Marks2CSV

A simple solution to convert tabular mark fields to CSV file

Mini Project Presentation: Zeroth Review

Guided by:
Dr. Deepa V.

Presented by:

Ajay T Shaju,	SJC20AD004
Emil Saj Abraham,	SJC20AD028
Justin Thomas Jo,	SJC20AD046
Vishnuprasad KG,	SJC20AD063

Outline

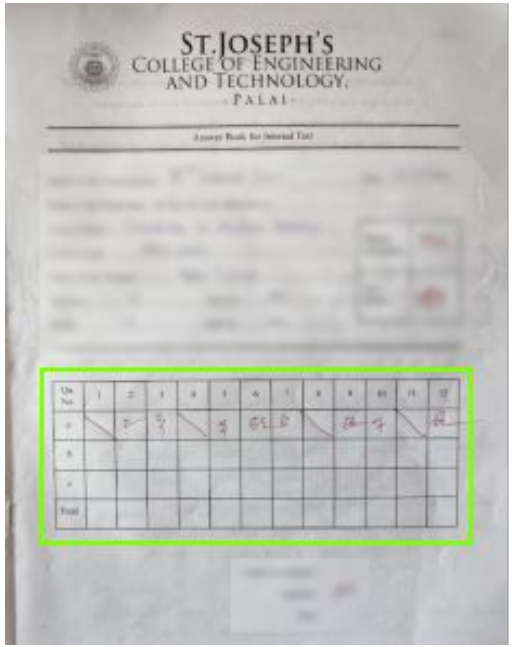
- Introduction
- Motivation
- Literature Review
- Objectives
- Block diagram
- Data Description
- Future Scope
- Conclusion
- References

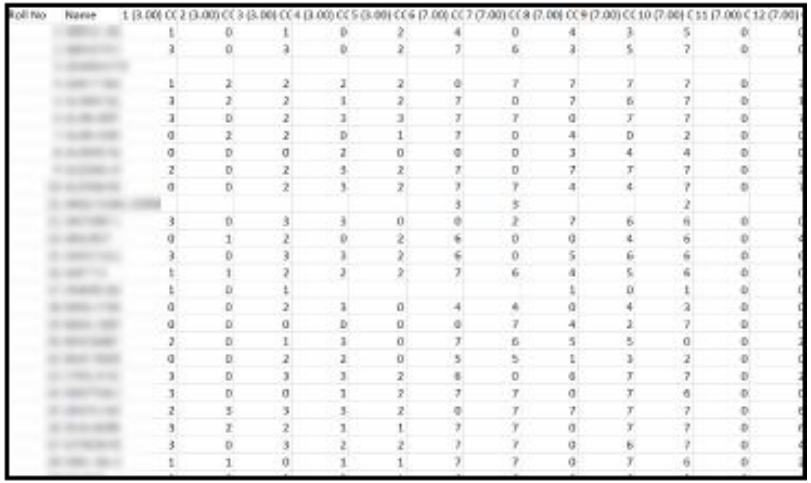
Introduction

"Technology will never replace great teachers, but technology in the hands of great teachers can be transformational." - The Innovator's Mindset, George Couros

- Introducing our time-saving idea of digitizing handwritten documents. It could automate various data entry processes.
- The key technology will be Optical Character Recognition.
- OCR, or Optical Character Recognition, is a technology that enables the conversion of scanned or photographed images of text into digital text that can be edited, and searched by a computer. OCR is commonly used to digitize handwritten or paper-based documents.
- We can apply a part of our project for automation of mark data entry of our teachers,

We could automate





We approximated that a <u>teacher teaching three subjects to about 60 students</u> in each class would take <u>4-6 hours to create Excel sheets</u>. But <u>with our idea, we can do it in under 30 minutes.</u>*

* Approximate calculation

Time calculation breakup*

- 60 students, each attended 10 questions = 600 marks
  - Questions may have split-ups (7a, 8b) = 600 + 100 = 700 marks
- Likewise 3 classes = 700 x 3 = 2100 marks
  - 1900 Mouse movements / keystrokes
    - Total 4000 keypresses (2100+1900)
- To words per minute (avg 35) = 4000 ÷ 35 = 114.28 min ~ 2 hours
  - Adding all miscellaneous works(error checking, retyping) = 2 hours
- So, a total of **4 hours** for an average person.
- This **4 hour** time frame can extend upto **6 hours** due to unforeseen circumstances.

