Optical Character Recognition of Jutakshars within Devanagari Script

Sheallika Singh Shreesh Ladha
Supervised by : Dr. Harish Karnick, Dr. Amit Mitra

UGP Presentation, 10 April 2016

Basic working of OCR

- Convert printed/scanned text into editable text

Basic working of OCR

- Convert printed/scanned text into editable text
- Starts off with some basic preprocessing of the image for uniformity

Basic working of OCR

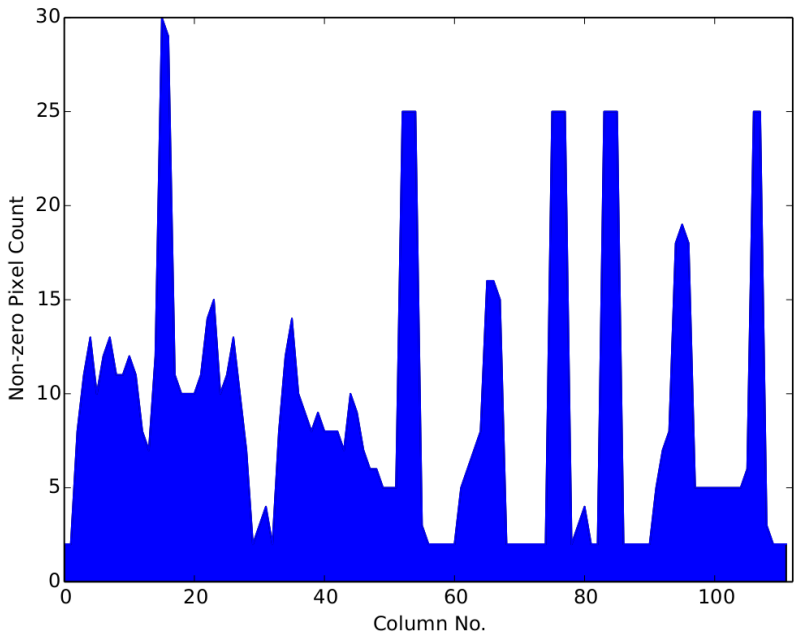
- Convert printed/scanned text into editable text
- Starts off with some basic preprocessing of the image for uniformity
- Segmentation of lines, words, characters based on their horizontal and vertical histograms of pixel intensities

Basic working of OCR

- Convert printed/scanned text into editable text
- Starts off with some basic preprocessing of the image for uniformity
- Segmentation of lines, words, characters based on their horizontal and vertical histograms of pixel intensities
- Classifiers are then used for prediction

Basic working of OCR

- Convert printed/scanned text into editable text
- Starts off with some basic preprocessing of the image for uniformity
- Segmentation of lines, words, characters based on their horizontal and vertical histograms of pixel intensities
- Classifiers are then used for prediction
- Composition using language rules and n-gram models



