OPTICAL CHARACTER RECOGNITION OF PRINTED ODIA DOCUMENTS

Thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Meenhaz MK Mishra
(Roll: 110CS0566)

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India

OPTICAL CHARACTER RECOGNITION OF PRINTED ODIA DOCUMENTS

Thesis submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Meenhaz MK Mishra
(Roll: 110CS0566)

Under the guidance of

Prof. Ratnakar Dash

NIT Rourkela

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Orissa, India.

Certificate

This is to certify that the work in the project entitled optical recognition of Odia characters by Meenhaz MK MIshra is a record of their work carried out under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

Prof. Ratnakar Dash
Dept. of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769008

ACKNOWLEDGEMENT

I take this opportunity to express my gratitude and regards to my guide Prof. RATNAKAR DASH for his exemplary guidance, monitoring and constant encouragement throughout the course of this project.

I also take this opportunity to express a deep sense of gratitude to my friends for their support and motivation which helped me in completing this task through its various stages.

I am obliged to the faculty members of the Department of Computer Science & Engineering at NIT Rourkela for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment. In particular I am thankful to Mr.TK Mishra for offering advice and help on important issues relating to the project. Lastly, I thank my parents for their constant encouragement without which this assignment would not have been possible.

Meenhaz MK Mishra

110cs0566

Author's Declaration

I hereby certify that all the work contained in this report is done by me unless otherwise acknowledged. Also, all of my work has not been previously submitted for any academic degree. All sources of quoted information have been acknowledged by means of appropriate references.

Meenhaz MK Mishra

Roll No - 110CS0566

NIT ROURKELA

ABSTRACT

Optical Character Recognition (OCR) is a document image analysis method that involves the mechanical or electronic transformation of scanned or photographed images of typewritten or printed text into text that can be easily read by the computer. OCR has been become a very widespread area of interest and research because of its ability to narrow the reading ability gap between computers and humans and because it improves human machine interaction in many applications. Example applications include cheque verification, and a large variety of banking, business and data entry applications. The project involved skew correction of odia documents, line segmentation and eventual segmentation of odia characters. The project involved segmentation of a document into its constituent lines, then treating the line as one entity, it segmented the words. Now, once the words are segmented, the characters are extracted one by one. The algorithms used here stand true for all the devnagri scripts. Hence examples of telgu word segmentation is also done just to show as an proof of the applied algorithm.

Contents

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

OCR is a process of recognition of characters from scanned and digital images. It has the scope of immense contribution towards advancing automation process and can improve the man- computer interaction in many applications. Some practical applications of OCR systems are:-

- Aiding the blinds in reading
- Automation of text input into the computer for desktop publication
- Automation of reading to sort postal mail and important documents.
- Document data compression: from document image to ASCII format.

Traditionally OCR systems are classified into two types

- Template based
- And Feature Based Approach

In template based methodology, an obscure example is superimposed on the perfect layout design and the level of correspondence between the two is utilized for the choice about grouping. Early OCR frameworks utilized this methodology. Yet they have gotten insufficient in vicinity of commotions, progressions of penmanship and so forth. Current frameworks along these lines join it with feature based methodology to acquire better results. Feature based models determines more critical properties of the test examples and utilizes them in a more modern order model.

1.1.1 STEPS IN AN OCR MODEL.

Pre-processing

Before any recognition or feature extraction technique is applied to a document, it needs to be made suitable for the same. Some of the preprocessing modules include :

- Binarisation
- Skew correction
- Segmentation

FEATURE EXTRACTION

In this step, various features, ranging from simple to complex are extracted from the image. Typical examples of such features are :

- Lines, ridges and edges
- Localized interest points like blobs and corners.
- Complex features like texture, gradient, curvature, shape etc.

CLASSIFICATION

Once the features are extracted from the input images, they are now compared with the existing database. And keeping in lines with the closeness or resemblance, they are classified. It only helps to make the next step, character matching easier.

