4 May 2022

IIIT KOTA

DR. BASANT AGARWAL

STUDENT

Ayush Kumar Mohanta
2019KUCP1033
Rishabh Chouhan
2019KUCP1071
Raghav Jajoo
2019KUCP1082

HINDI OCR

Problem Statement

OCR for Offline Hindi Handwritten Text

What is OCR?

Optical Character Recognition (OCR) is a technology that recognizes text within a digital image.

It is commonly used to recognize text in scanned documents and images. OCR software can be used to convert a physical paper document, or an image into an accessible electronic version with text.

II USE CASE

Digitizing of Official Documents

Converting official handwritten documents and old manuscripts and storing it digitally to ease access.

Handwriting Recognition

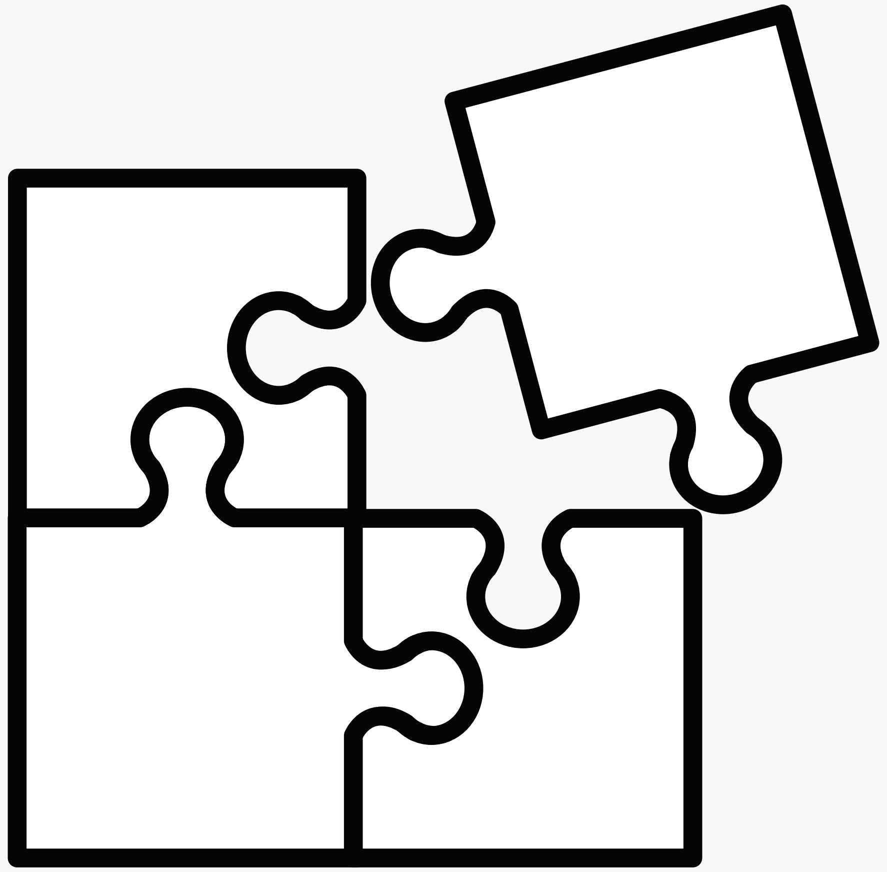
To recognize the person the document was written by by comparing his/her handwriting
.