***** Approximate calculation

Motivation

- We have observed efficient teachers entering marks manually into CSV files which made us empathise with them, and our plan to find a solution inspired the creation of this project.

- Using OCR technology, we can automatically convert handwritten marks into characters, eliminating manual data entry and speeding up the evaluation process.

- Our solution can help teachers save time, so that they can dedicate more of their time to plan new methods of teaching.

Literature Review

[3] Raajkumar G., Indumathi D., “Optical Character Recognition using Deep Neural Network”, Vol. 176 No. 41 pp:61-65, July 2020.

[5] B.Nunamaker, Syed S.Bukhari, D.Borth, A.Dengel, “A Tesseract-based OCR Framework For Historical Documents Lacking Ground-Truth Text”, IEEE ICIP pp: 3269-3273, 2016.

[6] V.Wu, R.Manmatha, Edward M.Riseman, “TextFinder: An Automatic System to Detect and Recognize Text In Images”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 21, No. 11., pp: 1224-1229, November 1999.

[7] R.Deepa, Kiran N.Lalwani, “Image Classification and Text Extraction using Machine Learning”, pp: 680-684, Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019].

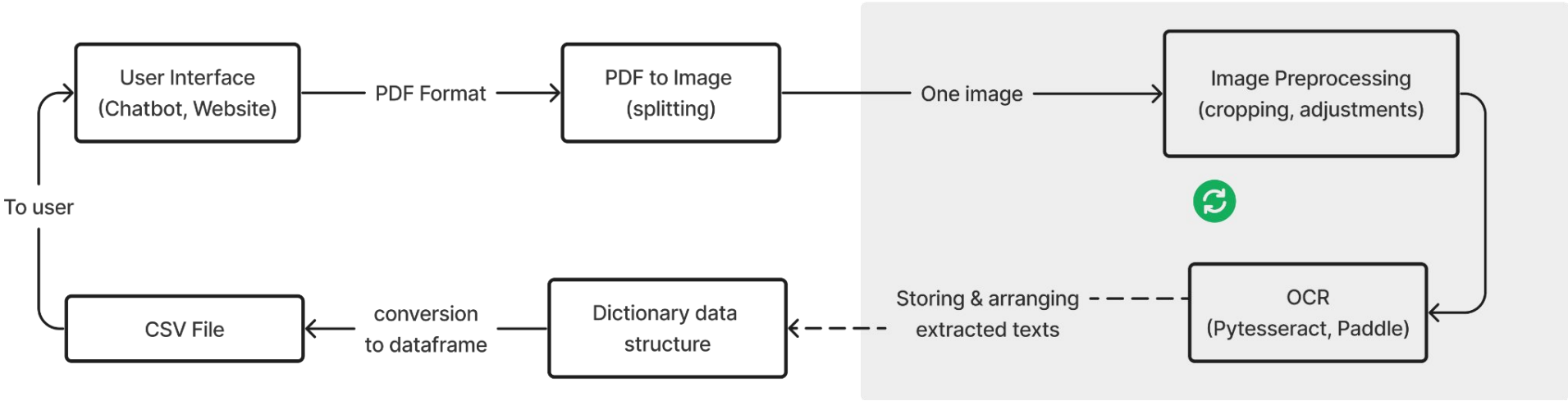
Inferences

- Deep Learning models Convolutional Neural Networks (CNN) & Long Short Term Memory (LSTM) are commonly used.
- Most papers have implemented OCR technologies like Tesseract.
- Cleaning noise and outliers can improving the accuracy.
- After comparing the performance of Support Vector Machines (SVM) and Tesseract for text recognition, it was found in [3] that Tesseract, which uses deep learning, outperformed SVM.

