Document Verification: Analysis of Overlapping Images & QR/Barcode Code

Neha Sharma, Avinash Markad   
*Ajeenkya DY Patil*   
*School of engineering*  
*Savitribai Phule Pune*   
*University*  
*Pune,India*

*Abstract*- This work presents two novel algorithms: one for recovering overlapping signatures in hyperspectral imaging (HSI) of documents and another for improving data storage in Quick Response (QR) codes using encrypted lossless compression technology. The first algorithm addresses the challenge of signature extraction in forensic document analysis when signatures overlap with typed text, stamps, or images. By utilizing pure pixel index approaches and spectral classification, the method effectively separates overlapping elements and validates the extracted signatures using the structural similarity index.

The second algorithm enhances QR code data storage by incorporating a combination of Huffman compression and XOR-based encryption. This technique overcomes the limitation of traditional QR codes’ size without altering the existing QR code structure, providing a secure method for encoding larger messages. Experimental results demonstrate the effectiveness and superiority of both methods, highlighting their potential applications in document authentication and secure data transmission.

*Keywords: High-Spectrum Imaging, Huffman Coding, Spectral Analysis, Security,* *quick response code (QR code), Huffman coding, security.*

*Introduction*- This paper presents two major contributions: a hyperspectral imaging (HSI) algorithm for document signature extraction and an improved methodology for secure data storage in Quick Response (QR) codes. HSI is a powerful imaging technology using a broad range of the electromagnetic spectrum to capture detailed information, making it valuable in various fields such as astronomy and chemical analysis. Inspired by the way certain animals utilize non-visible spectra, HSI captures spectral signatures that help identify overlapping elements, like signatures on documents with typed text or stamps. This work introduces a pure pixel index method for overlapping text segmentation and employs the Structural Similarity Index Measure (SSIM) to accurately validate the quality of the extracted signatures.

The second part of this research addresses the limitations of QR codes, a widely-used format for secure data transmission and document authentication. QR codes are structured in 2D matrices that store data using visible patterns, but their limited size can hinder the storage of large information. This paper proposes an enhanced QR code methodology incorporating lossless compression and encryption techniques, improving both data capacity and security without altering the existing QR structure. The new approach enables the inclusion of confidential information and secure message sharing, thereby addressing common QR code limitations. The results showcase the potential applications of these methods in document authentication and secure data transmission.

FIGURE 1. Illustrated the arrangement of the information in a 2D matrix visible as black and white.

*Literature Survey* –

A. RELATED WORK

Various methodologies have been proposed to address document authentication and verification in scenarios such as college admissions, scholarship applications, and employee recruitment, where the use of fake or altered documents is a prevalent issue. Traditionally, document verification has been performed manually, which is time-consuming, prone to human error, and less reliable for detecting subtle alterations. As a result, automated approaches leveraging image processing and machine learning have gained significant attention.

Several studies have explored hyperspectral imaging (HSI) as a means of detecting fraudulent modifications in documents. HSI, which captures spectral data across a wide range of wavelengths, allows for the analysis of fine details and subtle changes in document features that may be invisible to the human eye. Researchers have utilized spectral unmixing and pure pixel index methods to differentiate between overlapping text and graphical elements, achieving high precision in extracting hidden or overlaid information. Validation techniques such as the Structural Similarity Index Measure (SSIM) have been employed to evaluate the fidelity of the extracted data, highlighting the effectiveness of HSI in document forensics.

In addition to HSI, other studies have focused on using Quick Response (QR) codes for document integrity and secure data storage. QR codes, widely used for storing and transmitting information, face challenges due to their limited capacity and vulnerability to tampering. Past research has looked into optimizing the storage capacity of QR codes through error correction mechanisms and compression algorithms. Some methods propose encryption techniques to ensure data confidentiality, making it difficult for unauthorized users to manipulate or alter the encoded information.

A notable advancement is the integration of lossless compression and encryption schemes, which allow for greater data storage within a QR code without altering its visible structure. This approach not only enhances the amount of data that can be securely stored but also provides robust protection against forgery and unauthorized access. Such techniques have been successfully applied in scenarios requiring secure document sharing, digital signatures, and authentication systems.

Overall, prior research highlights the potential of HSI and QR code technologies in document authentication and security. However, there remains a need for more comprehensive systems that can combine these technologies to address the problem of document forgery in real-time applications, such as during document submission in colleges, scholarship portals, and employment verification processes. This paper builds on these foundations by proposing an improved HSI algorithm for extracting authentic document signatures and a novel QR code-based method for secure data storage and transmission.

