Project Phase III

OCR
Offline Malayalam Character Recognition

TEAM 12

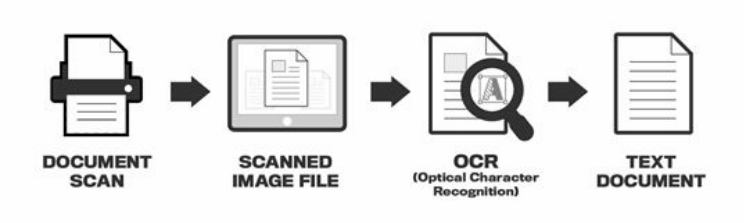
George Zachariah V	30
Harisankar M	33
Nikhil Narayanan	47
Aromal S	81

Guide : Dr. Ansamma John

TABLE OF CONTENTS

Introduction

- OCR - The electronic conversion of printed or handwritten text into a computer-readable form



- The handwritten script recognition is one of the most interesting and challenging areas of pattern recognition due to numerous variations in writing styles

- The character recognition system operates with an aim to replicate human reading ability by maintaining accuracy at a far higher speed.

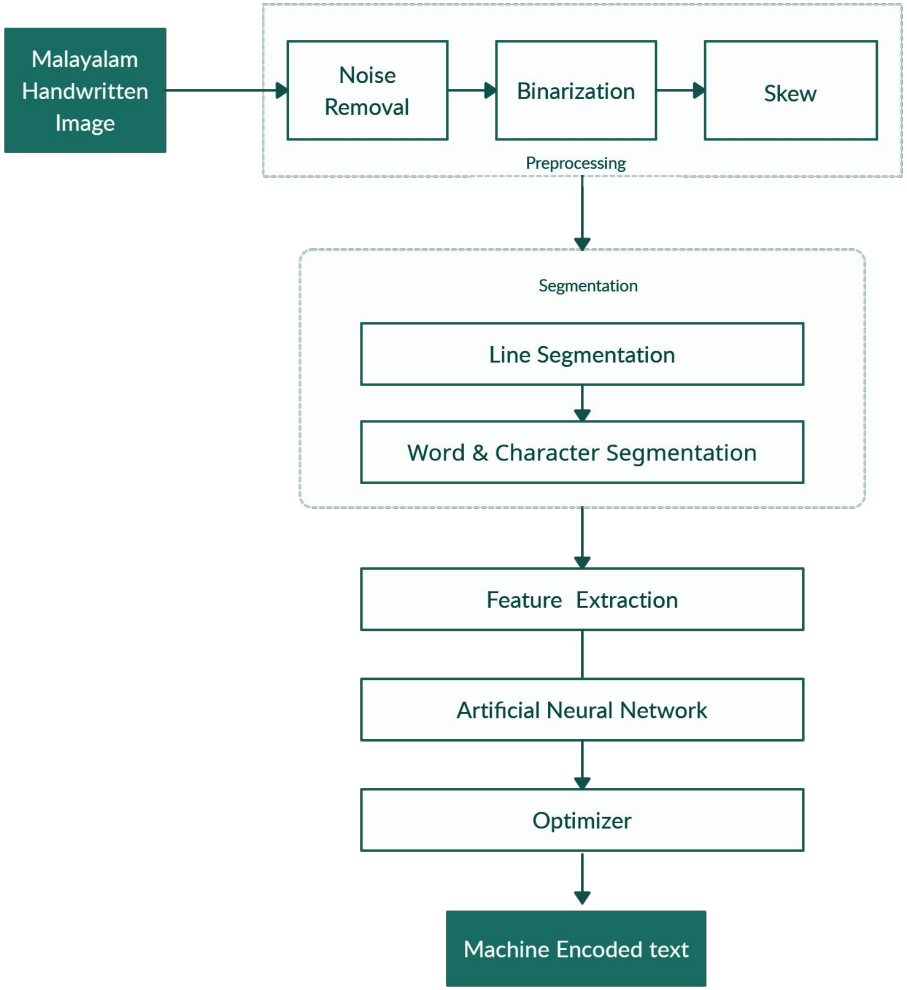
Motivation

- Demand for a robust and reliable character recognition system for Indic scripts
- Similarly shaped characters makes recognition quite difficult and challenging.
- Recognition of degraded historical and ancient archives will help several memory institutions to digitize their manuscript collection.
- Lack of benchmark and standard datasets for malayalam character recognition
- Existing systems do not use features like linguistic information for post-recognition error detection and correction

Objectives

- Implement a robust and reliable character recognition for malayalam language
- Develop an efficient method to segment a document into lines, words, and characters while preserving their order.
- Use linguistic information or script specific knowledge for post recognition error correction and detection