Handwriting Evaluatioin

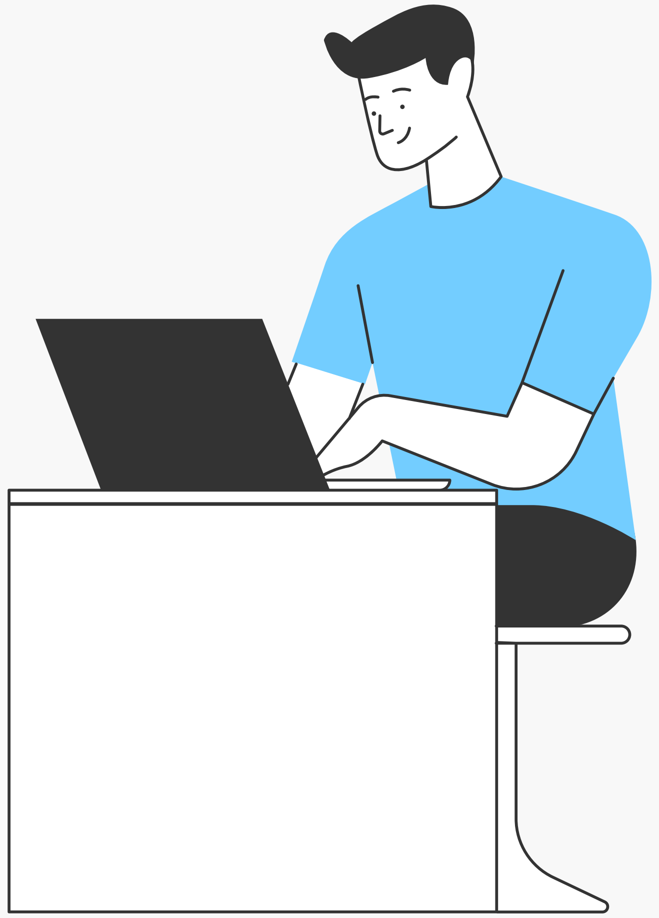
Evaluating handwriting for handwriting competitions based on the accuracy of the character.

III CHALLENGES

- Different Writing and Font styles.
- Non Standard way of writing.
- Confusion among similar characters.
- Shortage of DataSet.



IV PREVIOUS WORK



- Complete OCR for English Handwritten Documents.
- Hindi Handwritten OCRs have been made for character level with acceptable Character Error Rate (CER).
- OCR available for other Indian scripts such as Telgu & Bangla.
- OCR for Hindi Typed documents has been done with acceptable accuracy.

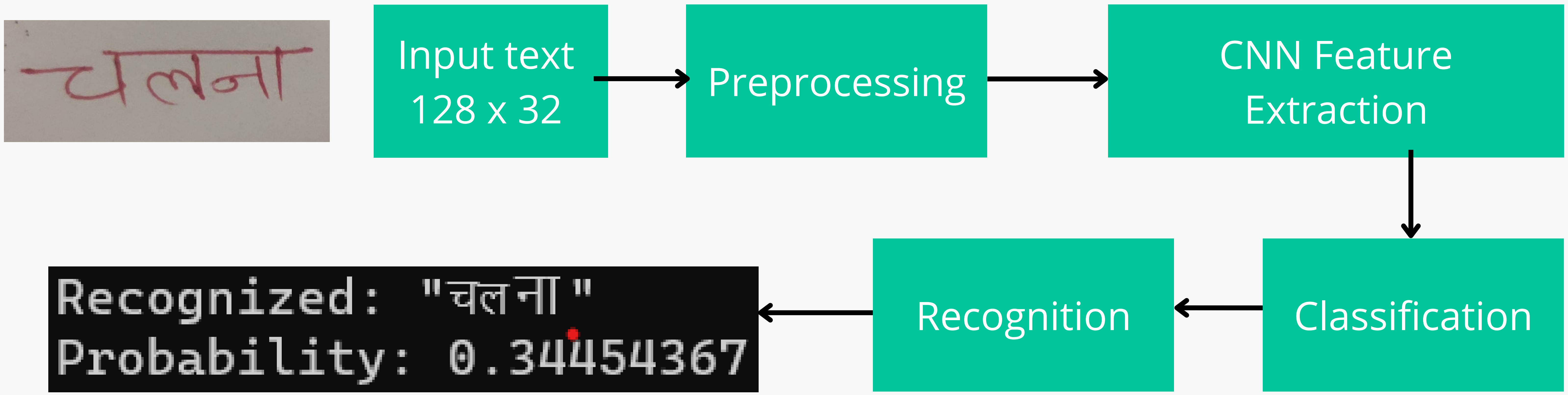
V PREVIOUS WORK & APPROACHES

Model	CER
CNN-LSTM	9.6%
SVM	20-30%
KNN (TYPED)	7%

Our Previous Work:

OCR of character level.
Planned to convert it into a word level OCR but failed due to availability of dataset of all characters and punctuations.

