Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30618

Batch Differential Desensitization and Visualization of Medical Images Based on OCR and FF3/SHA-256

Yuning Wang

School of Management and Economics, Tianjin University, Tianjin, China
18704168455@163.com

Abstract. In the medical field, data security is paramount. Medical images typically contain a wealth of patient personal information, such as names, ID numbers, and contact information, as well as sensitive medical data, such as diagnostic results. This data faces the risk of leakage during storage, transmission, and sharing. Once privacy information is leaked, it not only infringes on patient privacy but can also lead to serious legal and social problems. This paper primarily uses literature analysis, comparative analysis, and experimental methods to implement an automated batch classified desensitization and decryption system for medical images based on Optical Character Recognition (OCR) technology, a reversible encryption algorithm (FPE), and a hash function (SHA-256). This system applies different encryption techniques based on the intended use of the identified information, achieving partial reversibility so that data can be recovered for specific medical procedures under authorized scenarios. This effectively protects patient privacy, while automated batch processing improves the efficiency of medical staff and facilitates data sharing among researchers.

Keywords: Medical Imaging, Privacy Protection, Optical Character Recognition, Encryption Algorithm

1. Introduction

Optical Character Recognition (OCR) is a technology that converts text in an image into processable digital text. It involves a series of preprocessing steps on the image, followed by localization and feature extraction, and finally outputting the recognition result in text format. Format Preserving Encryption (FPE) is a general term for format-preserving encryption. When sensitive information is encrypted, the format and length of the ciphertext are completely identical to the original information. The reversible encryption algorithm used in this paper is FF3, a specific algorithm in the FPE field. It uses the AES-128 symmetric encryption algorithm as the core encryption module while adding format preservation functionality. This paper also uses the SHA-256 hash function to process some sensitive information. This hash function has one-way irreversibility, meaning that the original data cannot be deduced from the hash value.

In recent years, there have been many research achievements in the field of medical image desensitization. Some studies have tried to combine OCR with encryption technology. However, the

DOI: 10.54254/2755-2721/2026.TJ30618

recognition accuracy of traditional OCR tools is sometimes insufficient, and most of them use a uniform encryption algorithm, which lacks differentiated processing for sensitive information.

Based on this, this paper adopts a variety of research methods. First, literature analysis is used to analyze current OCR technology and mainstream encryption algorithms. Then, comparative analysis is used to compare the advantages and disadvantages of each technology and algorithm and make a selection. Finally, technology integration and experimental methods are used to build a complete desensitization system using Python 3.8 as the development language, and specific experiments are conducted to verify the effect.

This paper specifically studies three directions, possessing certain unique features. First, OCR recognition optimization: PaddleOCR is used to achieve high-precision localization and extraction of information from medical images, improving the accuracy and completeness of text recognition. Second, differentiated desensitization of sensitive information: The extracted information is divided into two categories. The SHA-256 algorithm is used for "core irreversible sensitive information" to ensure absolute security, while the FF3 algorithm is used for "non-core reversible information" to facilitate information decryption and restoration in authorized scenarios. Third, input and output optimization: It supports processing common image formats and can perform batch processing, ultimately outputting all images in common formats to a specified directory.

This research is applicable to the medical industry. Optimization of OCR effectively addresses the issue of low text resolution in medical images; the classification of sensitive information takes into account the characteristics of information in medical images, and the FPE algorithm avoids occlusion of the encrypted image, ensuring image integrity and facilitating subsequent use of medical images by medical personnel; the optimization of input and output meets the format compatibility requirements of the medical industry and takes into account the massive amount of information in the field, significantly improving efficiency. Furthermore, this research has social significance, meeting current high demands for privacy protection, supporting compliant operations of relevant medical institutions, and improving the usability of medical data.

2. Extraction of sensitive information from medical images

Text recognition in images is the first goal of this project, and its effect will directly affect the subsequent encryption effect. OCR can convert the text in the image into computer-editable text. However, one of the major challenges facing OCR technology is the complexity and diversity of application scenarios. OCR has a wide range of applications, including education, finance, transportation and medical care. However, the data styles in different scenarios are very different, and the user needs are also different [1]. Therefore, it is necessary to find the OCR technology that is most suitable for medical scenarios to improve the accuracy of recognition.