Architecture



MODULES

1. Preprocessing
2. Segmentation
3. Feature Extraction
4. Character Recognition

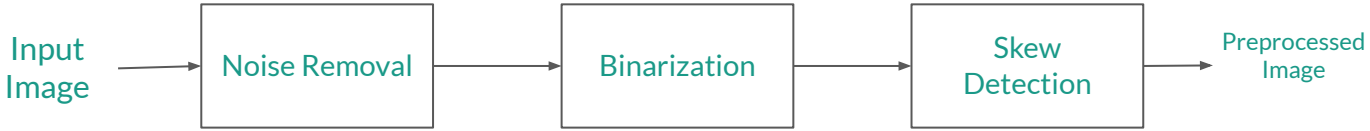
MODULE 1

PREPROCESSING

Preprocessing

Objective

To improve the image data to suppress unwilling distortions and noise or enhances some image features important for further processing



Noise Removal	:	Median Filter 2D + Morphological Operation
Binarization	:	Otsu thresholding
Skew Detection & Correction	:	Global Skew Detection and correction using WarpAffine Transformation

Comparison of Thresholding Algorithms

	2peaks	bernsen	bradleyRoth	entropyJohannsen	entropyKapur	entropyPun	feng
ssim	0.828676	0.709016	0.846650215	0.632062541	0.854760883	0.318639522	0.852983
psnr	12.61276	11.44084	13.70623932	9.016634775	13.98083613	4.418923374	15.17984
mse	3701.76	5500.771	2965.314002	16228.88234	2915.994704	24024.23281	2104.569
nrms	0.251285	0.295899	0.224245578	0.445436427	0.219872022	0.640489833	0.1891

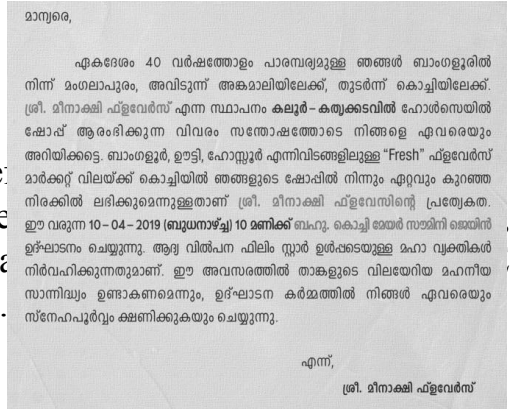
	localContrast	localMean	minError	niblack	nick	Otsu	tsuMultiT	p_tile	sauvola	singh	wolf
ssim	0.33550492	0.2272807	0.653836	0.269152	0.821632	0.791191	0.642306	0.330314	0.800043	0.746866	0.858109
psnr	6.301919222	5.2843094	9.090859	6.334507	12.35243	12.27723	11.8408	4.509033	11.35811	12.50872	13.86742
mse	16063.95273	19855.42	10771.98	15597.4	4042.68	5412.996	4723.003	23366.74	5074.176	4784.987	2955.074
nrms	0.518661915	0.580192	0.404838	0.514304	0.261927	0.282021	0.279522	0.63241	0.293524	0.269201	0.221673

The performance of 18 thresholding algorithms for DIBCO Dataset

Feng thresholding

- An improved method for binarizing document images by adaptively exploiting the local image contrast.

- Overcomes the common problems encountered in low quality images, such as uneven illumination, low contrast, and random noise

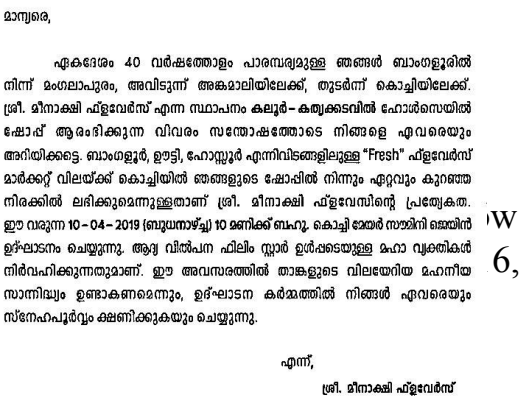


മാന്യരെ,

ഏകദേശം 40 വർഷത്തോളം പാരമ്പര്യമുള്ള ഞങ്ങൾ ബാംഗളൂരിൽ നിന്ന് മംഗലാപുരം, അവിടുന്ന് അങ്കമാലിയിലേക്ക്, തുടർന്ന് കൊച്ചിയിലേക്ക്. ശ്രീ. മീനാക്ഷി ഫ്ളവേർസ് എന്ന സ്ഥാപനം **കലൂർ – കത്രിക്കടവിൽ** ഹോൾസെയിൽ ഷോപ് ആരംഭിക്കുന്ന വിവരം സന്തോഷത്തോടെ നിങ്ങളെ എവരെയും അറിയിക്കട്ടെ. ബാംഗളൂർ, ഊട്ടി, ഹോസ്സൂർ എന്നിവിടങ്ങളിലുള്ള "Fresh" ഫ്ളവേർസ് മാർക്കറ്റ് വിലയ്ക്ക് കൊച്ചിയിൽ ഞങ്ങളുടെ ഷോപ്പിൽ നിന്നും ഏറ്റവും കുറഞ്ഞ നിരക്കിൽ ലഭിക്കുമെന്നുള്ളതാണ് ശ്രീ. മീനാക്ഷി ഫ്ളവേസിന്റെ പ്രത്യേകത. ഈ വരുന്ന **10 – 04 – 2019 (ബുധനാഴ്ച) 10 മണിക്ക്** ബഹു. കൊച്ചി മേയർ സൗമിനി ജെയിൻ ഉദ്ഘാടനം ചെയ്യുന്നു. ആദ്യ വിൽപന ഫിലിം സ്റ്റാർ ഉൾപ്പടെയുള്ള മഹാ വ്യക്തികൾ നിർവഹിക്കുന്നതുമാണ്. ഈ അവസരത്തിൽ താങ്കളുടെ വിലയേറിയ മഹനീയ സാന്നിദ്ധ്യം ഉണ്ടാകണമെന്നും, ഉദ്ഘാടന കർമ്മത്തിൽ നിങ്ങൾ എവരെയും സ്നേഹപൂർവ്വം ക്ഷണിക്കുകയും ചെയ്യുന്നു.