VI APPROACH



```
Recognized: "चलना"
Probability: 0.34454367
```

VII DataSet

Word level Handwritten datasets for Indic scripts

- A Devanagari dataset comprising of over 95K handwritten words.
- Datasets contain word images only and these images are in jpg format.
- CVIT IIIT Word Dataset.

Drawback of the dataset.

Only 12 Writers so less variation in data.

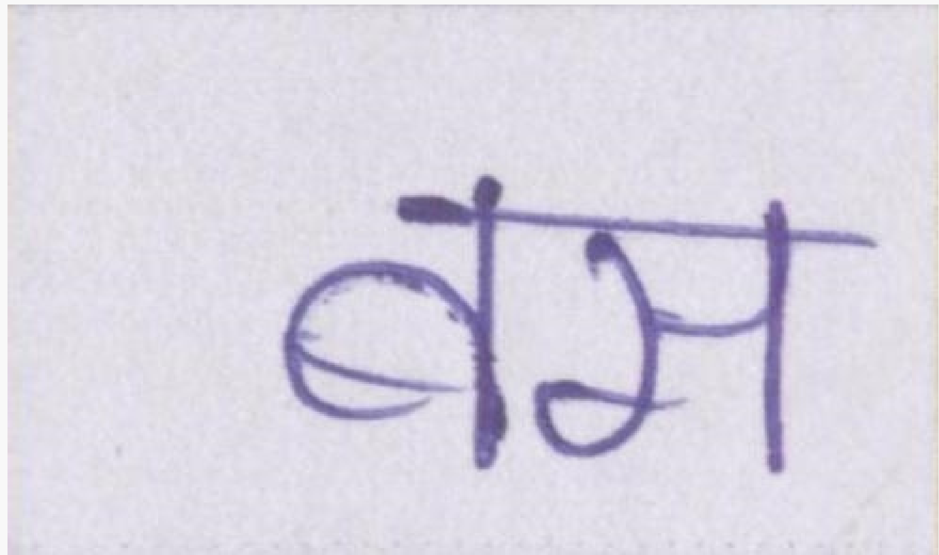
VIII Model

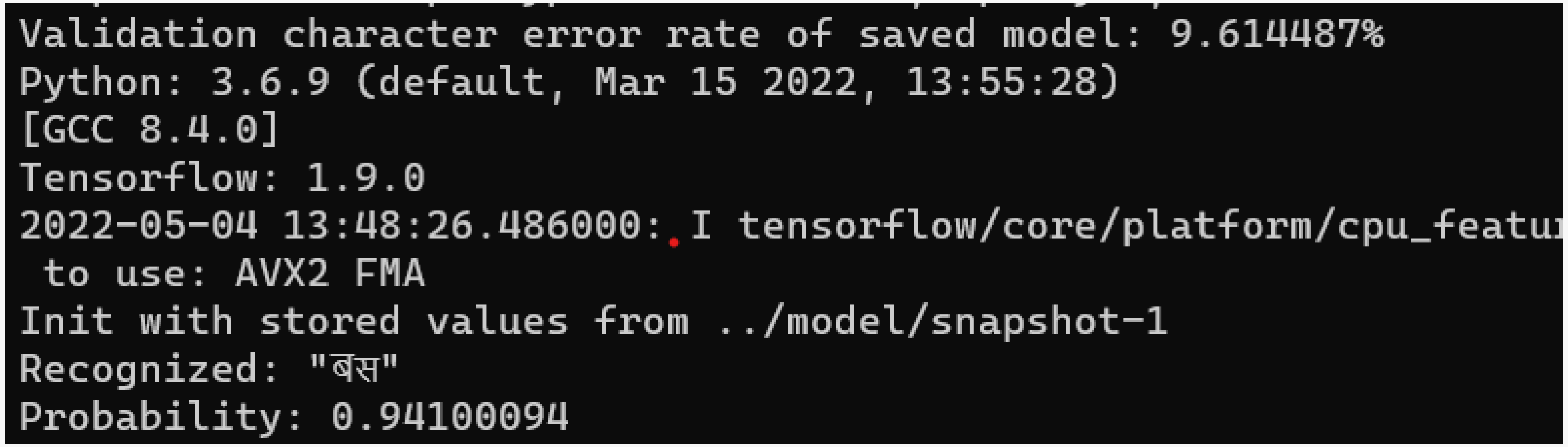
- Input image is resized, converted to grayscale in preprocessing and passed to the model in batch sizes.
- 5 Layer of CNN which extracts 256 feature maps of 32x1.
- Extracted feature vectors are passed through a 2 layer of LSTM form 1 BLSTM to maintain text dependencies form both directions.
- CTC Loss function is used to train the LSTM without need of transcription provided the correct spare ground truth texts and numeric value.
- Finally CTC decode is used on to decode the LSTM output and predict the written text in digital Image.

IX Model Accuracy

MODEL	WER	CER
CNN-BLSTM	**37.92**	**9.1**
SCNN-BLSTM	34.52	7.83
IAM-SCNN-BLSTM (Telegu)	23.98	4.58

X Result





```
Validation character error rate of saved model: 9.614487%
Python: 3.6.9 (default, Mar 15 2022, 13:55:28)
[GCC 8.4.0]
Tensorflow: 1.9.0
2022-05-04 13:48:26.486000: I tensorflow/core/platform/cpu_featu
 to use: AVX2 FMA
Init with stored values from ../model/snapshot-1
Recognized: "बस"
Probability: 0.94100094
```

XI Future Work & Further Improvements

- Use Data Augmentation before training to improve accuracy.

