

SecureVault: Intelligent File Sharing and OCR System

Mayank Dhodi

Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India
mayank.dhodi22@spit.ac.in

Sahil Gole

Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India
sahil.gole22@spit.ac.in

Uday Hiray

Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India
uday.hiray22@spit.ac.in

Anuharsh Pawar

Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India
anuharsh.pawar22@spit.ac.in

Sanjuktarani Jena

Department of Computer Engineering
Sardar Patel Institute of Technology
Mumbai, Maharashtra, India
sanjuktarani.jena@spit.ac.in

Abstract—With the rise of digital documents and online file-sharing platforms, keeping uploaded files secure and private has become a major challenge. Traditional file-sharing systems often offer limited control over how documents are accessed, viewed, or shared after distribution. This can lead to risks, including unauthorized access and the misuse of confidential information. This paper proposes a SecureVault, an intelligent file-sharing and OCR system, to create a safe and smart platform for uploading, managing, and sharing PDF documents and image files. The system includes secure authentication, permission-based access control, and Optical Character Recognition (OCR) techniques for automatic text extraction from uploaded documents. It is built using React.js for the frontend, Node.js and Express.js for backend services, and MongoDB for database management. Tesseract.js and PDF-Parser libraries are incorporated for OCR and text extraction functions. JWT authentication and encrypted password storage enhance security and privacy protection. Experimental results show that this system effectively blends secure file sharing with OCR-based document processing for modern digital document management.

Index Terms—Secure File Sharing, OCR, Tesseract.js, Access Control, Document Security, React.js, Node.js, MongoDB, PDF Processing, Authentication

I. INTRODUCTION

The fast development of digital communication technologies and cloud-based platforms has caused a dramatic increase in the prevalence of online document sharing systems across educational institutions, organizations, healthcare systems, government sectors, and enterprise applications. Users commonly share sensitive documents such as PDF files, scanned copies of ID cards, invoices, certificates, legal documents, and image-based records via digital systems for storage, communication, and collaboration. Due to the convenience and accessibility of existing cloud storage and document sharing systems, users still adopt these systems, though they have a multitude of security and privacy issues. These challenges for users include unauthorized access, uncontrolled downloading and sharing of documents, and

misuse of sensitive documents.

Once documents are shared with other users through conventional file sharing platforms, the systems provide little control to the original document owner regarding the shared documents. Owners of documents usually cannot control how documents are viewed, downloaded, copied, or shared after they are made accessible to others. This causes serious privacy problems, especially with confidential and sensitive documents. These problems are even worse with documents that include personal identification, financial, and health records, as well as data pertaining to corporate and business operations. Consequently, there is a clear demand for document sharing systems that are both secure and intelligent and that permit access control and document management with privacy concerns addressed.

Beyond the security issues, the rising popularity of digital files and scanned images has led to a need for advanced automated document processing systems that can extract and control text from uploaded files. Optical Character Recognition (OCR) technology is one of the many systems that provide the ability to convert scanned documents and text images into editable text. With OCR-based document processing systems, intelligent text extraction from documents, automated indexing, metadata creation, and automated retrieval of information from the uploaded documents are all possible.

We present SecureVault: Intelligent File Sharing and OCR System. SecureVault is an online platform for the management and sharing of sensitive documents that incorporates advanced security protocols. The system combines several new security tools into one seamless design. These new tools significantly enhance the security of the files. The user is given control over the permissions at the document level. The greater control over permissions means that the user is able to permit

access to the document as a secure view. The user can choose not to permit the document to be downloaded. This control means that the user can ensure that the document will not be distributed beyond the secure view. In addition the user can control the access to the document so that it is not permitted to be viewed at all.

The proposed application uses React.js for the frontend, Node.js and Express.js for the backend, and MongoDB for the database. Tesseract.js implements the OCR functionality, and PDF-Parser works to extract text from scanned images and PDF documents. We secured user sessions through the use of JWT for authentication and bcrypt for password hashing.