എന്ന്,

ശ്രീ. മീനാക്ഷി ഫ്ളവേർസ്



മാന്യരെ,

ഏകദേശം 40 വർഷത്തോളം പാരമ്പര്യമുള്ള ഞങ്ങൾ ബാംഗളൂരിൽ നിന്ന് മംഗലാപുരം, അവിടുന്ന് അങ്കമാലിയിലേക്ക്, തുടർന്ന് കൊച്ചിയിലേക്ക്. ശ്രീ. മീനാക്ഷി ഫ്ളവേർസ് എന്ന സ്ഥാപനം **കലൂർ – കത്രിക്കടവിൽ** ഹോൾസെയിൽ ഷോപ് ആരംഭിക്കുന്ന വിവരം സന്തോഷത്തോടെ നിങ്ങളെ എവരെയും അറിയിക്കട്ടെ. ബാംഗളൂർ, ഊട്ടി, ഹോസ്സൂർ എന്നിവിടങ്ങളിലുള്ള "Fresh" ഫ്ളവേർസ് മാർക്കറ്റ് വിലയ്ക്ക് കൊച്ചിയിൽ ഞങ്ങളുടെ ഷോപ്പിൽ നിന്നും ഏറ്റവും കുറഞ്ഞ നിരക്കിൽ ലഭിക്കുമെന്നുള്ളതാണ് ശ്രീ. മീനാക്ഷി ഫ്ളവേസിന്റെ പ്രത്യേകത. ഈ വരുന്ന **10 – 04 – 2019 (ബുധനാഴ്ച) 10 മണിക്ക്** ബഹു. കൊച്ചി മേയർ സൗമിനി ജെയിൻ ഉദ്ഘാടനം ചെയ്യുന്നു. ആദ്യ വിൽപന ഫിലിം സ്റ്റാർ ഉൾപ്പടെയുള്ള മഹാ വ്യക്തികൾ നിർവഹിക്കുന്നതുമാണ്. ഈ അവസരത്തിൽ താങ്കളുടെ വിലയേറിയ മഹനീയ സാന്നിദ്ധ്യം ഉണ്ടാകണമെന്നും, ഉദ്ഘാടന കർമ്മത്തിൽ നിങ്ങൾ എവരെയും സ്നേഹപൂർവ്വം ക്ഷണിക്കുകയും ചെയ്യുന്നു.