The main task during the development phase of the sensitive information recognition module was to conduct multiple rounds of testing on different OCR technologies to select the optimal solution that could meet the high-precision recognition requirements of text in medical images. Initially, the research team tested the EasyOCR third-party library, which showed low accuracy in recognizing text in medical images, with numerous instances of misidentification or missed recognition. Subsequently, the team attempted to build a custom image recognition model based on the PyTorch deep learning framework and trained the model using collected medical image samples. However, during testing, two core problems were discovered: first, the model training was inefficient and slow, failing to meet the project development cycle requirements; second, the model convergence was poor, with consistently high loss values during training, making it unsuitable for the high-precision text recognition requirements of medical scenarios.

DOI: 10.54254/2755-2721/2026.TJ30618

To address the aforementioned technical bottlenecks, the research team conducted comparative tests on mainstream open-source image recognition libraries such as Chineseocr, Chineseocr_lite, Tesserocr, and Paddle OCR. As shown in Table 1, Paddle OCR performed best in terms of ease of use and accuracy [2].

Table 1. Comparison of mainstream open-source OCR model frameworks [2]

Open Source OCR Project	F1-Score	Pre-trained Model Size	PIP Installation	Custom Training	Native Edge Deployment
PaddleOCR - v2	0.5224	3.5M	Supported	Supported	Supported
Chineseocr	/	/	Not supported	Supported	Not supported
Chineseocr_lite	0.3899	4.7M	Not supported	Not supported	Supported
Tesseract	/	/	Not supported	Supported	Supported

In addition, considering that this project is mainly applied to the medical industry, Paddle OCR has the following unique advantages: First, the tool library is open source and free, which reduces the cost in the medical industry. At the same time, it only requires selecting the required model and filling in a few parameters to achieve powerful OCR effects. The operation threshold is low and it is easy for medical staff who are not computer professionals to use.

Second, Paddle OCR achieved high accuracy in the experiment by using a pre-trained medical text-specific model, as shown in Figure 1, which can accurately extract key information such as names, ID numbers, and medical terms.

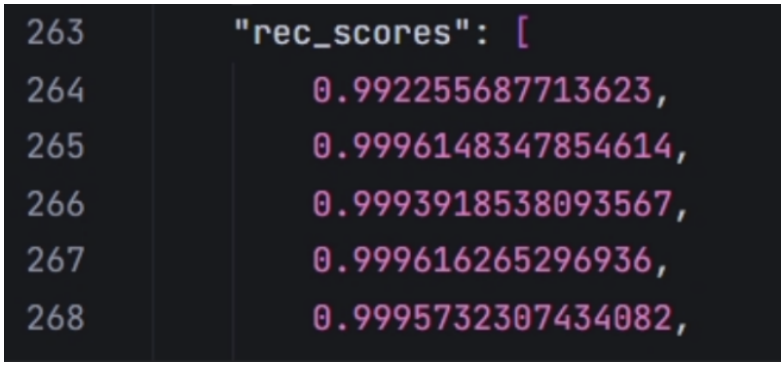


Figure 1. Recognition confidence score of Paddle OCR

Third, the tool can simultaneously output the coordinate position of the text block in the image, as shown in Figure 2, locating the positions of the four vertices in the text block, providing a precise spatial positioning reference for subsequent sensitive information encryption and desensitized image generation.

DOI: 10.54254/2755-2721/2026.TJ30618

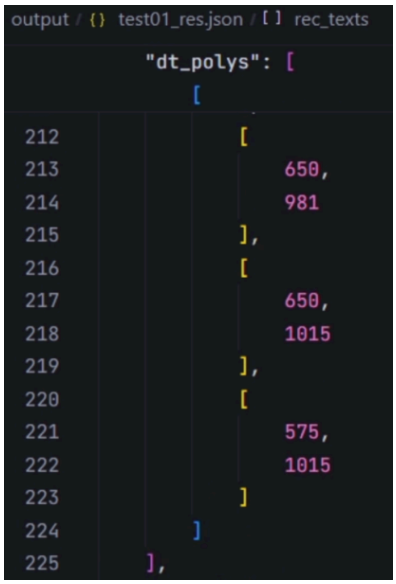


Figure 2. Four-corner localization of the text block recognized by Paddle OCR

In summary, Paddle OCR was ultimately chosen as the text recognition tool for this study, and as shown in Figure 3, it successfully located and extracted sensitive information from medical images.

