20CS713 PROJECT PHASE 1- C18

Converting Handwritten Images and Documents to text using Advanced OCR Techniques

TEAM MEMBERS
YEGNESHWARAN B
SATHYA MOORTHY S

MENTOR: Mr M.P.KARTHIKEYAN
Asst Professor, CSE

ABSTRACT

In today's digital age, the need to convert images and documents into text is essential. Handwriting recognition is a computer system that can recognize and identify characters and letters written by users. Over the years, researchers have used AI/machine learning tools to automatically analyze written and printed documents and convert them into electronic ones.
The purpose of this article is to summarize research in the behavioral literature and provide research suggestions. Optical character recognition (OCR) technology has made progress toward this goal, but typing remains a challenge. OCR is a technology that can be used to convert text into digital text, but current OCR can be inaccurate, especially when processing text. This project proposes a solution that uses advanced OCR technology to fill this gap and replace the content of the text. Many businesses benefit.

PROBLEM STATEMENT

Handwritten text is a valuable source of information, but it can be difficult to process and store electronically. OCR is a technology that can be used to convert handwritten text to digital text, but existing OCR systems can be inaccurate, especially when dealing with handwritten text.

The goal of this project is to develop an OCR system that is specifically designed to handle handwritten text. The system should be able to accurately recognize handwritten characters in a variety of fonts and styles, even in challenging conditions such as low light or poor quality images.

The system should also be able to identify and extract the logical structure of the handwritten document, such as paragraphs, headings, and lists. This will allow the converted text to be easily formatted and edited in a word processing program poor-quality

PROJECT OBJECTIVE

To develop a system that can accurately and reliably convert handwritten images to documents to text using Optical Character Recognition (OCR).

TECHNOLOGY USED

1. OCR (Optical Character Recognition)
2. DEEP LEARNING MODEL
3. OPENCV
4. PYTHON AND SQLITE

LITERATURE SURVEY

[1] Online recognition of multi-stroke handwritten Urdu characters

categorized the number of strokes into four classes. The study proposed online intelligent Urdu character recognition by considering a single-stroke character. Three different classifiers, namely, a correlation-based classifier, backpropagation neural network (BPNN), and probabilistic neural network (PNN), were fed with statistical features that were pre-extracted. The dataset contained 85 instances from 35 writers. The author claimed that the PNN achieved the best accuracy of 94%. PNN-based classifiers do not require pre-training, thus resulting in higher accuracy compared to the BNN-based classifier. A comprehensive summary related to Urdu handwritten character recognition

[2] The BBN Byblos Pashto OCR system

the Byblos Pashto OCR system was proposed for script-free OCR using HMMs. This system was also subsequently tested for Chinese, English, and Arabic text with success. As previously mentioned, Pashto Intelligent character recognition(ICR) and Optical character recognition (OCR) area are the least explored to date. Thus, due to the unavailability of a Pashto corpus, the authors of this paper collected 27,000 characters from faxed printed pages. They then scanned these pages using a 300 dpi scanner. This research work lacked in performance due to insufficient training data

[3] Efficient DNN Execution on Intermittently-Powered IoT Devices With Depth-First Inference

suggested the Deep convolutional neural networks (DCNN)model for the recognition of isolated handwritten Arabic characters. They proposed a dataset referred to as Arabic handwritten characters dataset AHCD. The model was trained with an optimization method for 30 epochs that resulted in a significant increase in performance and a 94.9% classification accuracy.

[4]HANDWRITTEN ARABIC CHARACTER RECOGNITION USING MULTIPLE CLASSIFIERS BASED ON LETTER FORM

Conventional approaches depend on manual feature extraction by experts whereas deep learning-based techniques automatically extract features from raw images. Conventional techniques cannot extract features from images in their raw form. Machine learning experts have struggled to design feature extractors that extract discriminative features from raw data into vectors as an input to classifiers for pattern recognition

[5] Pashto Isolated Character Recognition Using K-NN Classifier

The authors suggested a Pashto OCR (optical character recognition) system. A small dataset was compiled with 1125 entries. In the proposed approach, individual Pashto characters were recognized by utilizing both high- and low-level features. High-level features were based on the structural information from the characters and the resulting binary trees uniquely classified each of the characters. Although the approach was robust, it was affected by the variation in size, orientation, and writing style. An alternative low-level feature approach based on K-Nearest Neighbors was used giving an overall word recognition of 74.8%.

REFERENCES

[1] Khan, K.U.; Haider, I. Online recognition of multi-stroke handwritten Urdu characters. In Proceedings of the 2010 International Conference on Image Analysis and Signal Processing, Zhejiang, China, 9–11 April 2010; pp. 284–290 https://ieeexplore.ieee.org/document/5476113?denied=

[2] Decerbo, M.; MacRostie, E.; Natarajan, P. The BBN Byblos Pashto OCR system. In Proceedings of the 1st ACM Workshop on Hardcopy Document Processing, Washington, DC, USA, 12 November 2004; Association for Computing Machinery: New York, NY, USA, 2004; pp. 29–32. https://dl.acm.org/doi/10.1145/1031442.1031447#sec-ref

[3] El-Sawy, A.; EL-Bakry, H.; Loey, M. CNN for handwritten Arabic digits recognition based on LeNet-5. In Proceedings of the Proceedings of the International Conference on Advanced Intelligent Systems and Informatics, Cairo, Egypt, 24–26 October 2016; Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F., Eds.; Springer: Cham, Switzerland, 2017; pp. 566–575. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9874742

