Synopsis

on

“NoteScan”

Submitted in the partial fullfilment of the requirement for the award of degree of

Bachelor of Technology

in

Computer Science and Engineering

Batch

(2022-2026)

Submitted To:

Dr. Naveen Bilandi

Associate Professor

Submitted By

Geetanjally

12200821

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DAV UNIVERSITY

JALANDHAR-PATHANKOT NATIONAL HIGHWAY, NH44,

SARMASTPUR PUNJAB

144001

TABLE OF CONTENT

1. CHAPTER 1: INTRODUCTION

1.1 Features

1.2 Advantages

1.3 Disadvantages

1.4 Steps of Working

1.5 Architecture

1.6 Scope

1.7 Significance

2. CHAPTER 2: LITERATURE SURVEY

2.1 Existing Platforms

2.2 Limitations

2.3 Gap Analysis

3. CHAPTER 3: PROBLEM STATEMENT

3.1 Identification and Formulation of a Problem

3.2 Need and Significance of Proposed Work

4. CHAPTER 4: OBJECTIVES

5. CHAPTER 5: METHODOLOGY

5.1 Input Acquisition

5.2 Preprocessing (Computer Vision)

5.3 OCR & Text Extraction

5.4 Text Cleaning & NLP Processing

5.5 Summarization & Note-Making

5.6 PowerPoint Generation

5.7 User Interface

5.8 Tools & Technologies

5.9 Expected Outcomes

6. CHAPTER 6: CONCLUSION

CHAPTER 1: INTRODUCTION

In today's academic and professional environments, a large volume of information continues to be handwritten, particularly in classrooms, research discussions, and workshops. While digital transformation has advanced rapidly, converting handwritten notes into structured, reusable digital content remains a challenge. Traditional OCR (Optical Character Recognition) systems work effectively for printed text but face major limitations with handwritten inputs due to varied writing styles, inconsistent spacing, and noisy scanned backgrounds.

NoteScan is designed to bridge this gap by providing an AI-powered solution that transforms handwritten or scanned PDF notes into structured, study-ready formats. Unlike conventional OCR tools, NoteScan integrates Computer Vision (CV), Deep Learning (DL), and Natural Language Processing (NLP) techniques to go beyond raw text extraction. It converts scanned handwriting into cleaned, structured notes, generates concise summaries, and even creates ready-to-use PowerPoint slides. This makes it a complete academic assistant for students, teachers, researchers, and professionals.

1.1 Features

- Automated Handwriting Recognition: Advanced OCR and deep learning models to extract text from handwritten and scanned PDFs.
- Content Structuring: Intelligent grammar correction, segmentation, and formatting for readability.
- Summarization & Notes: NLP-based summarization for bullet points and concise study notes.
- PPT Generation: Automatic slide creation highlighting key points, diagrams, and flow of concepts.
- Multi-format Export: Outputs in TXT, DOCX, PDF, and PPTX formats for wide usability.
- User Interface: A Streamlit-based platform for easy uploads and direct downloads.

1.2 Advantages

- Time-Saving: Eliminates manual rewriting and structuring of handwritten notes.
- Multi-Format Support: Exports to multiple file formats for academic and professional use.
- Accessibility: Cross-device support ensures that notes can be accessed anytime, anywhere.
- Enhanced Usability: Summarization and presentation generation add value beyond text recognition.

1.3 Disadvantages

- Accuracy may vary with extremely poor handwriting or low-quality scans.
- Requires internet access for cloud-based processing and NLP integration.
- Initial computational cost for deep learning models may be high.

1.4 Steps of Working

1. User uploads a scanned handwritten PDF.
2. Preprocessing techniques clean the input for better recognition.
3. OCR and deep learning models extract raw handwritten text.
4. NLP modules clean, summarize, and structure the content.
5. PPT slides and formatted study material are auto-generated.
6. User downloads outputs in multiple formats.

1.5 Architecture

- Frontend: Streamlit-based interface for uploads, previews, and downloads.
- Backend: Python-based pipeline integrating OCR (Tesseract, Deep Learning models), NLP (transformers for summarization, text cleaning).
- Storage & Export: Files are processed and exported into TXT, DOCX, PDF, and PPTX formats.

1.6 Scope

- NoteScan is not limited to academic use but can also benefit researchers, corporate professionals, and institutions handling handwritten manuscripts or reports. Future improvements may include multi-language handwriting support, mobile app integration, and cloud storage synchronization.

1.7 Significance

This project represents a step forward in EdTech innovation, addressing the gap between handwritten input and digital usability. By delivering structured notes, summaries, and presentations, NoteScan enhances learning efficiency, reduces manual effort, and promotes digital accessibility. Its holistic approach makes it more powerful than traditional OCR tools and positions it as a valuable tool in both education and professional fields.

CHAPTER 2: LITERATURE SURVEY

This survey reviews key platforms that convert scanned PDFs to PowerPoint using Optical Character Recognition (OCR).

1. Adobe Acrobat

- Description: An industry-leading tool that uses robust OCR to convert scanned PDFs.
- Pros: High-quality conversion, excellent preservation of formatting.
- Cons: Requires a paid subscription for full features.
- Link: Adobe Acrobat PDF to PPT Converter

2. Smallpdf

- Description: An online, user-friendly platform with a combined PDF to PPT and OCR tool.
- Pros: Simple, all-in-one workflow; emphasizes security.
- Cons: OCR is a paid "Pro" feature.
- Link: Smallpdf PDF to PPT Converter

3. Nitro

- Description: A free online tool that uses OCR to convert scanned PDFs to PPT.
- Pros: Free to use, no registration needed, browser-based.
- Cons: Paid desktop version (Nitro Pro) is required for advanced features like batch processing.
- Link: Nitro PDF to PPT Converter

2.2 Limitations of Existing Systems

- Focus only on basic text extraction rather than contextual refinement.
- Struggle with noisy classroom notes, mixed diagrams, and inconsistent handwriting.
- Limited or no support for generating study-ready formats such as bullet points, summaries, or slides.