Objectives

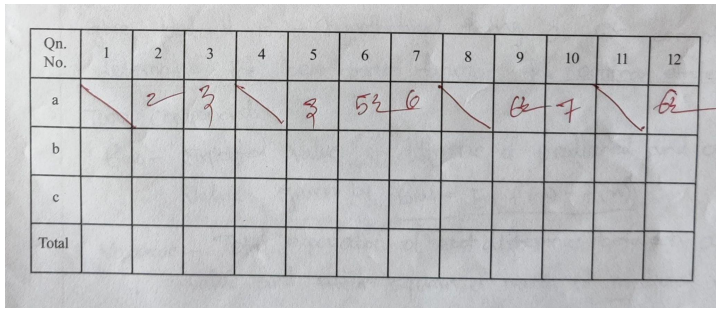
- The application needs to offer an easy-to-use interface and should accept a popular file format (like PDF) which has all mark sheets, making it effortless for teachers to send and receive files.
- Need to make a custom OCR tool that is trained on our datasets rather than importing big libraries.
- We aim to minimize processing time for large quantities of data by translating the codebase to a high-performance language like C++ (if it can reduce processing time).
- Project development should follow systematic approach and utilize top technologies such as GitHub for code management, Figma & LaTeX for better graphics and slides, and Notion for project tracking.

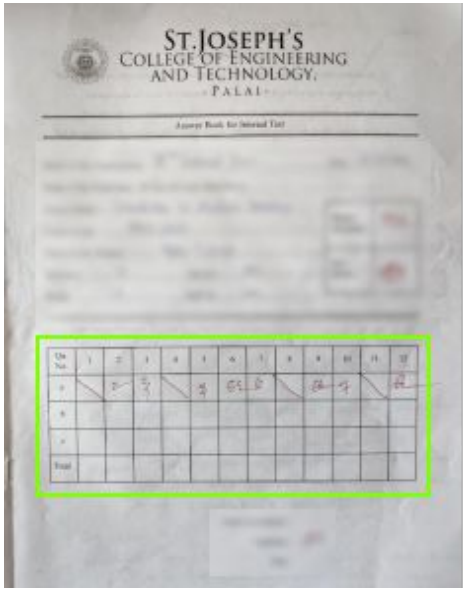
Block Diagram



Data Description

- Primary data - ‘Answer sheets of our college’ - image data of A4 size.
- The region of focus is the big mark table on the front page of the answer sheet.
- We require preprocessing techniques to improve the quality of the image data such as image resizing, noise reduction, and adjustments.





Future scope

- The model can be applied to different institutions for easy document-digitization.
- We plan to incorporate more powerful & less time consuming methods for the main project.
- We hope to make Marks2CSV as our main project or a startup, if it can be scaled as per our expectations.

Conclusion

- The application will be efficient and user-friendly for converting mark files to CSV format for report-making, data processing, and analysis.
- The development process will follow a systematic approach, including requirement gathering, design, development, testing, and deployment.
- Overall, this project can simplify the mark documenting process, providing a valuable solution and a great help for the teachers.

References

[1] Ali F.Biten, R.Tito, L.Gomez, E.Valveny, and D.Karatzas, “OCR-IDL: OCR Annotations for Industry - Document Library Dataset”, 25 Feb 2022.

[2] Ömer Aydin, “Classification of Documents Extracted from Images with Optical Character Recognition Methods”, Vol.6 No.2 pp:46-55, 01 Jun, 2021.

[3] Raajkumar G., Indumathi D., “Optical Character Recognition using Deep Neural Network”, Vol. 176 No. 41 pp:61-65, July 2020.

[4] Y.Yu, Y.Li, C.Zhang, X.Zhang, Z.Guo, X.Qin, K.Yao, J.Han, E.Ding, J.Wang, “StrucTexTv2: Masked Visual-Textual Prediction for Document Image Pre-Training”, ICLR Conference, 1 Mar 2023.

[5] B.Nunamaker, Syed S.Bukhari, D.Borth, A.Dengel, “A Tesseract-based OCR Framework For Historical Documents Lacking Ground-Truth Text”, IEEE ICIP pp: 3269-3273, 2016.

[6] V.Wu, R.Manmatha, Edward M.Riseman, “TextFinder: An Automatic System to Detect and Recognize Text In Images”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 21, No. 11., pp: 1224-1229, November 1999

[7] R.Deepa, Kiran N.Lalwani, “Image Classification and Text Extraction using Machine Learning”, pp: 680-684, Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019]

[8] Pranav P.Nair, A.James, C.Saravanan, “Malayalam Handwritten Character Recognition Using Convolutional Neural Network”, pp: 278-281, International Conference on Inventive Communication and Computational Technologies (ICICCT 2017)

[9] Christos N.E.Anagnostopoulos, Ioannis E. Anagnostopoulos, Vassili Loumos, Eleftherios Kayafas, “A License Plate-Recognition Algorithm for Intelligent Transportation System Applications”, IEEE Transactions On Intelligent Transportation Systems, Vol. 7, No. 3, pp: 377-392, September 2006

Questions?