B. SPECTRAL UNMIXING OF HSI DOCUMENT IMAGES

Spectral unmixing is a key process in hyperspectral imaging (HSI) that allows for the separation and identification of multiple components present within a single pixel by analyzing its unique spectral signature. In the context of document authentication, spectral unmixing helps to differentiate between various overlapping elements in a document, such as handwritten signatures, typed text, stamps, or seals. The complexity of document forgery scenarios, like modifying names, dates, or other sensitive information, often results in a mixture of spectral signals, making it challenging to identify authentic content. Therefore, spectral unmixing is essential for isolating genuine features and detecting alterations that might be invisible in standard imaging techniques.

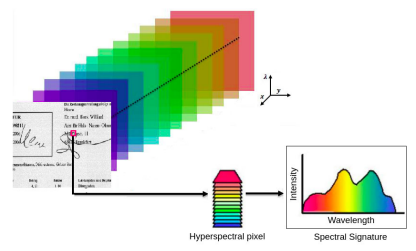
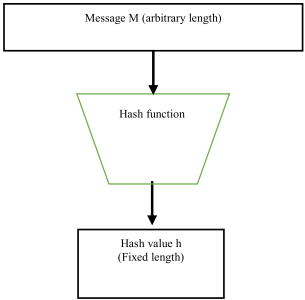


FIGURE 2. A hyperspectral camera captures images consisting of a large number of channels. Each pixel has its own spectral signature.

In the proposed system, spectral unmixing is applied to HSI document images to segment and extract overlapping content, thereby addressing the problem of fraudulent document submissions in college admissions, scholarship applications, and employee recruitment. By using pure pixel index approaches, the system identifies the presence of unique spectral properties within each hyper-pixel and separates the mixed components based on their spectral characteristics. Each material, whether it is the ink of a signature or the printed text, has a distinct spectral signature that HSI can capture across hundreds of wavelengths. The process involves three main steps:

1. Spectral Decomposition: Each pixel is analyzed across its spectral bands to determine the unique combination of materials or inks present in it. This is particularly useful when a signature is overwritten by typed text or another graphical element, as it allows the system to distinguish between the spectral properties of the different inks used.

2. Spectral Classification: After decomposition, the system groups the spectral components into different classes, representing unique elements like original signatures, stamps, or edited content. This classification is critical for pinpointing the authentic parts of the document and identifying potential alterations.

3. Validation with Structural Similarity Index Measure (SSIM): To ensure that the extracted elements closely match the authentic features of the document, the Structural Similarity Index Measure (SSIM) is used to validate the fidelity of the segmented data. SSIM compares the structural similarity of the recovered signature with a reference signature, highlighting even minor discrepancies that could indicate forgery.

Through these steps, the proposed method leverages HSI’s capability to capture detailed spectral data, enabling the detection of complex alterations and the recovery of authentic signatures with high precision. This makes it particularly effective in scenarios where document tampering involves overlapping modifications, thereby ensuring a robust and reliable document verification system.

C. COMPRESSION USED IN THE PROPOSED SYSTEM

The proposed system integrates advanced data compression techniques to enhance secure document storage and transmission. Given the constraints associated with limited storage capacity and the need for efficient and reliable verification, lossless compression algorithms are employed to maximize data utilization without compromising integrity. The system specifically utilizes “Huffman Coding”, a well-established lossless compression method, alongside secure hashing functions to ensure both storage efficiency and data authenticity.

FIGURE 2. Show the integrity of data using hash function.

1. Huffman Coding for Data Compression

Huffman coding is a fundamental lossless data compression technique that reduces the size of input data by assigning variable-length codes to different characters based on their frequency of occurrence. In the context of the proposed system, Huffman coding optimizes the storage of document information by encoding frequently used characters with shorter binary codes and less frequent characters with longer ones. This results in significant reduction of the overall data size, making it ideal for embedding large amounts of information into QR codes without altering their structure.

The compression process begins by building a Huffman tree, where each character is represented as a leaf node, and the tree is constructed based on the frequency of characters in the input data. After constructing the tree, the system traverses it to assign variable-length codes to each character. This approach is particularly effective when dealing with repetitive patterns in text data, such as names or document identifiers, which often appear multiple times in admission, scholarship, and hiring documents.

The efficiency of Huffman coding can be expressed with a time complexity of \(O(n \log n)\), making it a suitable choice for real-time document processing scenarios. Once compressed, the data is integrated into the QR code structure, allowing secure and compact storage of sensitive information, such as digital signatures or encrypted document details.

2. Data Integrity Using Hash Functions

To ensure that the integrity of the compressed data is preserved, the system incorporates a robust hash function that generates a unique hash value for each document. The hash function takes an arbitrary length input (e.g., compressed document data) and outputs a fixed-length hash value, which acts as a digital fingerprint for the original document. Any modification to the document, whether accidental or malicious, would result in a completely different hash value, thereby alerting the system to potential tampering.