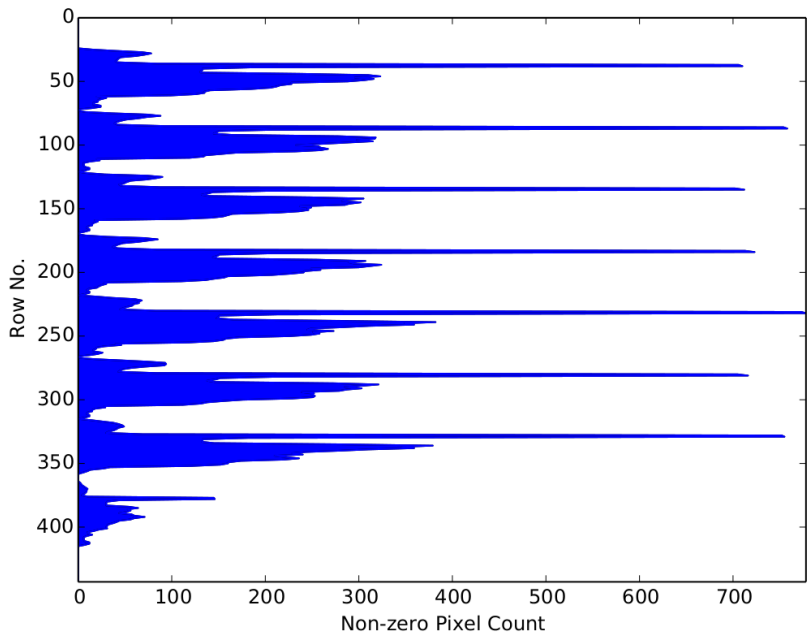


Figure: Horizontal and vertical histograms

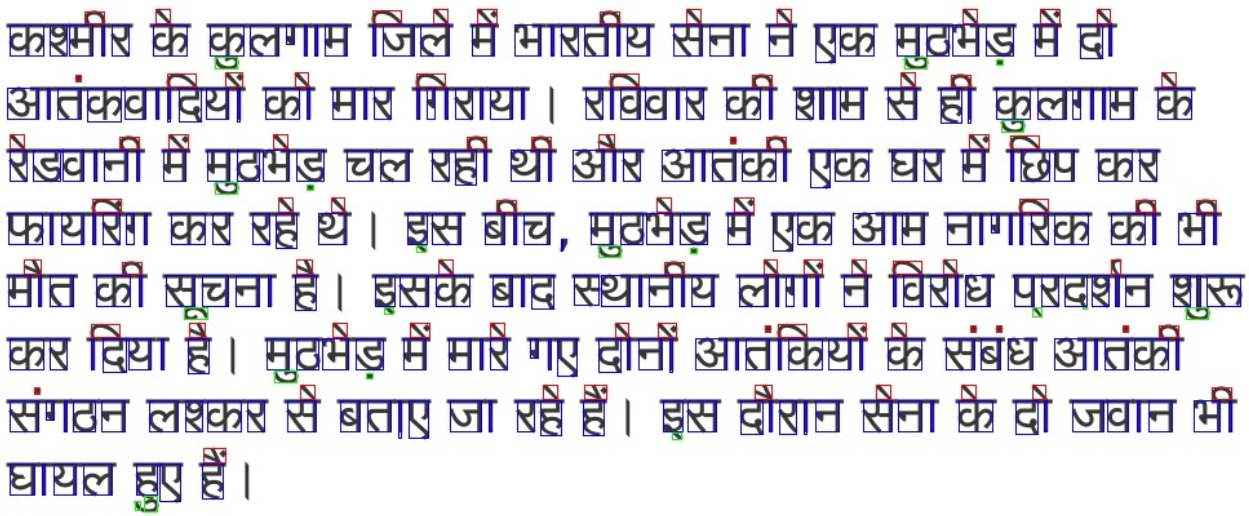


Figure: Example of Segmentatin

Challenges with Devanagari

- Robust OCR systems already in place for Roman Scripts

Challenges with Devanagari

- Robust OCR systems already in place for Roman Scripts
- Lack of such software for Hindi Languages because of highly complicated structure and composition

Challenges with Devanagari

- Robust OCR systems already in place for Roman Scripts
- Lack of such software for Hindi Languages because of highly complicated structure and composition
- An eg. of such composition is a form of character conjunction referred to as “Jutakshar”

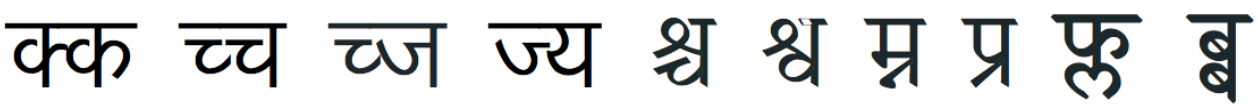


Figure: Classes of Jutakshars. *Note the model was developed to understand Jutakshars similar in type with the first four

Our Work

- Words containing Jutakshar are quite popular and can be found in many document

Our Work

- Words containing Jutakshar are quite popular and can be found in many document
- Most OCR systems do not take care to handle such characters, which in turn reduces their precision

Our Work

- Words containing Jutakshar are quite popular and can be found in many document
- Most OCR systems do not take care to handle such characters, which in turn reduces their precision
- Low accuracy due to Jutakshars being recognized as a single character

Our Work

- Words containing Jutakshar are quite popular and can be found in many document
- Most OCR systems do not take care to handle such characters, which in turn reduces their precision
- Low accuracy due to Jutakshars being recognized as a single character
- We have tried to build upon existing frameworks to better handle these special strings