എന്ന്,

ശ്രീ. മീനാക്ഷി ഫ്ളവേർസ്

MODULE 2

SEGMENTATION

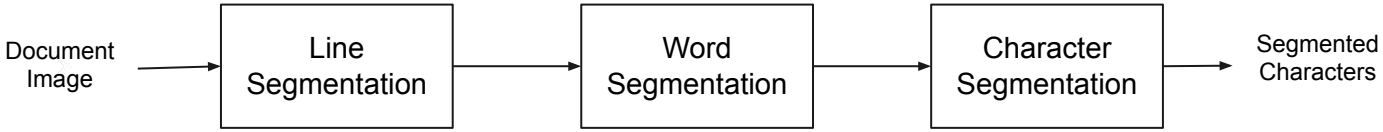
Segmentation

Objective

To extract characters from the input image of handwritten text

Input : Preprocessed Image

Output : Segmented characters



Steps in segmentation

Methods used

- Morphological Closing
- RLSA

Contour extraction with masking algorithm

Tools

OpenCV

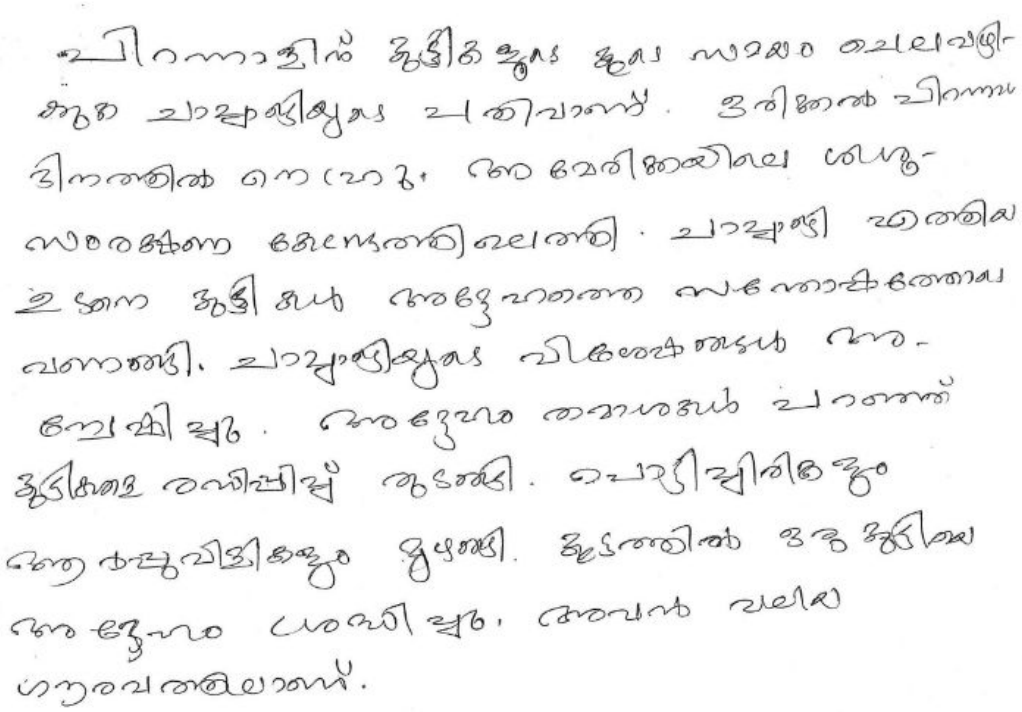
Dataset

PHD_Indic11 - Malayalam

Dataset

PHD_Indic11
(Malayalam)

- 107 page-level handwritten images



p_mal_0042.tiff

Challenges


Challenges
Overlapping regions
Touching regions
Curved lines
Classification of diacritics like chandrakala
Gaps in characters
Varying character size

Line Segmentation

Objective

To extract lines from the input image of handwritten text

Input : Preprocessed Image
Output : Segmented lines

Method used

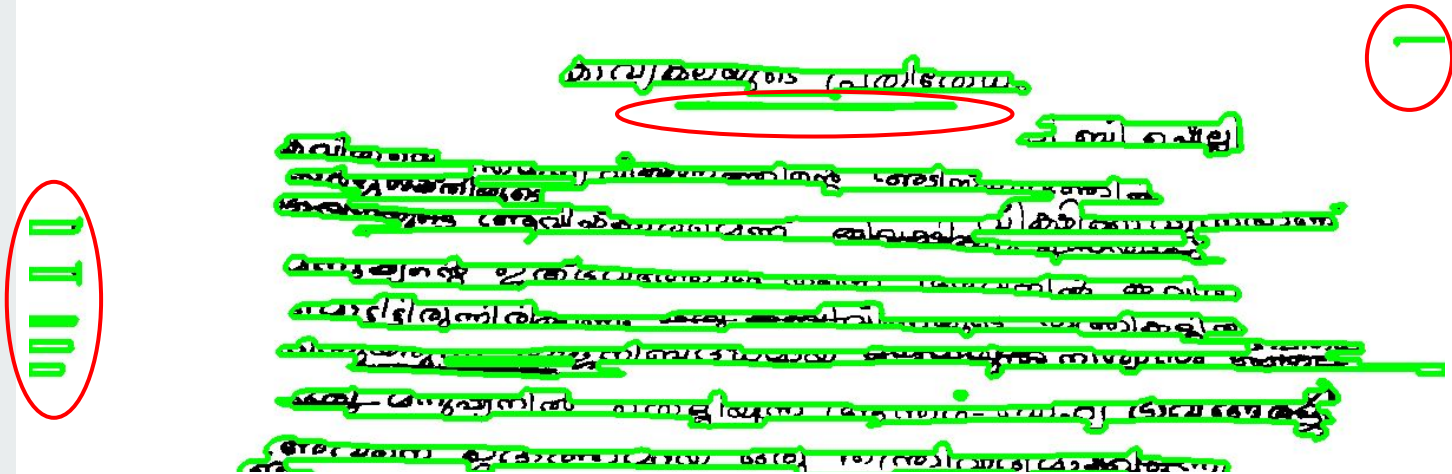
Morphological Closing followed by contour extraction with masking along with horizontal projection profile to handle closely spaced lines.