1.1.2 ODIA SCRIPT AND CHALLENGES FACED

Despite the fact that quick progressions have occurred in the optical character recognition arrangement of numerous remote dialects and a percentage of the odia scripts, the exploration levels in Odia are falling behind. The issue proclamation of the task is to distinguish the diverse properties of Odia script and discover intends to handle them.

Odia is a language used in the state of Odisha, India. The odia script developed from the kalinga script which is a descendent of Brahmi script of ancient India. The modern Odia script has 11 vowels and 41 consonants. These are called the basic characters.

Composing of the odia script is from left to right and the idea of upper case and easier case letters is truant. The characters are roundish in nature and look practically like one another. That makes it troublesome for grouping. However there are no flat lines like in Bengali and hindi scripts which makes division a ton less demanding.

At the point when a vowel takes after a consonant, it takes an adjusted shape. Contingent upon the vowel utilized, it is dead set whether it is put at the left, right, best or at the bottom of the consonant. The adjusted shapes are called Matras. Also when a consonant or vowel takes after the consonant, it off and on again takes a changed shape called compound character. There are almost 200 compound characters in the odia script and around 300 classes must be built. A standout amongst the most testing parts of the script is the vicinity of similitude fit as a fiddle.

1.2 MY CONTRIBUTION

In this project, I have implemented the preprocessing stage of the OCR system of odia characters. New algorithms have been implemented in this paper. The character recognition techniques for telgu scripts was implemented on odia scripts also, and found to give better results.

1.3 Organization of thesis

The first chapter of the thesis deals with the introduction of OCR systems and explains it's working. Details have been provided so as to ensure that the reader gets a historical perspective of evolution of OCR systems. In the 2nd chapter we carry out a detailed analysis of the scholarly articles that I had read before implementing the project. Three papers have been analyzed in all, though a bit of information is added

through other sources that I had studied. The 3rd chapter deals with the implementation of the algorithms that were studied. Here

CHAPTER 2

LITERATURE SURVEY

In this chapter, we will analyze all the references that have been taken up by us. This will help us understand the various implementation issues in the existing OCR techniques.

2.1 Review of the existing OCR techniques.

In this chapter, we have analyzed the past techniques of ocr systems and their evolution into the modern times.

2.1.1 History

Traditionally OCR systems are classified into two types

- Template based
- And Feature Based Approach

2.1.2 Types of Approaches

In template based methodology, an arbitary example is superimposed on the perfect format design and the level of relationship between the two is utilized for the choice about grouping. Early OCR frameworks utilized this methodology. Anyhow they have gotten incapable in vicinity of commotions, progressions of penmanship and so on. Advanced frameworks accordingly consolidate it with characteristic based methodology to acquire better comes about.

Feature based models determines more imperative properties of the test examples and utilizes them in a more advanced order model.

2.1.3 PREPROCESSING

The fundamental standard of recognition-based character division is to utilize a portable window of variable width to give the speculative divisions which are affirmed (or not) by the grouping. Picture obtaining and preprocessing are the two moderately basic stages, which are introduced first. Picture securing is at the picture representation level of pattern recogntion(PR). It is the methodology of securing a machine representation of an archive or an article to be perceived. A flatbed scanner is utilized at this stage to secure 200 dpi, 8-bits light black-level pictures. Preprocessing is at the picture-to-picture conversion level. It is the procedure of repaying a low quality-unique and/or low quality-examining. There are two techniques to improve the obtained picture in the proposed framework, which are binarization and smoothing. Character division could be performed by either the dissection or recognition-based procedure. Dissection implies the decay of the picture into a grouping of sub-pictures utilizing general characteristics. It includes investigation of the picture to discover the sub-picture division ways. Each one sub-picture is dealt with as a character for recognition. It is worth saying that arrangement of characters is completed at a later stage. Projection analysis, connected component preparing, and white space and pitch discovering are a percentage of the regular dissection strategies utilized by OCR frameworks. These procedures are suitable for scripts which have large character spaces in between then. In the event that a dissection system is utilized for cursive scripts, a more `intelligent and particular dissection strategy for the specific script is required. Be that as it may, there is still no surety that high segmentation exactness might be accomplished. The essential guideline of recognition-based character segmentation is to utilize a portable window of variable width to give the experimental divisions which are affirmed (or not) by the order. Characters are by- results of the character recognition for frameworks utilizing such a guideline to perform character segmentation. The principle point of interest of this strategy is that it sidesteps genuine character segmentation issues. On a basic level, no particular segmentation algorithms for the particular script is required and recognition failures are principally because of disappointments throughout the classification stage. Consequently, more cursive script OCR frameworks utilize this procedure for enhancing the recognition correctness. This