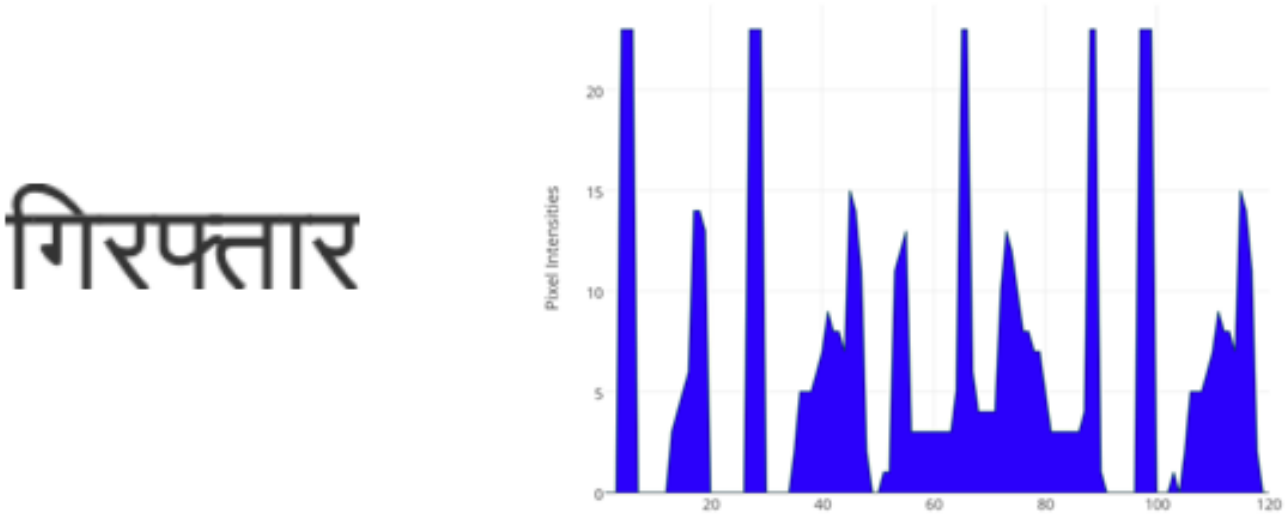


Figure: Where segmentation fails

Approach

- Jutakshar Detection :
  Figure out whether a character is a jutakshar or not

Approach

- Jutakshar Detection :
  Figure out whether a character is a jutakshar or not
- Jutakshar Identification :
  Figure out the jutakshar and in turn the entire word

Jutakshar Detection

- Created dataset of jutakshars using 8 fonts and added gaussian noise

Jutakshar Detection

- Created dataset of jutakshars using 8 fonts and added gaussian noise
- Trained classifiers to predict whether a given character was a Jutakshar or not

Jutakshar Detection

- Created dataset of jutakshars using 8 fonts and added gaussian noise
- Trained classifiers to predict whether a given character was a Jutakshar or not
- Experimented with three different types of models

Jutakshar Detection

- Model I : SVM trained using hog features

Jutakshar Detection

- Model I : SVM trained using hog features
- Model II : Logistic regression using features extracted from penultimate layer of a pretrained convolutional neural network

Jutakshar Detection

- Model I : SVM trained using hog features
- Model II : Logistic regression using features extracted from penultimate layer of a pretrained convolutional neural network
- Model III : Classification using the width of the character boxes

अंतरराष्ट्रीय योग दिवस का सबसे बड़ा सूबा उत्तर प्रदेश रविवार को देश की इस सांस्कृतिक विरासत का महापर्व मनाने में बढ़चढ़ कर आगे रहा। सुबह से ही योग के आयोजन स्थलों पर आम से लेकर खास, बच्चे से लेकर बुजुर्ग-महिलाएं और सीआरपीएफ व सेना के जवान सभी पूरे उत्साह के साथ योग में रमे। राजधानी में सबसे बड़े आयोजन स्थल केडी सिंह बाबू स्टेडियम में केंद्रीय गृहमंत्री राजनाथ सिंह ने भी कई आसन किए। इस दौरान मुस्लिम धर्म के लोग भी योग में शामिल हुए। राजनाथ सिंह ने ने किसी का नाम लिए बगैर कहा कि योग भारत की सांस्कृतिक विरासत है। इसे जाति, मजहब व धर्म की सीमा में नहीं बांधा जा सकता।