Tools

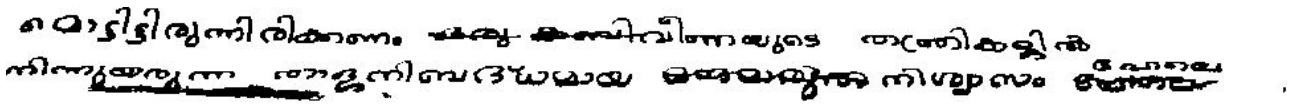
OpenCV

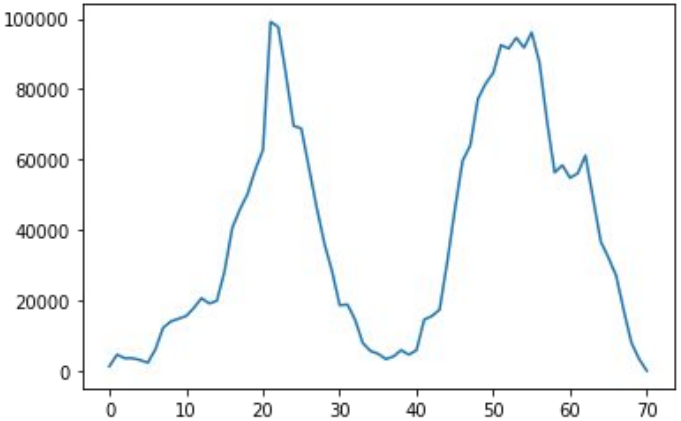
Removing unwanted lines

- Unwanted segmented lines (underlines and other small noise are) identified as lines .
- Such lines are removed by classifying the segmented lines according to their height and selecting lines with a height greater than a threshold value

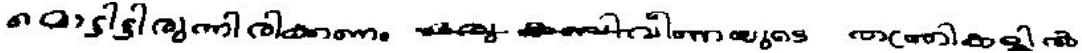


Segmenting closely spaced lines





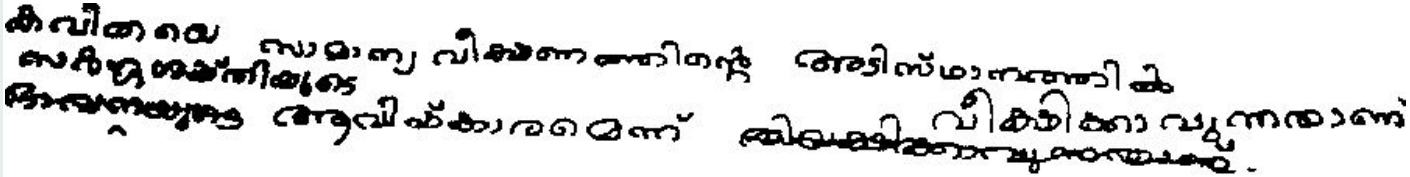
Line 1:



Line 2:

New challenge identified

- Segmentation of closely spaced curved lines



Word Segmentation

Objective

To extract words from the segmented lines

Input : Segmented lines
Output : Segmented words

Method used

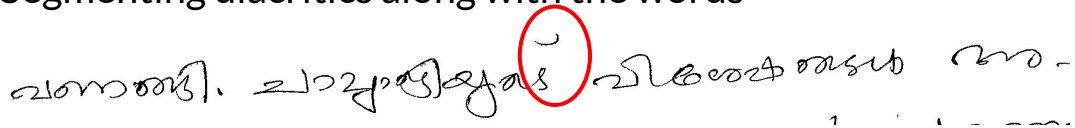
Morphological Closing or RLSA followed by contour extraction

Tools

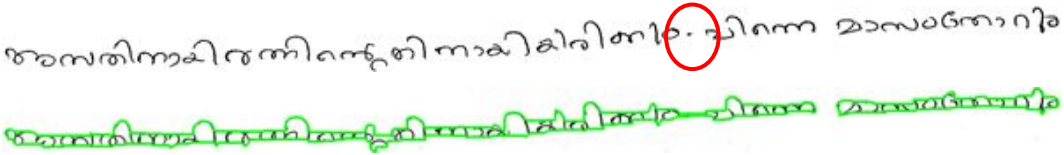
OpenCV

New Challenges Identified

- Segmenting diacritics along with the words



- Presence of fullstop between closely spaced words may result in undersegmentation.



RLSA

- Run Length Smoothing Algorithm

- Replaces a sequence of background pixels with foreground pixels if the number of background pixels in the sequence is smaller than or equal to a predefined threshold

- The algorithm transforms a binary sequence x into an output sequence y according to the following rules:
  1) 1's in x are changed to 0's in y if the number of consecutive 1's is less than or equal to a predefined limit C
  2) 0's in x are unchanged in y

RLSA

Function used

```
def rlsa(image: numpy.ndarray, horizontal: bool =
True, vertical: bool = True, value: int = 0)->
numpy.ndarray:
def rlsa(image: numpy.ndarray, horizontal: bool =
True,  vertical: bool = True, hvalue: int = 0,
vvalue: int = 0) -> numpy.ndarray:
```

How lines, words, characters can be segmented?

hvalue

- Large value - lines
- Intermediate value - words
- Small value - characters

hvalue	vvalue
>60	22
width/42 to 55	22
3	3

RLSA Comparison

- **Performing HRLSA and VRLSA separately on the input image and performing bitwise AND operation.**

```
image_rlsa_horizontal = rlsa(image_binary.copy(),True, False , 50, 0)
image_rlsa_vertical = rlsa(image_binary.copy(),False, True, 0, 22)
image_and = cv2.bitwise_and(image_rlsa_horizontal, image_rlsa_vertical,
mask=None)
```

- **Performing HRLSA on input image followed by VRLSA on the output obtained**

```
image_rlsa = rlsa(image_binary.copy(), True,True ,image.shape[1]//42,22
```