The proposed system primarily contributes the following:

- A file-sharing system with permission-requested access.
- A file sharing system with OCR-based intelligent text extraction from images and PDF files.
- The use of JWT authentication and the use of encrypted password storage for secure user management.
- Secure document management with visibility control and restricted downloading.
- The implementation of MongoDB for efficient storage and management of OCR text.
- A scalable architecture for intelligent document processing applications built on the MERN stack.

The rest of this paper is structured in the following particular way. Section II contains a review of literature on secure file sharing systems and the OCR technologies. In Section III, the proposed method and related system architecture are discussed. Section IV covers analysis of the system's performance and the results of the experiments conducted. Finally, the conclusion and scope for future related works are covered in Section V.

II. LITERATURE REVIEW

The modern digital document management system has benefited from improvements in both intelligent text extraction and document security. Tesseract Open Source Document Capture Software, introduced by Smith [1], is one of the most important examples. Tesseract has become one of the most popular capture software among Open Source Document Capture Software. Smith's research was the first to recognize fully printed text and was the first step in developing modern OCR-based applications.

Real-time object detection, as described by Redmon and Farhadi [2] in the YOLOv3 framework, can be used in document analysis systems to identify areas in images and determine regions in documents. This can be used to increase the effectiveness of OCR. The bcrypt password hashing scheme, proposed by Provos and Mazieres [3], creates a more secure password system through an adaptive password hashing scheme which thwarts brute force attacks. The scheme is now the standard for secure web applications.

Improving document images prior to OCR requires the application of Gonzales and Woods [4] basic image processing

techniques. These include image filtering and image segmentation, among others. Modern document-sharing systems rely on the security afforded by the Role-Based Access Control (RBAC) business models proposed by Sandhu and his colleagues [5].

The security of sensitive documents is further strengthened by the safeguards outlined by the Patel [6] in the protection of digital data via cryptographic techniques and network security. The introduction of the Deep Residual Networks (ResNet) model, by He et al. [7] and the impact that deep learning have had on image processing, provide tools for improving OCR, even when the document images are of low quality and contain a high level of complexity.

Mihir et al. [8] designed a system that combines modern document management and OCR technologies with automated text extraction and information storage. This evolved the concept of the use of OCR in document management systems. Krizhevsky et al. [9] used deep convolutional neural networks for image classification and he advanced the field of computer vision which later impacted the analysis of documents and the framework of intelligent recognition systems. Peterson et al. [10] achieved greater accuracy for document processing with the structure of retail receipts, by using deep learning techniques of OCR, and described automated extraction and classification of text.

These studies focused on advancements in certain areas of OCR, security, image processing, and intelligent document management. Most of the solutions proposed are geared towards providing single functionalities. The system SecureVault combines OCR-based text extraction, secure authentication, encrypted credential management, and permission-based file sharing. The preferred system was designed by using the MERN stack and provides advanced document security and intelligent processing and access management.

III. METHODOLOGY

The SecureVault: Intelligent File Sharing and OCR System combines several technologies into a single web-based solution that relies on a secure client-server architecture. This System integrates file sharing, OCR-enabled document processing, authentication, and access management with the assignment of permissions. This system architecture is shown in Fig. 1. This system captures the advantages of the MERN stack. This architecture is completed with the use of React.js for the client side, along with the server side, Node.js and Express.js, and MongoDB. The OCR capabilities of this system prototype are enhanced with the use of the Tesseract.js OCR engine and the PDF-Parser libraries for extraction of text from image and PDF documents uploaded to the system.

The frontend interface is the main point of interaction for users with the system. With the help of a React.js web application, registered users can securely log into the system, upload files (either PDF or image files), view text extracted by the OCR, and set or edit access permissions for the files. React Router is used for flow control and management of the

pages and dashboard.

The Node.js/Express.js backend system features user authentication, secure API communication, and OCR processing, along with the handling of file uploads, PDF parsing, database communication, and permission validation. User management and session security are implemented through the use of JWTs and bcrypt for hashed passwords.