अंतरराष्ट्रीय योग दिवस का सबसे बड़ा सूबा उत्तर प्रदेश रविवार को देश की इस सांस्कृतिक विरासत का महापर्व मनाने में बढ़चढ़ कर आगे रहा। सुबह से ही योग के आयोजन स्थलों पर आम से लेकर खास, बच्चे से लेकर बुजुर्ग-महिलाएं और सीआरपीएफ व सेना के जवान सभी पूरे उत्साह के साथ योग में रमे। राजधानी में सबसे बड़े आयोजन स्थल केडी सिंह बाबू स्टेडियम में केंद्रीय गृहमंत्री राजनाथ सिंह ने भी कई आसन किए। इस दौरान मुस्लिम धर्म के लोग भी योग में शामिल हुए। राजनाथ सिंह ने ने किसी का नाम लिए बगैर कहा कि योग भारत की सांस्कृतिक विरासत है। इसे जाति, मजहब व धर्म की सीमा में नहीं बांधा जा सकता।

अंतरराष्ट्रीय योग दिवस का सबसे बड़ा सूबा उत्तर प्रदेश रविवार को देश की इस सांस्कृतिक विरासत का महापर्व मनाने में बढ़चढ़ कर आगे रहा। सुबह से ही योग के आयोजन स्थलों पर आम से लेकर खास, बच्चे से लेकर बुजुर्ग-महिलाएं और सीआरपीएफ व सेना के जवान सभी पूरे उत्साह के साथ योग में रमे। राजधानी में सबसे बड़े आयोजन स्थल केडी सिंह बाबू स्टेडियम में केंद्रीय गृहमंत्री राजनाथ सिंह ने भी कई आसन किए। इस दौरान मुस्लिम धर्म के लोग भी योग में शामिल हुए। राजनाथ सिंह ने ने किसी का नाम लिए बगैर कहा कि योग भारत की सांस्कृतिक विरासत है। इसे जाति, मजहब व धर्म की सीमा में नहीं बांधा जा सकता।

Jutakshar Identification

- Using the previous model, the character boxes predicted as containing jutakshars, are now split into the corresponding half and full character

Jutakshar Identification

- Using the previous model, the character boxes predicted as containing jutakshars, are now split into the corresponding half and full character
- Used vertical histograms for splitting but restricted our search in the middle part of the character

Jutakshar Identification

- Using the previous model, the character boxes predicted as containing jutakshars, are now split into the corresponding half and full character
- Used vertical histograms for splitting but restricted our search in the middle part of the character
- The position of the minima was used to split the the Jutakshar into the half and the full character forming it

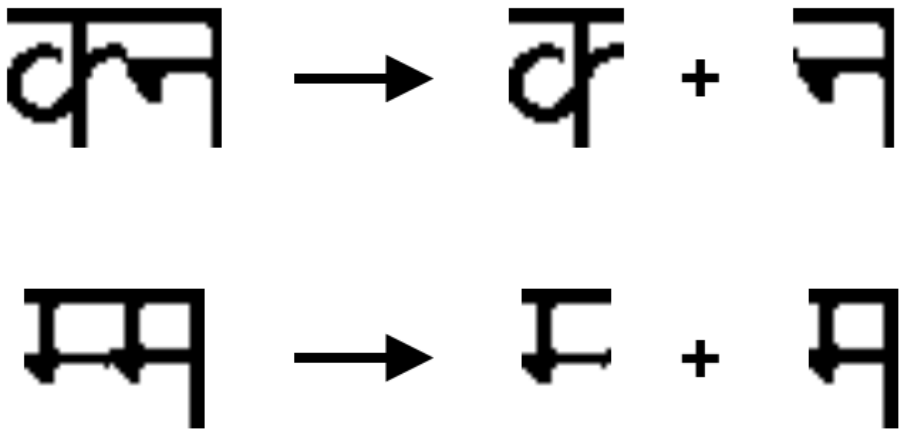


Figure: Example showing how splitting works on Jutakshars

Jutakshar Identification

- Segmented word is passed onto the character classifiers to predict the respective characters