Thank You

Presentation Setting

Opening(title) - Ajay

Intro, automate, time calc - Ajay

Motivation - Justin

Review of Lit - Justin

Objectives - Vishnu

Block diagram - Vishnu

Data Description - Emil

Future Scope - Emil

Conclusion - Ajay

References - Justin

Questions, Thank you - All members of the team

Don’t Present this

Don’t Present this

Introduction

Extra copy

"Technology will never replace great teachers, but technology in the hands of great teachers can be transformational." - The Innovator's Mindset, George Couros

Introducing our time saving idea of a mark recognition system named Marks2CSV. It could automate the data entry process of our teachers, like manually typing in all the marks scored by students in an exam.

For our system, we harness the technology of Optical Character Recognition for mark(digit) recognition.

OCR, or Optical Character Recognition, is a technology that enables the conversion of scanned or photographed images of text into digital text that can be edited, and searched by a computer. OCR is commonly used to digitize written or paper-based documents.

OCR software works by analyzing the patterns and shapes of characters in an image and then matching them to a database of known characters.

Marks2CSV

A simple solution to convert tabular mark fields to CSV file

Extra copy

Mini Project Presentation: Zeroth Review

Guided by:
Dr. Deepa V.

Presented by:
Ajay T Shaju, SJC20AD004
Emil Saj Abraham, SJC20AD028
Justin Thomas Jo, SJC20AD046
Vishnuprasad KG, SJC20AD063

Literature Review

[3] Raajkumar G., Indumathi D., “Optical Character Recognition using Deep Neural Network”, Vol. 176 No. 41 pp:61-65, July 2020

- Data is preprocessed through filtering, morphological operations, normalization, segmentation etc.
- OCR conversion performed using CNN and LSTM architecture.
- Implementation using PyTesseract.
- Performance analysis using confusion matrix and visualization graphs.

[5] B.Nunamaker, Syed S.Bukhari, D.Borth, A.Dengel, “A Tesseract-based OCR Framework For Historical Documents Lacking Ground-Truth Text”, IEEE ICIP pp: 3269-3273, 2017.

- Using Tesseract OCR system for training the model.
- Cleaning noise and outliers and improving the accuracy.
- Using mean squared error technique to evaluate the similarity of the individual character images of a subset and the resulting model’s performance
- Implementation of random selection models and MSE models

[6] V.Wu, R.Manmatha, Edward M.Riseman, “TextFinder: An Automatic System to Detect and Recognize Text In Images”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 21, No. 11., pp: 1224-1229, November 1999

- Text Segmentation Module and its importance
- Chip - Generation Module and its various steps
- Chip Scale Fusion Module
- Text Cleanup Module and Chip Refinement Module

[7] R.Deepa, Kiran N.Lalwani, “Image Classification and Text Extraction using Machine Learning”, pp: 680-684, Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019]

- Basic understanding of overfitting and underfitting situations
- Image classification using CNN
- Text Classification using PyTesseract and ROI Extraction
- Performance Analysis - Comparing accuracies of CNN and SVM, comparing accuracies of Tesseract OCR and OCRopus OCR

Data Description

Excluded
May be useful in future

- Primary data are ‘Answer sheets of our college’, which are image data of A4 size
- The data comes under ‘college documents’ domain
- Initial data can be collected from our department, for more data variability, we could request it from other departments of our college itself.
- Data can be compressed to file format like PDF for easy portability
- Data size: Input data can have 5-10MB (PDF), and output data can have 4KB - 2MB (CSV)
- Data may be unstructured and papers may be damaged or faded.

Time Breakup

Excluded

May be useful in future

For processing one page:

- Upload time: 10-30 seconds (depending on the file size and internet speed)
- OCR software processing time: 1-5 seconds
- Text extraction and formatting time: 2-10 seconds
- Total processing time per page: 3-15 seconds

For processing 60 pages:

- OCR software processing time: 1-5 seconds per page
- Text extraction and formatting time: 2-10 seconds per page
- Total processing time for 60 pages: 3-15 minutes