- Convert it into a full paragraph reading model with a few more layers added on it.

XII References

- [1] D. Yadav, S. Sanchez-Cuadrado, and J. Morato, “Optical Character Recognition for Hindi Language Using a Neural-network Approach,” Journal of Information Processing Systems, vol. 9, no. 1, pp. 117–140, Mar. 2013.

- K. Dutta, P. Krishnan, M. Mathew and C. V. Jawahar, "Towards Spotting and Recognition of Handwritten Words in Indic Scripts," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 32-37, doi: 10.1109/ICFHR-2018.2018.00015.

- Nafiz Arica , Student Member , Fatos T. Yarman-vural , Senior Member, "Optical Character Recognition for Cursive Handwriting", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 6, JUNE 2002

XII References

- Gunna, S., Saluja, R., Jawahar, C.V. (2021). Transfer Learning for Scene Text Recognition in Indian Languages. In: Barney Smith, E.H., Pal, U. (eds) Document Analysis and Recognition – ICDAR 2021 Workshops. ICDAR 2021. Lecture Notes in Computer Science(), vol 12916. Springer, Cham.

- Abhishek Mehta, Dr. Subhashchdra Desai, Dr. Ashish Chaturvedi, 2021, Hindi Handwritten Character Recognition from Digital Image using Deep Learning Neural Network, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICRADL – 2021 (Volume 09 – Issue 05)

- Nisha Goyal1 , Er. Shilpa Jain, "Optimized Hindi Script Recognition using OCR Feature Extraction Technique", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 8, August 2015.

XII References

- Bairagi, Prasanta Pratim. “Optical Character Recognition for Hindi.” (2018).

Thank you!