Jutakshar Identification

- Segmented word is passed onto the character classifiers to predict the respective characters
- The predicted word is passed to a dictionary which gives a list of suggested correct words for an incorrect word

Jutakshar Identification

- Segmented word is passed onto the character classifiers to predict the respective characters
- The predicted word is passed to a dictionary which gives a list of suggested correct words for an incorrect word
- A heuristic is defined to find the correct prediction from the suggested list

Jutakshar Identification

- Suggested list of words refined to have only words which contain Jutskshars and having length as close as possible

Jutakshar Identification

- Suggested list of words refined to have only words which contain Jutskshars and having length as close as possible
- Levenshtein distance (computes the minimum number of steps required for converting one string into another by insertions, deletions or substitutions of single characters) between the predicted and the suggested word

Jutakshar Identification

- Suggested list of words refined to have only words which contain Jutskshars and having length as close as possible
- Levenshtein distance (computes the minimum number of steps required for converting one string into another by insertions, deletions or substitutions of single characters) between the predicted and the suggested word
- Number of substrings that match in the predicted word with the jutakshars removed and the suggested word

Jutakshar Identification

- Suggested list of words refined to have only words which contain Jutskshars and having length as close as possible
- Levenshtein distance (computes the minimum number of steps required for converting one string into another by insertions, deletions or substitutions of single characters) between the predicted and the suggested word
- Number of substrings that match in the predicted word with the jutakshars removed and the suggested word
- Length of the longest common substring between the predicted word with the jutakshars removed and the suggested word in the list