[4]Abandah, G.A.; Younis, K.; Khedher, M.Z. Handwritten Arabic character recognition using multiple classifiers based on letter form. In Proceedings of the 5th International Conference on Signal Processing, Pattern Recognition, and Applications (SPPRA), Innsbruck, Austria, 13–15 February 2008 https://dl.acm.org/doi/10.1145/1031442.1031447#sec-ref

[5]Ahmad, N.; Khan, A.A.; Abid, S.a.R.; Yasir, M.; Ullah, N. Pashto Isolated Character Recognition Using K-Nn ClassifieR. Sindh Univ. Res. J. SURJ 2013, 45, 679–682 https://www.researchgate.net/publication/260981290_Pashto_Isolated_Character_Recognition_Using_K-NN_Classifier

REFERENCES

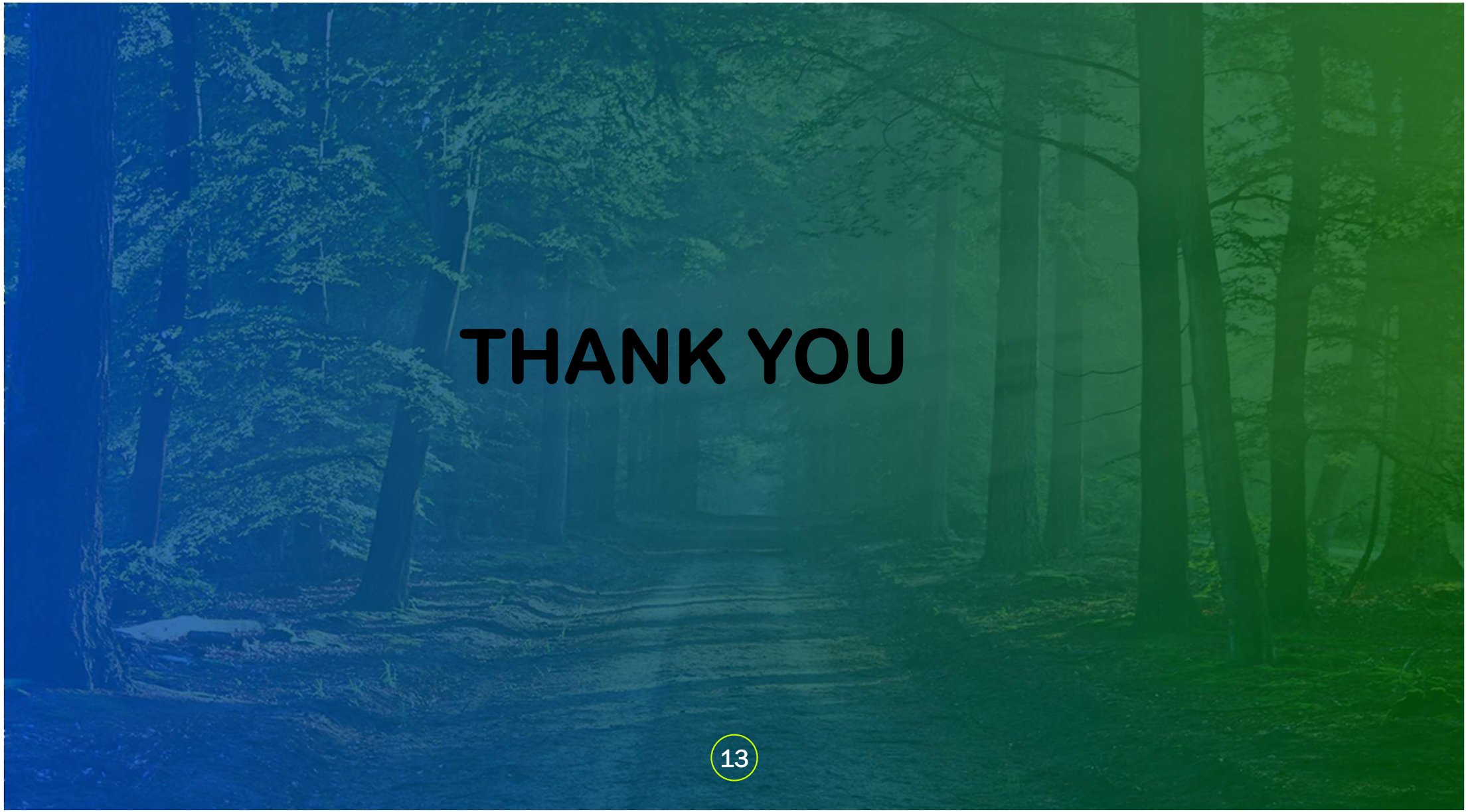
[6] S. Mori, C. Y. Suen, and K. Yamamoto, ‘‘Historical review of OCR research and development,’’ Proc. IEEE, vol. 80, no. 7, pp. 1029–1058, Jul. 1992.
https://ieeexplore.ieee.org/document/156468

[7] L. M. Lorigo and V. Govindaraju, ‘‘Offline Arabic handwriting recognition: A survey,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 5, pp. 712–724, May 2006.
https://ieeexplore.ieee.org/document/1608035

[8] H. S. Baird, H. Bunke, and K. Yamamoto, Structured Document Image Analysis. Berlin, Germany: Springer, 2012.
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9928191

[9] J. Memon, M. Sami, R. A. Khan, and M. Uddin, ‘‘Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR),’’ IEEE Access, vol. 8, pp. 142642–142668, 2020.
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9151144

[10] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez, ‘‘KNN and ANN-based recognition of handwritten pashto letters using zoning features,’’ 2019, arXiv:1904.03391.
https://thesai.org/Publications/ViewPaper?Volume=9&Issue=10&Code=ijacsa&SerialNo=69


THANK YOU