Figure 3. Example of medical image text detection and recognition results based on Paddle OCR

3. Differentiated encryption and decryption of sensitive information

3.1. Basis for classified encryption of sensitive information

At present, there is a basic technical framework of "identification-encryption" for desensitizing medical images, but most studies still adopt a "one-size-fits-all" encryption mode, ignoring the privacy preferences of information owners for different information, resulting in excessive encryption sacrificing usability or insufficient encryption causing privacy leaks [3].

DOI: 10.54254/2755-2721/2026.TJ30618

Based on research on user privacy preferences, we can formulate different privacy protection levels and adopt different encryption methods to achieve classified desensitization. Studies have shown that patients are willing to share some medical application information to a certain extent in exchange for more convenient medical services or health management when personal information is not leaked [3]. Among personal information, patients are most sensitive to their ID card number, and most people believe that the consequences of leaking the ID card number are the most serious. At the same time, in certain authorized specific scenarios, some encrypted sensitive information needs to be restored so that medical personnel can study the condition or contact the patient, but the ID card number is not helpful for this.

Based on this, an irreversible encryption algorithm is used for ID card numbers in sensitive information to ensure the security of highly sensitive privacy information; for other sensitive information such as patient names, contact information and pathology information, a reversible encryption algorithm is used so that the data can be recovered for specific medical operations in authorized scenarios.

3.2. Reversible encryption of information such as names

3.2.1. Comparison and selection of reversible encryption algorithms

Reversible encryption algorithms refer to the technology that can be reversed to recover the plaintext after encryption using the corresponding key and algorithm. According to the "key usage method", they can be divided into two main categories: symmetric encryption algorithms and asymmetric encryption algorithms.

The most widely used symmetric encryption algorithm is AES, while FPE is a secondary category under symmetric encryption algorithms. A key characteristic of FPE technology is that the ciphertext of information encrypted with FPE schemes shares the same characteristic format as the plaintext. Therefore, when the FPE scheme is used in fields where the plaintext to be encrypted has professional requirements, it has the advantages of reducing the maintenance cost of the encrypted database and improving the security of data analysis [4]. At the same time, for medical scenarios, FPE has more obvious advantages. It avoids the situation where the encrypted ciphertext is too long and occludes the medical image, ensuring the integrity and availability of the image. FF3 is the mainstream FPE algorithm recommended by NIST. It uses AES-128 as the core encryption module and inherits the high security of AES.

The most widely used asymmetric encryption algorithm is RSA, whose security is based on the "large integer factorization problem".

Based on this, this study used FF3 and RSA for comparative analysis, and the results are shown in Table 2:

Table 2. Multidimensional comparison of FF3 and RSA

Comparison Dimensions	FF3 (Based on AES-128)	RSA
Ciphertext Format	Same length as plaintext	Random binary scrambling
Computational Efficiency	High	Low(large integer factorization)
Medical System Compatibility	High	Low
Key Management	Simple (single key, no negotiation required	Complex (public key distribution, private key secrecy)

DOI: 10.54254/2755-2721/2026.TJ30618

Through the comparison of the above dimensions, it can be seen that FF3 is better than RSA in terms of scene adaptability. Therefore, this project chooses the FF3 algorithm in the reversible encryption module.

3.2.2. Implementation of the FF3 reversible encryption algorithm

In the FF3 encryption process, the plaintext is first encrypted using a specified adjustment value to break the mapping relationship between plaintext and ciphertext. Then, after a series of length tests, the information extracted by Paddle OCR is split into two parts, which are then encrypted using the following core code and finally merged into a ciphertext:

```
# 8 rounds of Feistel encryption
  for i in range(NUM_ROUNDS):
    if i % 2 == 0:
      m = u
      W = Tr
    else:
      m = v
      W = Tl
    P = calculate_p(i, self.alphabet, W, B) # Calculate wheel input P
    revP = reverse_string(P)
    S = self.aesCipher.encrypt(bytes(revP)) # AES encryption
    S = reverse_string(S)
    y = int.from_bytes(S, byteorder='big') # Convert to integer
    c = decode_int_r(A, self.alphabet) + y # Calculate the intermediate value c
    c = c % modU if i % 2 == 0 else c % modV
    C = encode_int_r(c, self.alphabet, int(m)) # Encode as a string
    A, B = B, C # Swap A and B
  return A + B
```