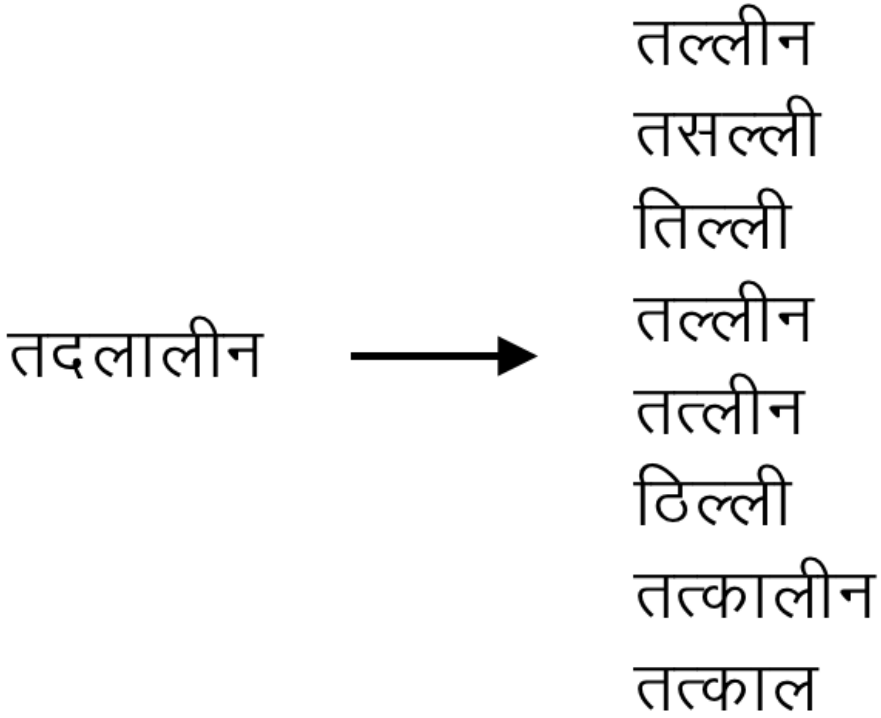


Figure: Example of suggestions returned by the dictionary

Figure: Correction in the predicted word after applying our heuristic

Results for Jutakshar Detection Model

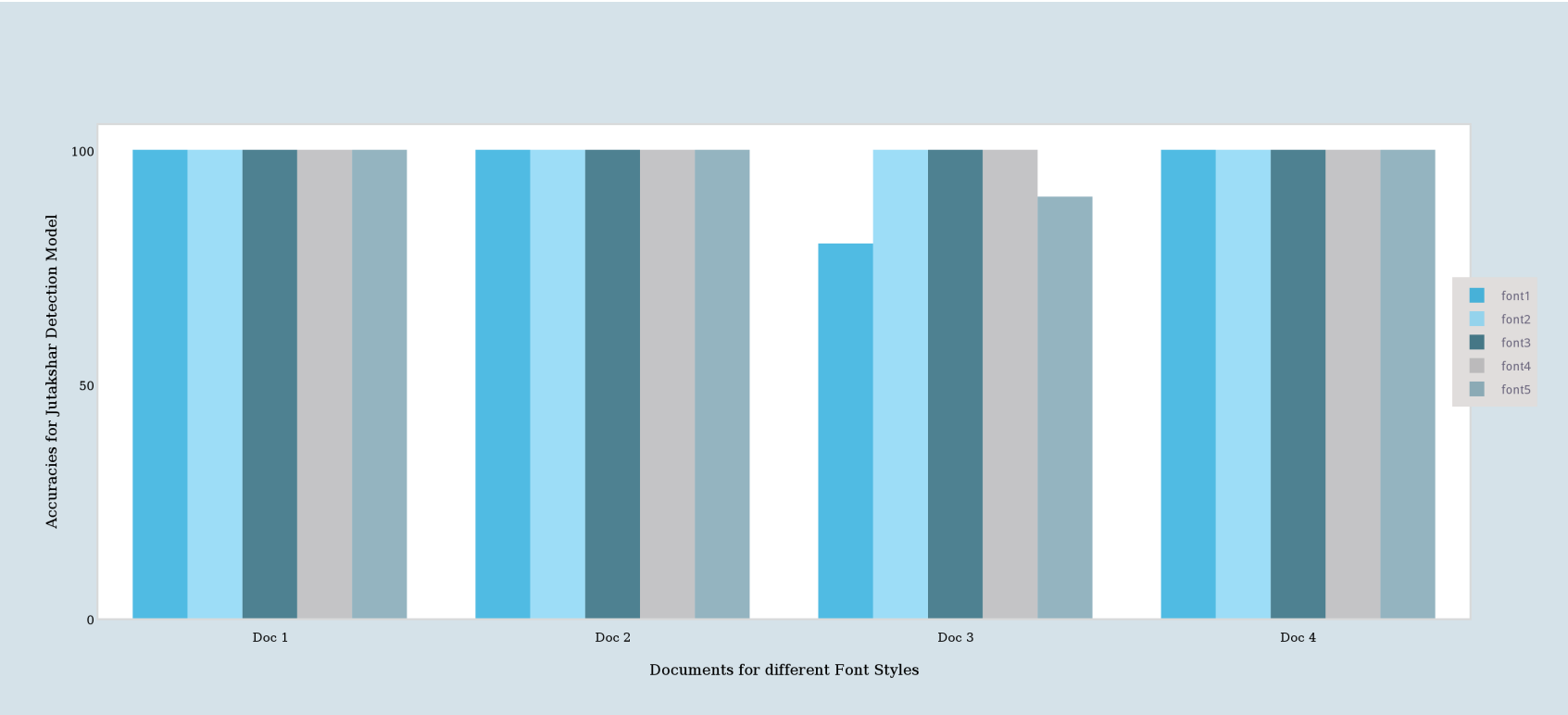


Figure: Plot of accuracies of Jutakshar detection model on different fonts and documents

Results : Jutakshar Detection

Document	Total Words	# Words containing Jutakshars	# True Positives	# False Positives
Doc 1	166	6	6	0
Doc 2	359	20	20	0
Doc 3	131	5	4	1
Doc 4	138	9	9	0

Table: Jutakshar Detection (Font 1)

Results : Jutakshar Detection

Document	Total Words	# Words containing Jutakshars	# True Positives	# False Positives
Doc 1	145	6	6	0
Doc 2	218	4	4	0
Doc 3	138	4	4	0
Doc 4	167	7	7	2

Table: Jutakshar Detection (Font 2)

Results : Jutakshar Detection

Document	Total Words	# Words containing Jutakshars	# True Positives	# False Positives
Doc 1	98	5	5	1
Doc 2	171	10	10	1
Doc 3	243	10	9	0
Doc 4	210	7	7	0

Table: Jutakshar Detection (Font 3)

