande, "Digital Watermarking Algorithm for Color Images", IOSR Journal of Engineering, vol. 3, **(2013).**
[18] C.-M. Pun and C.-F. Wong, "Fast and robust color feature extraction for content-based image etrieval", International Journal of Advancements in Computing Technology, vol. 3, no. 6, **(2011).**
[19] K. Iqbal, M. O. Odetayo and A. James, "Content-based image retrieval approach for biometric security using colour, texture and shape features controlled by fuzzy heuristics", Journal of Computer and System Sciences, vol. 78, no. 4, **(2012).**
[20] K. Yoshinari, Y. Hoshi and A. Taguchi, "Color image enhancement in hsi color space without gamut problem", Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on IEEE, **(2014**), pp**.** 578-581**.**
[21] C. Wu and X. Tai, "Application of gray level variation statistic in gastroscopic image retrieval", Computer and Information Science, 2009. ICIS 2009. Eighth IEEE/ACIS International Conference on IEEE, **(2009**), pp. 342-346**.**
[22] N. Otsu, "A threshold selection method from gray-level histograms", Automatica, vol. 11, **(1975),** pp**.** 285-296**.**
[23] B. Halder and U. Garain, "Color feature based approach for determining ink age in printed documents", Pattern Recognition (ICPR), 2010 20th International Conference on IEEE, **(2010**), pp. 3212-3215**.**
[24] Y. Zhang and L. Wu, "Classification of fruits using computer vision and a multiclass support vector machine", Sensors, vol. 12, no. 9, **(2012).**
[25] U. Garain and B. Halder, "Machine authentication of security documents", Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on IEEE, **(2009**), pp. 718-722**.**
[26] Z.-C. Huang, P. P. Chan, W. W. Ng and D. S. Yeung, "Content-based image retrieval using color moment and Gabor texture feature", 2 Machine Learning and Cybernetics (ICMLC), 2010 International Conference on IEEE, **(2010**), pp. 719-724**.**
[27] J. Han and K.-K. Ma, "Rotation-invariant and scale-invariant Gabor features for texture image retrieval", Image and vision computing, vol. 25, no. 9, **(2007).**
[28] K. Choudhary, M. Pundlik and D. Choukse, "An integrated approach for image retrieval based on content", IJCSI International Journal of Computer Science Issues, vol. 7, no. 3, **(2010).**
[29] A. Khaparde, B. Deekshatulu, M. Madhavilatha, Z. Farheen and S. Kumari, "Content based image retrieval using independent component analysis", IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 4, **(2008).**
[30] K. Iqbal, "Image Detection and Retrieval for Biometric Security from an Image Enhancement Perspective", in, Coventry University, (**2011**).

[31] Z. Huang and J. Leng, "Analysis of Hu's moment invariants on image scaling and rotation", 7 Computer Engineering and Technology (ICCET), 2010 2nd International Conference on IEEE, V7-476-V477-480; **(2010).**
[32] M. Yang, K. Kpalma and J. Ronsin, "A survey of shape feature extraction techniques", Pattern recognition, **(2008).**
[33] S. Voloshynovskiy, O. Koval, R. Villan, E. Topak, J. E. V. Forcén, F. Deguillaume, Y. Rytsar and T. Pun, Information-theoretic analysis of electronic and printed document authentication", Electronic Imaging 2006 International Society for Optics and Photonics, 60721D-60721D-60720; **(2006).**
[34] F. Ahmed, M. Y. Siyal and V. U. Abbas, "A secure and robust hash-based scheme for image authentication", Signal Processing, vol. 90, no. 5, **(2010).**

Authors

Mohammed AL-Gawda, he had a B.Sc. Honours in computer science and Information Technology, from Al-Neelain University College in Khartoum 2004. He had M.Sc. in computer science from Al-Neelain University in Sudan 2010. He served as Head of Development Department in YemenSoft Company from 2006 to 2010, He served as Head of the Computer Science department in the College of Computer Science in Queen Arwa University from 2010 to 2011. He is currently pursuing a Ph.D. in computer science at the School of Information Science and Engineering, Central South University, Changsha, China. His research interest includes Image processing, Printed Document Authentication Using Image Processing Techniques.

Beiji Zou, 1978 – 1982 Zhejiang University, Hangzhou, China BSc Computer Software.1982 – 1984 Qinghua University, Beijing, China. MSc Computer Application. 1997 – 2001 Hunan University, Changsha, China, Ph.D. Control Theory and Engineering Research Experience. 2002 – 2003 Qinghua University Post Doctor. 2003 – 2004 Griffith University, Australia. Visiting Scholar. Vice Dean of School of Information Science and Engineering, Central South University, China. His research interest includes Digital image processing, Computer Graphics, Multimedia Technology and Software Engineering.

Nurudeen Mohammed, he received his BSc. in Mathematical Science from the University for Development Studies , Ghana in 2008, had his Masters of Engineering in Computer and Technology in 2013 form Central South University and currently pursuing his PhD at the school of Information Science and Engineering, Central South University, China. His major research interest includes Digital Image Processing, Wireless Sensor Networks, Surveillance Engineering and Biometrics. Email: nurudeen_saeed@yahoo.com(N. Mohammed).