The proposed system utilizes hash functions adhering to strict avalanche criteria (SAC) and bit independence criteria (BIC) to minimize any correlation between input changes and the resulting hash value. This ensures that even a minor alteration in the document data results in a significantly different hash output, making it easier to detect forgery or tampering. By combining Huffman coding with hash-based integrity checks, the proposed system achieves a high degree of security, ensuring that all stored or transmitted documents are both efficiently compressed and reliably authenticated.

Overall, the integration of Huffman coding and hash functions provides a comprehensive solution for secure and compact data storage in QR codes, enabling efficient document verification across various applications such as college admissions, scholarship submissions, and employee recruitment. This dual-layered approach ensures that sensitive information is not only compressed but also safeguarded against unauthorized alterations, making the system robust and reliable for real-world deployment.

D. SPECTRAL SIMILARITY METHODS

Spectral similarity methods compare the similarity between spectral signatures. The two most commonly used algorithms for spectral similarity are Spectral Angle Mapping (SAM) and Spectral Information Divergence (SID).

1. SPECTRAL ANGLE MAPPING In Spectral Angle Mapping a pixel spectrum is represented as a vector with dimension equal to the number of spectral channels. A pixel vector has two components: its magnitude (length) and its angle. SAM compares two pixels by measuring the angle between its two corresponding pixel vectors so it is insensitive to errors caused by illumination effects.

2. SPECTRAL INFORMATION DIVERGENCE Spectral Information Divergence (SID) considers a pixel as a random variable and defines a probability distribution for it according to its spectrum. The similarity between two pixels is computed as a relative entropy of their random variables.

E. OVERVIEW OF QR CODE STRUCTURE AND BENEFITS

QR codes are two-dimensional barcodes introduced in 1990 that encode more data than traditional 1D barcodes. Developed by Denso Wave in 1994, QR codes are widely used for various applications, including confidentiality and copyright protection. A typical QR code structure consists of five main components:

1. Finder Pattern: Square patterns on three corners that help identify the position, size, and angle of the QR code.

2. Alignment Pattern: Used to correct distortions and maintain the QR code's integrity.

3. Timing Pattern: A sequence of alternating black and white cells along the horizontal and vertical axes, ensuring stability and accurate symbol delivery.

4. Quiet Zone: A blank region surrounding the QR code, essential for proper readability.

5. Data Area: Encodes the actual information using binary values (0s and 1s) based on Reed-Solomon codes.

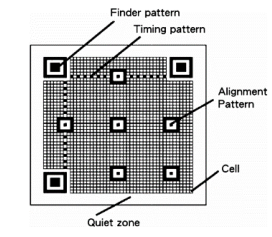
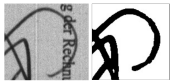


FIGURE 4. Show the main five parts of QR code structure

BENEFITS AND DATA CAPACITY

The QR code has 40 versions, each defined by a different module configuration, ranging from 21×21 modules (Version 1) to 177×177 modules (Version 40). The data capacity varies depending on the type of data: up to 7,089 characters for numeric, 4,296 for alphanumeric, or 2,953 for binary formats, making it superior to traditional barcodes, which can store only up to 20 digits. QR codes are versatile and can be read easily using both physical scanners and mobile devices.

### F. SIGNATURE EXTRACTION FROM HSI DOCUMENT IMAGES

The proposed methodology for extracting overlapping signatures in HSI images involves four primary steps:

**PRE-PROCESSING:**  
The dataset consists of 59 HSI document images, each with 240 channels and a resolution of 800x640 pixels. After visual inspection, channels 0-15 and 200-240 were discarded as they did not contain relevant signature information. A slight blur was applied to the remaining 184 channels using a low-pass filter to reduce noise and improve signature extraction.

**SPECTRAL-UNMIXING:**  
three endmember extraction algorithms—flipped, n-finder, and atop—were employed to identify unique spectral signatures. The resulting endmembers were grouped into classes representing typed text, overlapped text and signature, standalone signatures, and stamps or seals.

**HSI SPECTRAL IMAGE REGENERATION:**  
Signatures were extracted by comparing document pixels to the endmembers using Spectral Angle Mapper (SAM) and Spectral Information Divergence (SID) algorithms. Pixels with high similarity to the endmembers were classified as part of the signature. A manually tuned threshold determined which pixels were included in the final output.

**POST-PROCESSING:**  
The extracted signatures often contained noise, which was removed using a connected-component analysis. This method grouped pixels based on proximity and intensity similarity, resulting in a cleaner extracted signature.

FIGURE 5. Magnified view of a blurred image showing overlapped text and signature, and the recovered signature

Overall, the proposed approach effectively isolates overlapping signatures in complex HSI document images using a combination of spectral unmixing and image regeneration techniques.