3.3. Irreversible desensitization of valid ID information

3.3.1. Identification of valid ID information

Valid ID numbers need to be distinguished from other sensitive information for irreversible hashing. To ensure accurate differentiation, a standardized verification process is required to filter valid ID numbers. This study designs a multi-dimensional verification mechanism based on the People's Republic of China ID number coding standard, and implements automated verification of valid ID numbers through programming. The core verification logic is as follows:

First, a format compliance check is performed, namely, the ID number must be 18 digits long, the first 17 digits must be numbers and the last digit must be a number or X [5].

After format verification passes, the checksum logic is executed. This verification is based on the weighted summation algorithm specified in the national standard. First, a 17-digit array of corresponding weighting factors is defined [7,9,10,5,8,4,2,1,6,3,7,9,10,5,8,4,2], and a checksum

DOI: 10.54254/2755-2721/2026.TJ30618

mapping table ['1','0','X','9','8','7','6','5','4','3','2']. Then calculate the sum of the products of the first 17 digits and their corresponding weighting factors, and take the remainder of this sum divided by 11. The theoretical checksum is then obtained from the mapping table based on the remainder. The calculated theoretical checksum is compared with the 18th digit of the extracted text. If they match, the ID number is considered valid; otherwise, a prompt message is returned.

Finally, irreversible desensitization will only be performed on the information once it is verified as a valid ID number.

3.3.2. Choice of irreversible desensitization

Previous research has shown that the main characteristics of hash functions are:

a. One-way irreversibility. Information M is hashed to obtain a fixed-length digest D, but M cannot be derived from D.

b. No repetition. Two different pieces of information, M1 and M2, yield different digests, D1 and D2, and D1 and D2 must be different.

c. Simple computation and excellent encryption performance [6].

The most widely used hash algorithms are the MD series and SHA series algorithms. Among them, the MD5 algorithm and SHA-256 algorithm have high security, wide application and are the most representative. SHA-256 is very sensitive to the initial value and can enhance the randomness of encryption [7]. Therefore, this study uses the SHA-256 hash function to realize irreversible desensitization of valid ID card numbers.

3.4. Decryption process of FF3 algorithm

In authorized scenarios, encrypted sensitive information can be decrypted to recover data for specific medical research or communication. The decryption module herein targets information processed by the FF3 algorithm can restore it to plaintext; however, it is ineffective for ID card numbers processed by SHA-256.

The decryption process is similar in principle to the encryption process, but the order of rounds is reversed and the operation is subtraction. A summary of the key code is as follows:

```
# Reverse 8-round Feistel decryption
  for i in reversed(range(NUM_ROUNDS)):
    if i % 2 == 0:
      m = u
      W = Tr
    else:
      m = v
      W = Tl
    P = calculate_p(i, self.alphabet, W, A)
    revP = reverse_string(P)
    S = self.aesCipher.encrypt(bytes(revP))
    S = reverse_string(S)
    y = int.from_bytes(S, byteorder='big')
```