2.3 Gap Analysis

The current market lacks a dedicated solution that transforms handwritten content into educationally usable resources. Existing tools stop at text recognition but fail to provide semantic structuring and multi-format outputs tailored for students and educators.

CHAPTER 3: PROBLEM STATEMENT

The problem addressed in this project is the lack of an end-to-end AI-powered system that converts handwritten or scanned PDFs into structured, study-ready digital content. Traditional OCR tools are effective for printed text but perform poorly on handwriting due to variations in style, image quality, and noise. Moreover, most tools stop at raw text extraction and do not provide meaningful summarization, structuring, or presentation features.

NoteScan is proposed as a comprehensive solution that integrates Computer Vision (CV) for preprocessing, Deep Learning models for handwriting recognition, and Natural Language Processing (NLP) for cleaning, summarization, and structuring. The system will also generate PowerPoint slides and export results in multiple formats such as TXT, DOCX, PDF, and PPTX.

3.1 Identification and Formulation of a Problem

Handwritten notes remain common in classrooms, workshops, and research settings, yet digitizing them into usable formats is still challenging. Current limitations include:

- Inconsistent handwriting styles, lowering recognition accuracy.
- Low-quality scans with noise or skew.
- Unstructured outputs, where extracted text is not ready for study or presentation.

3.2 Need and Significance of Proposed Work

The proposed work is significant because it provides a complete AI-driven pipeline rather than a single-step OCR tool. NoteScan enhances productivity, ensures multi-format export and cross-device access, and caters specifically to educational and professional needs, making it more impactful than existing systems.

CHAPTER 4: OBJECTIVES

1. What to Do

The project aims to automate the process of converting handwritten or scanned PDFs into structured, summarized, and presentation-ready formats.

2. Tools and Technologies Used

The system will be developed using Python with libraries like Tesseract OCR, OpenCV, and PyTorch/TensorFlow for handwriting recognition, NLTK/Transformers for NLP tasks, and Streamlit for the web interface.

3. How to Do It

- Extracting handwritten text with OCR and deep learning models.
- Cleaning and structuring the extracted text into readable formats.
- Summarizing notes using NLP models.
- Automatically generating PowerPoint slides from summarized key points.

4. Expected Outcome

The project will deliver a user-friendly platform where students or professionals can upload handwritten PDFs and download processed results in TXT, DOCX, PDF, and PPTX formats.

CHAPTER 5: Methodology

5.1 Input Acquisition

- Users upload scanned handwritten PDFs.
- PDFs converted into images using PyMuPDF or pdf2image.

5.2 Preprocessing (Computer Vision)

- Applied using OpenCV.
- Techniques:
  - Noise removal
  - Thresholding
  - Skew correction
- Purpose: To improve OCR accuracy.

5.3 OCR & Text Extraction

- Baseline Model: Tesseract OCR (for printed text).
- Advanced Models: CRNN and TrOCR (for handwritten text).
- Enables recognition across varied handwriting styles.

5.4 Text Cleaning & NLP Processing

- Libraries: NLTK, SpaCy.
- Tasks performed:
  - Tokenization
  - Stopword removal
  - Lemmatization & POS tagging
  - Named Entity Recognition (NER)
- Outcome: Clean, structured, and grammatically consistent text.

5.5 Summarization & Note-Making

- Extractive Models: TextRank, BertSum.
- Abstractive Models: BART, T5.
- Outputs:

        - Bullet points
        - Concise summaries
        - Detailed notes

5.6 PowerPoint Generation

- Tool: python-pptx.
- Converts structured text into slides with headings, subpoints, and diagrams.

5.7 User Interface

- Framework: Streamlit.
- Features:
    - Upload PDF option
    - Download outputs (TXT, DOCX, PDF, PPTX)
    - Clean and user-friendly interface.

5.8 Tools & Technologies

- Programming Language: Python
- Computer Vision & OCR: OpenCV, PyMuPDF, pdf2image, Pytesseract
- Deep Learning Models: CRNN, TrOCR (HuggingFace)
- NLP Libraries & Models: NLTK, SpaCy, BERT, BART, T5
- Summarization Frameworks: HuggingFace Transformers
- PPT Generation: python-pptx
- Interface Development: Streamlit
- Environments/IDEs: Jupyter Notebook, VS Code, Google Colab

5.9 Expected Outcomes

- Clean, editable text extracted from handwritten/printed PDFs.
- Structured notes in bullet point and detailed formats.
- Automatically generated PowerPoint slides.
- Export options: TXT, DOCX, PDF, PPTX.
- Simple and accessible Streamlit interface for deployment.

CHAPTER 6: CONCLUSION

NoteScan provides an innovative step forward in digitizing and structuring handwritten and scanned notes. Unlike traditional OCR systems that only extract raw text, NoteScan leverages Computer Vision, NLP, and Deep Learning to deliver accurate recognition along with structured and summarized outputs. This combination reduces manual effort, making study materials and professional documents easier to prepare and share.

The system's ability to generate clean text, concise summaries, and even PowerPoint presentations highlights its versatility. By integrating advanced summarization models like BART and T5, NoteScan ensures that the content is not only digitized but also transformed into meaningful, study-ready formats.

With its user-friendly interface, NoteScan caters to students, educators, and professionals alike. As AI in EdTech continues to expand, this project demonstrates how technology can improve productivity, accessibility, and efficiency, ultimately contributing to smarter and more effective content digitization solutions.

CHAPTER 7: REFERENCE

1. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
2. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33418.pdf