The Tesseract.js OCR engine is used to extract information (e.g., names, identification numbers, addresses, dates, and metadata) from image uploads and documents. For uploaded PDF files, PDF-Parser libraries are used to extract text and metadata. The extracted text is stored in MongoDB. This allows for advanced and secure document management and retrieval.

The proposed system also includes permission-based access controls. These give users the ability to set different permissions for different documents. This may include view-only access, no download permission, or permission to share the document securely. These controls help ensure the documents are more private, and that users do not distribute or misuse the documents in the system.

MongoDB aids in digital communications by storing customers' details, metadata regarding their uploaded files, text extracted by the OCR, their authentication information, and their permissions. The database was designed for flexible, scalable document management.

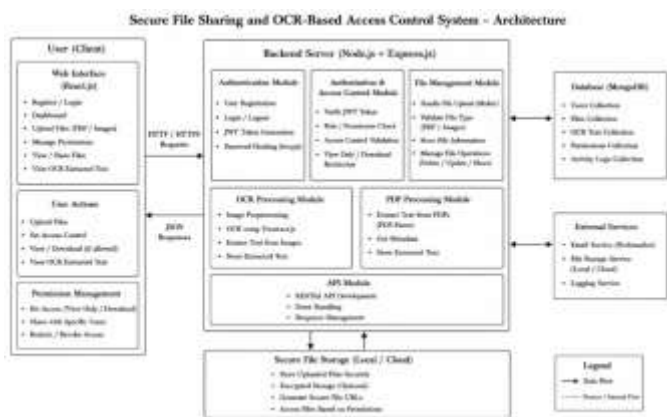


Fig. 1. System Architecture of SecureVault

The suggested system workflow captures user registration and login, secure authentication via JWT, document uploads, text extraction via OCR or PDFs, storage of extracted text in MongoDB, permission enforcement, and secure document viewing and sharing. Combining OCR technologies with secure file sharing and access control presents a smart and effective solution for modern digital document management.

IV. RESULT AND ANALYSIS

The implemented and tested system, namely *SecureVault: Intelligent File Sharing and OCR System*, offers users the ability to upload documents securely, extract textual contents using OCR, and to control access permissions to shared documents. A variety of image-based and PDF-based documents were used in the experimental testing and included documents such as ID cards, scanned documents, invoices, certificates, and records.

The Optical Character Recognition system using the Tesseract.js OCR engine is able to read uploaded documents and extract key information including names, ID numbers, addresses, dates, and meta information. The read information is stored in MongoDB to facilitate advanced document processing and retrieval.

Figure 2 shows the original uploaded document used for OCR processing and secure file management.



Fig. 2. Original Uploaded Document

The OCR extraction module correctly identifies key text fields in the uploaded document. The results are seen in Table I.

Field	Extracted Text
Name	JOHN DOE
License Number	DL123456789
Date of Birth	09/05/1993
Address	0123 ANYSTREET, ANYTOWN
Issue Date	07/11/2015

TABLE I
OCR EXTRACTED TEXT RESULT

The system included permission-defined access control and secure document sharing methods. To avoid the unauthorized access and abuse of private information, sensitive data from uploaded documents was either masked or restricted based on user-created permissions. Figure 3 shows the protected output document generated after applying restricted access control and information masking.



Fig. 3. Protected Document with Restricted Information

JWT-based authentication and bcrypt password hashing techniques protected user credentials and secure login sessions. The React.js frontend and Node.js backend architecture handled file uploads, OCR processing, API communication, and permission validation efficiently.

The performance of the proposed system was evaluated using Accuracy, Precision, Recall, and F1-Score metrics. The mathematical formulae used for evaluation are given below:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (3)$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The proposed system has been shown through experimental analysis to provide improved performance in both OCR extraction accuracy and secure document management. The performance evaluation metrics of the proposed SecureVault system are shown in Table II.