### G. PROPOSED METHOD FOR SECURE QR CODE DATA STORAGE

The proposed system improves the security and storage capacity of QR codes through a multi-level approach using Huffman coding and XOR encryption. The method includes:

**DATA CONVERSION & COMPRESSION:** Convert the data (text or image) into a binary format. Transform the binary data into hexadecimal representation. Compress the resulting data using Huffman coding to create a smaller, encoded binary file.

**ENCRYPTION PROCESS:** Apply XOR encryption using a symmetric key of the same length as the text, creating a stream cipher where each plaintext bit is encrypted with a corresponding key bit. Reverse the compressed binary file and encrypt again using XOR with the secret key to generate a secure encrypted file.

**KEY GENERATION:** A dynamic and secure key is generated to balance encryption strength and computational efficiency. The execution time for key generation was optimized for various file sizes.

**SECURE QR CODE STORAGE:** The encrypted file is then encoded into a QR code, enhancing its confidentiality and storage capacity. This approach addresses limitations of previous QR methods by combining lossless compression and advanced encryption techniques, significantly reducing computational complexity and increasing QR code storage capacity.

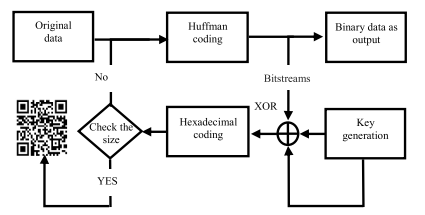


FIGURE 5. Demonstrated the proposed model for data compression.

### **Conclusion: -** The proposed system addresses the issue of document authenticity and security in various scenarios, such as college admissions, scholarship applications, and employee recruitment, where fraudulent document submission is a prevalent problem. By leveraging advanced techniques in hyperspectral imaging (HSI) and secure Quick Response (QR) code methodologies, the system ensures robust verification and protection against document forgery.

For document authentication, the HSI-based method utilizes endmember extraction algorithms combined with hyperspectral similarity measures to extract overlapping signatures accurately, even from complex documents. By incorporating pre-processing techniques like blurring and validating the results using the Structural Similarity Index Measure (SSIM), the system significantly improves the precision of signature recovery.

To secure data transmission and storage, the proposed method enhances QR code capacity and confidentiality using lossless Huffman compression and XOR encryption. This novel approach allows for the integration of large amounts of secure data within the QR code, overcoming the limitations of conventional QR code storage while resisting known security threats.

Overall, the proposed framework provides a comprehensive solution for document verification and secure data storage, making it suitable for real-time applications in education, employment, and other sectors that require rigorous document validation and data protection.

***References:-***

1] R. Mandal, P. P. Roy, and U. Pal, ‘‘Signature segmentation from machine printed documents using conditional random field,’’ in Proc. Int. Conf. Document Anal. Recognit., Sep. 2011, pp. 1170–1174.

[2] R. Mandal, P. P. Roy, U. Pal, and M. Blumenstein, ‘‘Signature segmentation and recognition from scanned documents,’’ in Proc. 13th Int. Conf. Intellient Syst. Design Appl., Dec. 2013, pp. 80–85.

[3] R. Zhu, F. Zhou, and J.-H. Xue, ‘‘MvSSIM: A quality assessment index for hyperspectral images,’’ Neurocomputing, vol. 272, pp. 250–257, Jan. 2018.

[4] V. K. Gunjan and J. M. Zurada, Eds., Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough. Cham, Switzerland: Springer, 2021.

[5] P. Barlas, S. Adam, C. Chatelain, and T. Paquet, ‘‘A typed and handwritten text block segmentation system for heterogeneous and complex documents,’’ in Proc. 11th IAPR Int. Workshop Document Anal. Syst., Apr. 2014, pp. 46–50.

[6] S. Liu, Z. Fu, and B. Yu, ‘‘A two-level QR code scheme based on polynomial secret sharing,’’ Multimed Tools Appl, vol. 78, no. 15, pp. 21291–21308, Aug. 2019.

[7] X. Zhang, H. Luo, J. Peng, J. Fan, and L. Chen, ‘‘Fast QR code detection,’’ in Proc. Int. Conf. Frontiers Adv. Data Sci. (FADS), Oct. 2017, pp. 151–154.

[8] R. Kr and B. M. Sagar, ‘‘Quick response code for fast detection and recognition of information,’’ Int. J. Emerg. Technol. Adv. Eng., vol. 4, no. 2, pp. 271–275, 2014.

[9] L. Belussi and N. Hirata, ‘‘Fast QR code detection in arbitrarily acquired images,’’ in Proc. 24th SIBGRAPI Conf. Graph., Patterns Images, Aug. 2011, pp. 281–288.