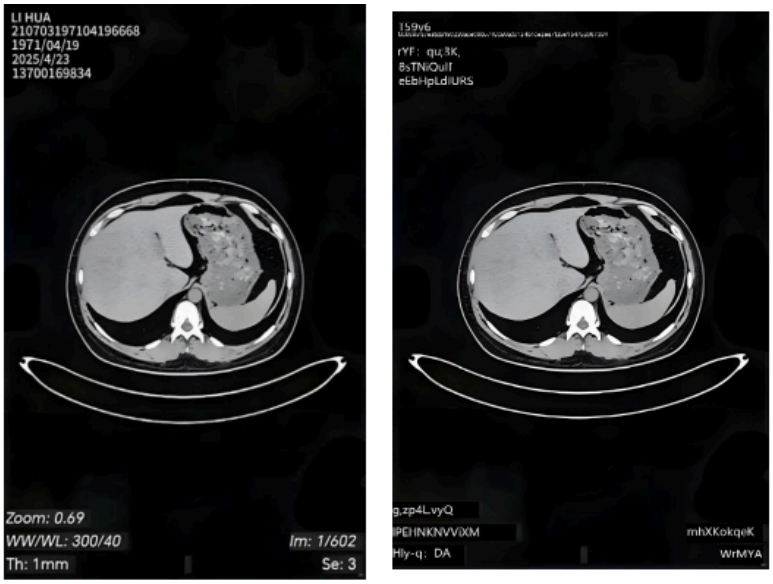
DOI: 10.54254/2755-2721/2026.TJ30618

```
    c = decode_int_r(B, self.alphabet) - y # Use subtraction instead of
addition.
    c = c % modU if i % 2 == 0 else c % modV
    C = encode_int_r(c, self.alphabet, int(m))
    B, A = A, C # Swap B and A
  return A + B
```

4. Demonstration of encryption/decryption effects and batch processing of medical images

The core feature of the encryption/decryption process in this project is visualization. It directly performs pixel-level overlay processing on the regions corresponding to sensitive information in the original medical images, drawing the ciphertext or decrypted plaintext directly onto the original images and generating complete output files without the need to split the images or store the text separately.

During the encryption phase, the coordinates of sensitive areas located by Paddle OCR are used to overlay the original text in a "black background with white text" mode. The ciphertext of the ID number after SHA-256 hashing and the ciphertext of other sensitive information after FF3 encryption are accurately drawn in their original positions. The font size is dynamically adapted according to the size of the text box to ensure that the ciphertext is displayed completely and does not exceed the original area. The effect is shown in Figure 4.

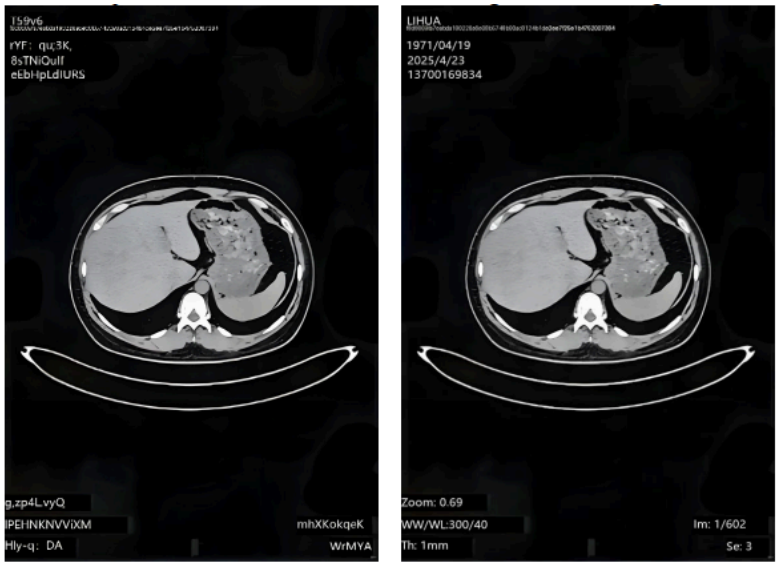


（a）Original medical images　　（b）Desensitized medical images

Figure 4. Effect of medical image encryption

During the decryption phase, for reversible sensitive information encrypted with FF3, the plaintext is restored using the key and Tweak value, and then the ciphertext is replaced with the same overwrite method. The image file containing the restored plaintext is directly output. The ID number remains in ciphertext state due to irreversible hashing, as shown in Figure 5.

DOI: 10.54254/2755-2721/2026.TJ30618



（a）Encrypted medical images　　（b）Decrypted medical images

Figure 5. Medical image decryption effect diagram

Meanwhile, it allows users to customize directories and generate complete paths by traversing all image files in the directory, thereby batch reading images and improving efficiency [8]. After processing by the above modules, the desensitized and restored medical images are batch stored in common formats, such as JPG to the specified output folder, as shown in Figure 6, and the system automatically checks whether the format of the stored files is correct and whether the file content is complete, ensuring that the desensitized medical images can be saved and used normally.