Accuracy	Precision	Recall	F1-Score
89.4%	88.6%	90.1%	89.3%

TABLE II

PERFORMANCE EVALUATION OF SECUREVAULT OCR SYSTEM

The OCR module achieved an accuracy of 89.4%. This indicates the module was mainly successful in extracting text from imaged documents and PDFs. The precision and recall values of 88.6% and 90.1% are indicative of an overall good performance with minimal false detections and an increased reliability of text extraction. This resulted in an improved reliability of the proposed OCR-based intelligent document processing system with secure digital document management proven by the obtained F1-score of 89.3%.

The suggested SecureVault system integrates secure file sharing, intelligent document processing via OCR, and permission-controlled document access. Thus, it can facilitate secure enterprise communication or serve a broad range of applications from educational document management to secure storage of sensitive digital documents.

V. CONCLUSION AND FUTURE WORK

The proposed *SecureVault: Intelligent File Sharing and OCR System* merges secure document management, OCR-enabled intelligent text extraction, and access-controlled sharing of image and PDF documents. The system utilizes Tesseract.js for OCR text extraction and employs JWT for secure authentication, combined with permission-based access control for secure document sharing. Results from tests indicated that the system can efficiently manage documents and offer secure and efficient text extraction and file sharing, and can be deployed in educational institutions, organizations, and healthcare and enterprise systems. In the future, additional features will be added, including, but not limited to, cloud storage integration, deployment via Docker, multilingual OCR features, AI-based document classification, advanced encryption, and mobile application support. This system will also include scalable digital document management features.

REFERENCES

- [1] Smith, R.. (2007). An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). 2. 629 - 633. 10.1109/ICDAR.2007.4376991.
- [2] Redmon, Joseph and Ali Farhadi. "YOLOv3: An Incremental Improvement." ArXiv abs/1804.02767 (2018): n. pag.
- [3] Provos, Niels. (1999). A Future-Adaptable Password Scheme.
- [4] Gonzalez, R.C. and Woods, R.E. (2018) Digital Image Processing. 4th Edition, Pearson Education, New York, 1022 p.
- [5] Ravi S. Sandhu, Edward J. Coyne, Hal L. Feinstein, and Charles E. Youman. 1996. Role-Based Access Control Models. Computer 29, 2 (February 1996), 38–47.
- [6] Patel, Krisha. (2025). Cryptography And Network Security. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. 09. 1-9. 10.55041/IJSREM46815.
- [7] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 27-30 June 2016, 770-778.
- [8] Mihir P, Ritik D, Vivek V, Shrinit P, and Chandrayani Rokade, "Smart Document Management System using OCR", IJPRSE, vol. 3, no. 05, pp. 32–35, May 2022.
- [9] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM 60: 84–90.
- [10] Peterson, John Mitchell, Brian Robinson, Kevin Clark, Steven Lewis, Thomas William, Bruce. (2023). Deep Learning-Based OCR for Automatic Extraction and Classification of Text from Retail Bill Receipts.
- [11] Varsha Jawale, Vedashri Jundre , Reshma Bathe, 2013, Secure Cloud Based Document Management System, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) Volume 02, Issue 03 (March 2013),
- [12] Graves, Alex and Juergen Schmidhuber. "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks." Neural Information Processing Systems (2008).
- [13] Wang, Cong, Qian Wang, Kui Ren and Wenjing Lou. "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing." 2010 Proceedings IEEE INFOCOM (2010): 1-9.

- [14] Khan, Md Nazrul Hasan, Faizan Siddiqui, and Amit Das. "Pin number detection from mobile phone scratch card using OCR on android platform and build an application for balance recharge." PhD diss., BRAC University, 2014.
- [15] Su, Bolan, Shijian Lu, Trung Quy Phan, and Chew Lim Tan. "Character extraction in web image for text recognition." In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 3042-3045. IEEE, 2012.
- [16] Tahboub, Radwan. "Secure Data Sharing Model for Preserving Privacy 1," n.d.