Results : Jutakshar Detection

Document	Total Words	# Words containing Jutakshars	# True Positives	# False Positives
Doc 1	72	3	3	1
Doc 2	99	3	3	1
Doc 3	114	8	8	1
Doc 4	133	5	5	0

Table: Jutakshar Detection (Font 4)

Results : Jutakshar Detection

Document	Total Words	# Words containing Jutakshars	# True Positives	# False Positives
Doc 1	246	10	10	3
Doc 2	110	3	3	1
Doc 3	89	2	2	2
Doc 4	129	9	9	2

Table: Jutakshar Detection (Font 5)

Results : Jutakshar Detection

False Positive type of cases

OCR Evaluation

Word : सहजन

Output : असहज

Accuracy : 0% or 50% ?

Error rates

- WER: #Erroneous words divided by total word count in ground-truth text
- CER: #Erroneous characters divided by total character count in ground-truth text
- Evaluation tool: `https://github.com/impactcentre/ocrevalUAtion`

Error rates

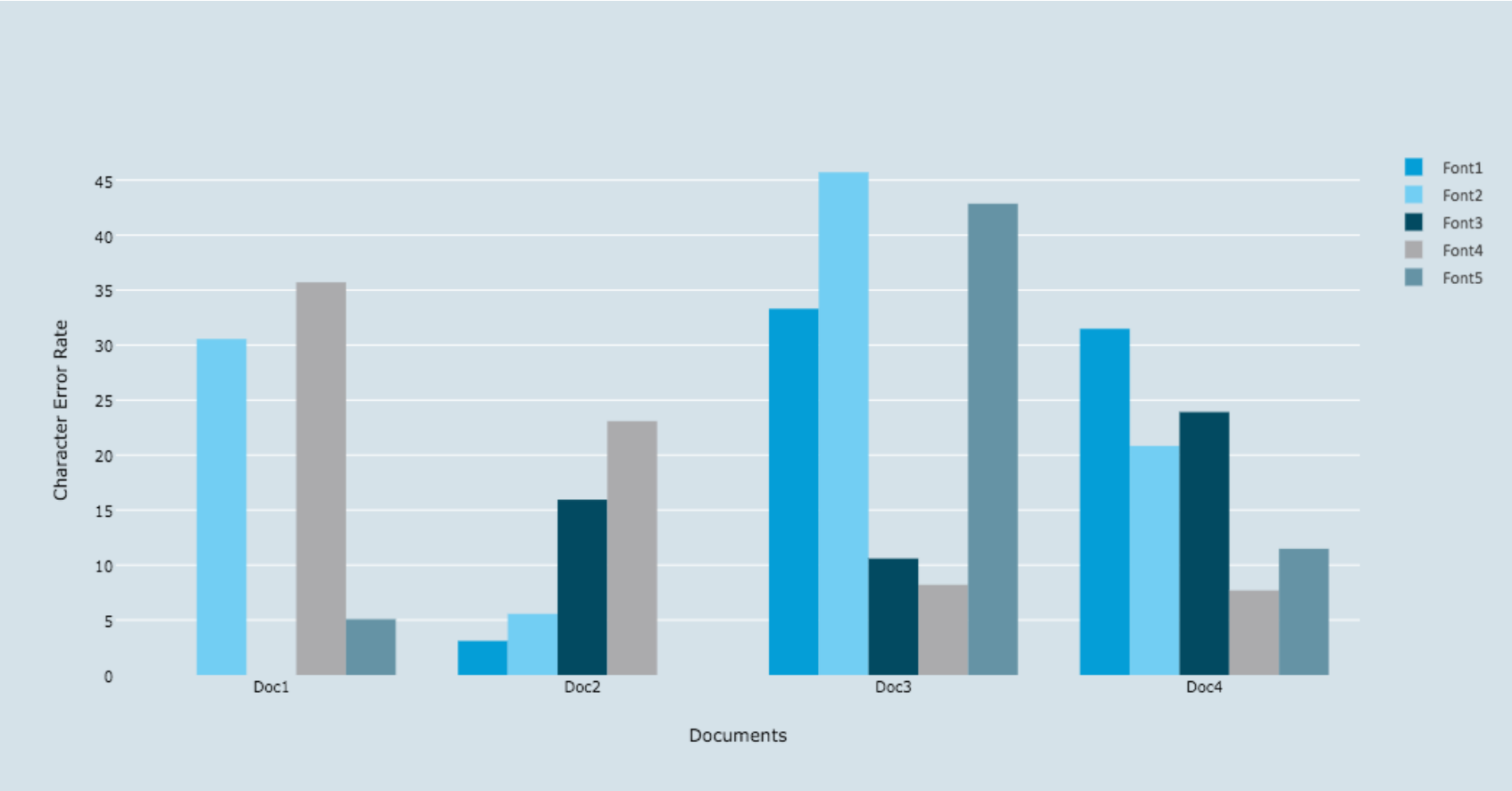
- WER: #Erroneous words divided by total word count in ground-truth text
- CER: #Erroneous characters divided by total character count in ground-truth text
- Evaluation tool: `https://github.com/impactcentre/ocrevalUAtion`