RLSA Comparison - Special Case

1. Input line

2. Performing HRLSA on input image

3. Performing VRLSA on input image

A. Performing AND operation 2 & 3

B. Performing HRLSA followed by VRLSA on input image without AND operation

Character Segmentation

Objective

To extract isolated characters from the segmented words

Input	:	Segmented words
Output	:	Segmented characters

Method used

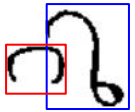
RLSA followed by contour extraction along with masking algorithm to handle overlapping bounding boxes

Tools

OpenCV

Contour extraction with masking algorithm

- Contours of each character are obtained and they are then sorted from left to right. Then, for each contour, its rectangular bounding box is calculated.
- The **Region of Interest (ROI)** for each contour is obtained by selecting that region of the whole grayscale document which corresponds to the stored values of the bounding box.
- When ROI is obtained, the bounding boxes of closely spaces characters may overlap

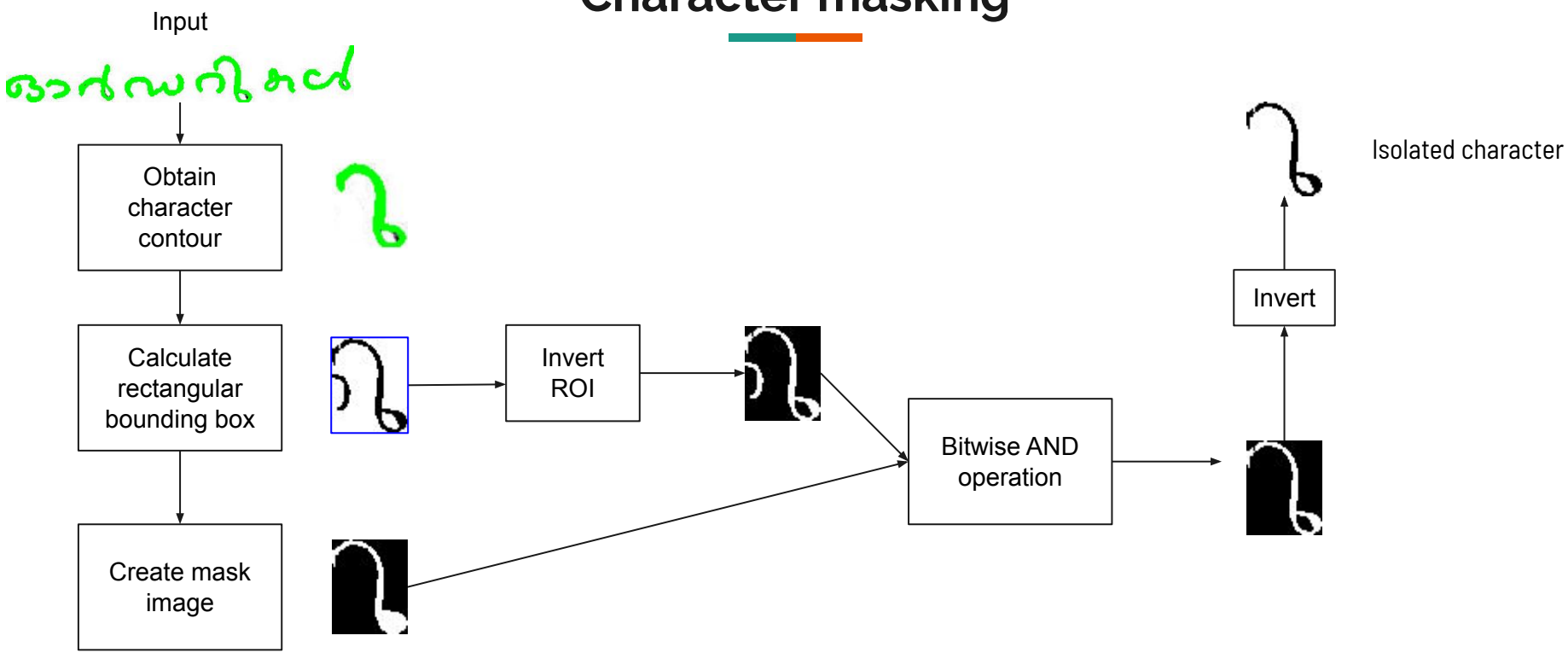


- To tackle this problem, a masking operation is performed to isolate the character