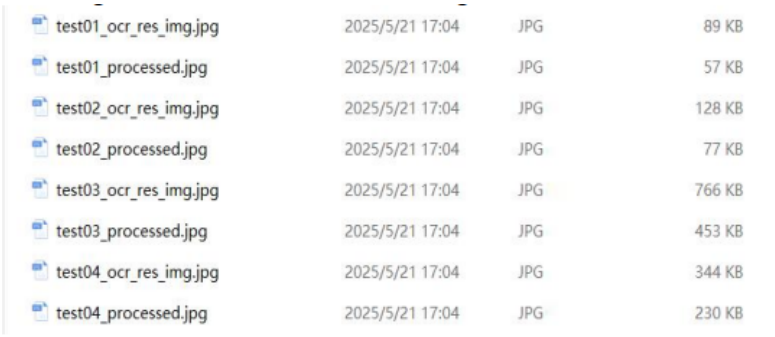


Figure 6. Effect of batch output of medical images

5. Conclusion

This paper addresses the need for desensitizing sensitive information in medical images, exploring a batch processing scheme based on OCR recognition, FF3 reversible encryption, and SHA-256 irreversible hashing. The scheme achieves accurate desensitization, restoration, and visualization of sensitive information. Research shows that this scheme can automatically locate and distinguish sensitive information in images, ensure data security through differentiated encryption strategies, and support directly outputting both ciphertext and decrypted plaintext by overwriting the original image without changing the file format. It is fully compatible with the requirements of medical systems such as HIS and PACS, and both batch processing efficiency and desensitization security meet clinical application standards.

DOI: 10.54254/2755-2721/2026.TJ30618

There is still room for improvement in this study. For example, the robustness of the currently used OCR technology needs to be enhanced, and it does not support the desensitization of sensitive information in multiple languages, and lacks a simple and intuitive visual operation interface.

Future research can focus on three aspects: first, combining deep learning to optimize OCR algorithms and improve recognition accuracy in complex scenarios; second, developing desensitization modules that are compatible with various languages and realizing dynamic desensitization processing during image transmission; and third, designing corresponding user interfaces to further reduce the usage threshold for non-technical personnel and enhance the practicality and universality of the project.

References

[1] Li Hongliang, Liu Yuliang, Liao Wenhui, et al. Optical character recognition in the era of large-scale models: current status and prospects [J].Journal of Image and Graphics, 2025, 30(06): 2023-2050.

[2] Zheng Hongkun. Application and Practice of PaddleOCR-based Image Recognition Technology in Epidemic Prevention and Control [J].Digital Technology & Application, 2022, 40(12): 117-119.DOI: 10.19695/j.cnki.cn12-1369.2022.12.37.

[3] Zhu Hongcan, Chen Jili, Zheng Kaidi, et al. A Study on Users' Privacy Preferences for Digital Medical and Health Information from the Perspective of Sensitivity—Based on Discrete Choice Experiments [J].Scientific Information Research, 2025, 7(04): 1-12.DOI: 10.19809/j.cnki.kjqbyj.2025.04.001.

[4] Zhang Xuewei. Research and Application of Identity-Based Format-Preserving Encryption Schemes [D].Northwest Normal University, 2020.DOI: 10.27410/d.cnki.gxbfu.2020.001893.

[5] Gao Lixia. Application and Reflections on ID Card Number Verification in Auditing [J].Internal Auditing in China, 2017, (06): 72-76.

[6] Li Ni. Research and FPGA Implementation of MD5 and SHA-256 Algorithms [D].Hunan University, 2021.DOI: 10.27135/d.cnki.ghudu.2021.001027.

[7] Yang Yuguang, Wang Jiawei. A Quantum Logistic Mapping Image Encryption Algorithm Based on SHA-256 and Arnold Mapping [J].Journal of Anhui University(Natural Sciences), 2024, 48(01): 35-42.

[8] Zhang Jingchen, Wang Jiayang, Zhao Yuanzhi, et al. A Custom Medical Image De-identification System Based on Data Privacy [J].Frontiers of Data and Computing, 2025, 7(03): 122-135.