Mathematically,

$$WER(\%) = \frac{(S_w + I_w + D_w) \times 100}{N_w}$$

$$CER(\%) = \frac{(S_c + I_c + D_c) \times 100}{N_c}$$

Results : Jutakshar Identification



Plot for Character Error rates with which word containing jutakshars are identified

Overall Accuracies

Document	CER(%) (Prev Model)	WER(%) (Prev Model)	CER(%) (Our Model)	WER(%) (Our Model)
Doc 1	6.8	16.5	6.7	8.6
Doc 2	8.2	19.8	6.3	8.5
Doc 3	8.1	19.0	7.6	8.8
Doc 4	9.8	22.7	9.5	12.1
Overall	8.2	19.6	7.5	9.5

Table: OCR Accuracy : Comparing Models (Font 1)

Overall Accuracies

Document	CER(%) (Prev Model)	WER(%) (Prev Model)	CER(%) (Our Model)	WER(%) (Our Model)
Doc 1	13.5	29.1	17.6	25.4
Doc 2	8.9	20.3	9.9	13.0
Doc 3	11.7	31.2	15.7	19.6
Doc 4	6.9	15.9	6.9	11.8
Overall	9.4	23.6	12.5	17.4

Table: OCR Accuracy : Comparing Models (Font 2)

Overall Accuracies

Document	CER(%) (Prev Model)	WER(%) (Prev Model)	CER(%) (Our Model)	WER(%) (Our Model)
Doc 1	9.3	20.8	10.0	6.9
Doc 2	10.4	22.1	11.8	16.8
Doc 3	8.2	19.7	8.7	12.8
Doc 4	9.8	23.3	10.2	11.8
Overall	9.3	21.5	7.8	12.1

Table: OCR Accuracy : Comparing Models (Font 3)

Overall Accuracies

Document	CER(%) (Prev Model)	WER(%) (Prev Model)	CER(%) (Our Model)	WER(%) (Our Model)
Doc 1	8.8	18.4	13.6	17.6
Doc 2	9.6	18.7	10.1	14.3
Doc 3	6.3	16.6	7.6	13.3
Doc 4	7.1	16.9	9.3	13.1
Overall	7.8	17.5	10.1	14.6

Table: OCR Accuracy : Comparing Models (Font 4)

Overall Accuracies

Document	CER(%) (Prev Model)	WER(%) (Prev Model)	CER(%) (Our Model)	WER(%) (Our Model)
Doc 1	6.3	16.0	5.6	8.1
Doc 2	4.5	9.8	5.5	7.5
Doc 3	6.7	17.9	7.8	12.2
Doc 4	8.3	17.3	8.1	7.9
Overall	6.5	15.4	6.7	8.9

Table: OCR Accuracy : Comparing Models (Font 5)

Future Work

- Our model relies heavily on the dictionary. Hence a more powerful dictionary would reduce quite a few errors

Future Work

- Our model relies heavily on the dictionary. Hence a more powerful dictionary would reduce quite a few errors
- The dataset can be made more representative by adding more fonts and possible jutakshars

Future Work

- Our model relies heavily on the dictionary. Hence a more powerful dictionary would reduce quite a few errors
- The dataset can be made more representative by adding more fonts and possible jutakshars
- Two deepnet models could be trained - One for all half characters and the other for all different types of full characters

Questions?