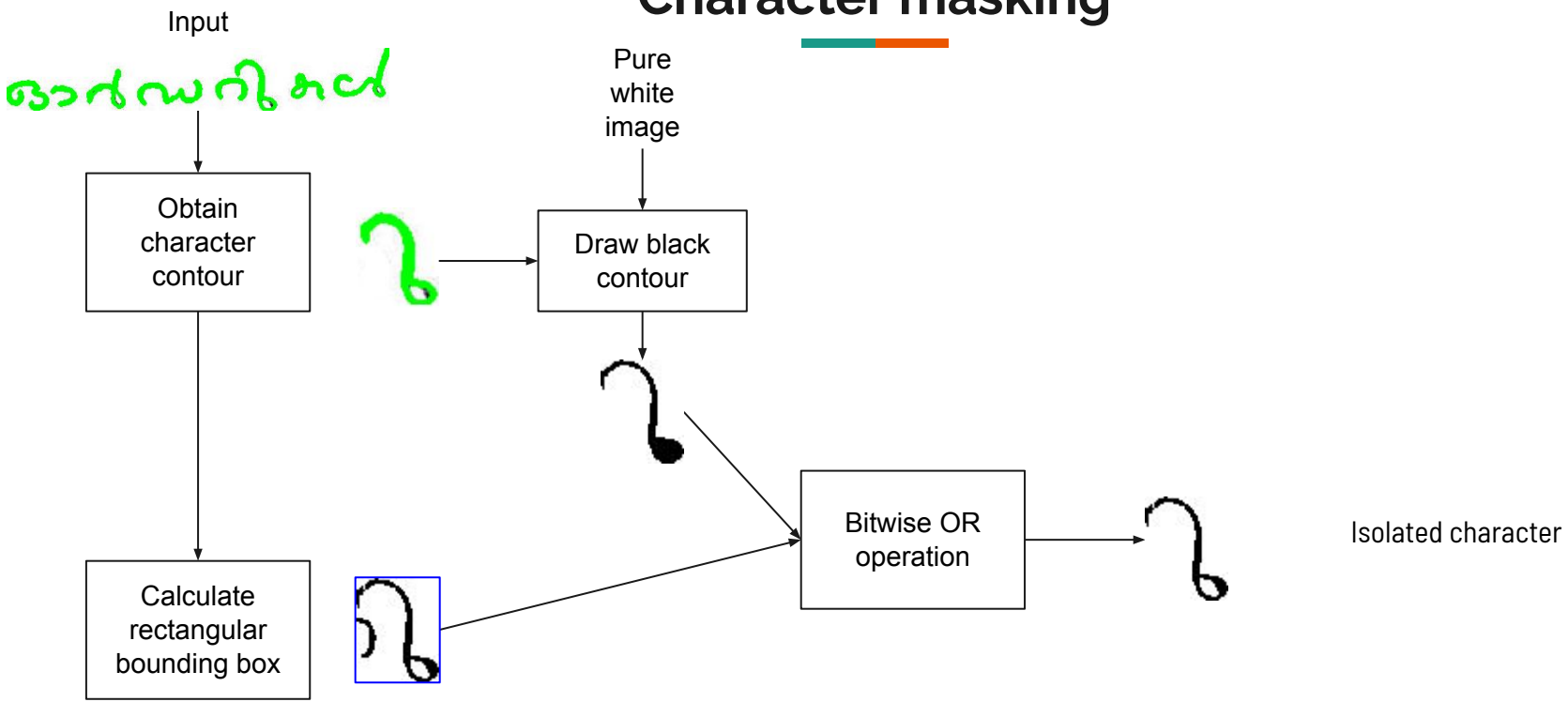
Reference :

Hashrin C.P., Amal Jossy, Sudhakaran K., Thushara A., Ansamma John, "Segmenting Characters from Malayalam Handwritten Documents", *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT),* 20190

Character masking



Character masking



EAST

Reference :
Xinyu Zhou et. al, “EAST: An Efficient and Accurate Scene Text Detector”
Xinyu Zhou, 2017

Future Work (Segmentation)

- Handling segmentation of closely spaced and touching lines
- A global parameter value for morphological closing, RLSA that effectively segments all input images
- Handling connected characters and masking of characters
- Combine all segmentation stages and evaluate segmentation accuracy

MODULE 3

Feature Extraction

Feature Extraction

Objective

To scale down the original document set by evaluating certain features which are capable of distinguishing an input character from the other.

Input	:	Segmented characters
Output	:	Feature vectors

Method used

Structural Features, Zoning, Wavelet Transform

Tools

OpenCV

Dataset

P-ARTS Handwritten Malayalam characters

Challenges

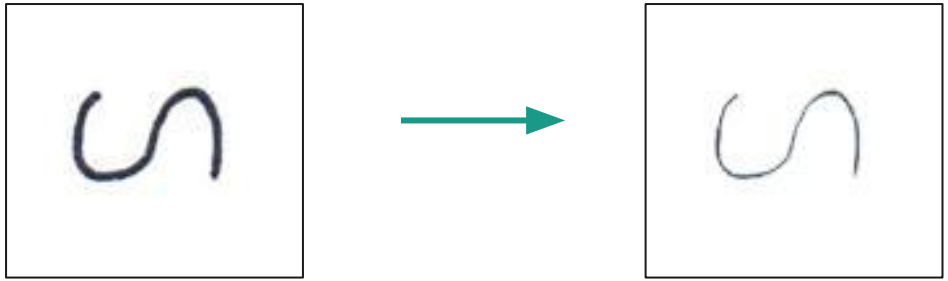
- Presence of Identical Features for some characters
- Efficient methods to utilize curves as a feature for the characters

Parabola Curve Fitting Based Features

- Parabola curve fitting is a feature extraction method which is used to extract the curve features of the malayalam character.
- The main aim of feature extraction phase is to detect various features of character image, which maximizes the recognition accuracy.
- The extracted features when given to a classifier should be able to recognise the character accurately.

Parabola Curve Fitting Based Features

- The first step in this method is to thin the character images as shown below.

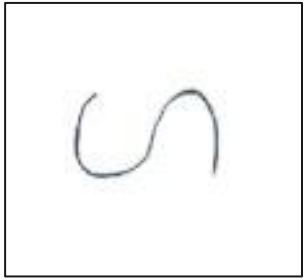


Character image

Thinned Character image

- The Thinned character image is then binarized and then it is inverted.

Parabola Curve Fitting Based Features



Thinned Character image

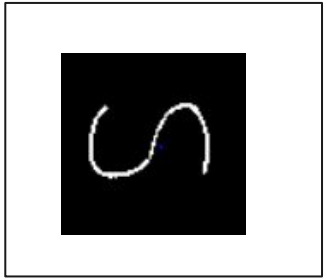


Image after thresholding and inversion

- The next step is to divide the thinned inverted image into different zones. Here the image is divided into 36 zones of equal sizes.

Parabola Curve Fitting Based Features

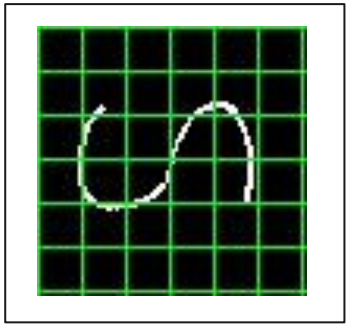


Image divided into different zones

- For each zone, fit a parabola using the least square method and calculate the values of a, b and c. Here a, b and c are the coefficients of the equation of the curve.

- Corresponding to the zones that do not have a foreground pixel, set the values of a, b and c to zero.

Parabola Curve Fitting Based Features

Output example :

```
X-coordinates
[24 24 24 25 25 25 26 26 26 26 26 27 27 27 27 27 28 28 28 28 28 29 29 30
 30 31 31 31 32 32 32]

Y-coordinates
[30 31 32 30 31 32 25 26 27 28 29 25 26 27 28 29 25 26 27 28 29 21 22 21
 22 18 19 20 18 19 20]

a= -0.035614032105772964  b= 0.36376825485893427  c= 42.991149309431236
```


Reference :

Kumar, M., Sharma, R.K. & Jindal, M.K.. “Efficient Feature Extraction Techniques for Offline Handwritten Gurmukhi Character Recognition”, *Natl. Acad. Sci. Lett,* 381–391 (2014).

Character Recognition

Objective

To recognise the character by passing the extracted features to a classifier

Input	:	Extracted Features
Output	:	Classification score

Implementation

Deep Learning Model + Optimizer

Tools

Tensorflow/Keras

Work Distribution

Phase II

Preprocessing	**Aromal**	70%
Line Segmentation	**Harisankar**	70%
Word & Character Segmentation	**George**	50%
Feature Extraction	**Nikhil**	60%
Recognition	**(parallel)**	Not yet started
Total		50%

- Scheduled 8 meetings with guide and incorporated the suggestions.

Work Distribution

Preprocessing	**Aromal**	90%
Line Segmentation	**Harisankar**	90%
Word & Character Segmentation	**George**	90%
Feature Extraction	**Nikhil**	70%
Recognition	**(parallel)**	Not yet started
Total		68%

- Scheduled 13 meetings with guide and incorporated the suggestions.

Project Timeline

Phases/Month	OCT		NOV						DEC			JAN				FEB
Dates	21	31	1	10		20	30	MILESTONE 1 COMPLETED	1	15	31	1	10	20	31	MILESTONE 2 COMPLETED
Feature Extraction																
Classifier																
Preprocessing																
Segmentation																

Learning Phase | Development Phase | Completion | Testing | Further research

Project Timeline

Phases/Month	OCT	NOV		DEC		JAN		FEB	MAR	APR		MAY
Dates		21	31	15	31	20	31	28	31	10	30	31
Preprocessing												
Segmentation												
Feature Extraction												
Classifier												

Learning Phase Development Phase Completion Testing Further research

Project Timeline

Phases/Month	NOV	DEC		JAN		FEB	MAR		APR		MAY		JUN
Dates		15	31	20	31	28	15	31	15	30	15	31	15
Preprocessing													
Segmentation													
Feature Extraction													
Classifier													

Learning Phase Development Phase Completion Testing Further research

Paper Status

Introduction	Preprocessing completed
Literature Survey	Ongoing
Methodology	Ongoing
Experimental Results	Collecting Results

References

1. K, MANJUSHA, et al. “Implementation Of Rejection Strategies Inside Malayalam Character Recognition System Based On Random Fourier Features And Regularized Least Square Classifier.” *Journal of Engineering Science and Technology*, vol. 13, no. 1, 2018, pp. 141 - 157. 20.

2. Hashrin C.P., Amal Jossy, Sudhakaran K., Thushara A., Ansamma John, “Segmenting Characters from Malayalam Handwritten Documents”, *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT),* 2019

3. Subhash Panwar, Neeta Nain, “Handwritten Text Documents Binarization and Skew Normalization Approaches” , *IEEE Proceedings of 4th International Conference on Intelligent Human Computer Interaction*, 2012