methodology is otherwise called without division recognition because of the virtual nonattendance of the character partition stage.

Binarization is an extraordinary instance of thresholding, of which there are just two states of yields in the ensuing picture, either dark or white. It diminishes the computational necessities of the framework and may empower evacuation of some noise. A document could be binarized comprehensively or adaptively. Unless the archive is printed on an uneven shaded paper, worldwide thresholding is sufficient to do the binarization. Two worldwide thresholding calculations were concentrated on and executed. A smoothing procedure was taken. It is essential to note that this calculation can smooth the picture as well as restore missing pixels.

2.2 Review of Coupled snakelets for curled text line segmentation Scheme:

2.2.1 Text line segmentation

Camera-caught, distorted report pictures generally hold twisted content-lines due to twists brought on by camera point of view and page twist. Distorted record pictures could be changed into planar archive pictures for enhancing optical character recognition precision and human intelligibility utilizing monocular DE twisting procedures. Twisted text-lines division is a urgent beginning venture for the vast majority of the monocular DE twisting methods. Existing twisted content- line division methodologies are touchy to geometric and viewpoint mutilations. In this paper, we present a novel twisted content-line division calculation by adjusting dynamic form (snake).our calculation performs content-line division by assessing sets of x-line and pattern. It evaluates a neighborhood pair of x-line and standard on each one joined part by mutually following top and base purposes of neighboring associated parts, and finally each one gathering of cover- ping sets is recognized as a sectioned content-line.

2.2.2 Monocular Dewarping

Dewarping is moderately another report picture preprocessing step when contrasted with others which are said here. It is a methodology of correcting Polaroid-caught report pictures that experience the ill effects of point of view and geometric mutilations. It is possible either by applying stereo vision strategies or by utilizing monocular dewarping methods—a dewarping method that is produced for pictures which are caught by single camera is known as a monocular dewarping system. A large portion of the detail-of-the-craftsmanship monocular dewarping systems are focused around text-line segmentation.

2.2.3 Snakelet Coupling

The top and lowest part snakes are made out of the same number of focuses with comparative qualities of x-directions. For every regular quality of x-direction of the top and base snakes, outright separation is figured from the comparing qualities of y-directions. At that point, normal separation is processed. Presently, for every normal worth of x-direction of both snakes, the comparing qualities of y-directions are expanded or diminished relatively such that the separation between them gets equivalent to the normal separation.

2.3 Review of OCR techniques used for Telgu character segmentation

This paper talks in detail about the problems faced in line segmentation and character segmentation of Indian scripts.

2.3.1 Line Segmentation

The different lines of content are identified by figuring a histogram of the amount of dark pixels in each one column and after that finding the valleys in the projection profile. The spaces between the lines could be identified by the vicinity of columns with no dark pixels identified in the histogram. The lines with the non-zero number of dark pixels demonstrate the columns of content. The circumstances in Telugu content is different. The vicinity of modifier images above and underneath the primary character suggests that the continuous columns with non-zero values in the histogram are the consequence of three varieties

The amount of persistent columns with histogram number >0 demonstrates the stature of the characters in that set of lines. The greatest of this number for the first three sets of columns is taken as the stature of the fundamental character. Along these lines, if the line hole is short of what or equivalent to a third of the measure of the primary character, it is not treated as a detachment between the lines. Else, it is dealt with as a partition between the lines.

2.3.2 Word Segmentation And Character Segmentation

To break a line into its constituent words, point of interest is taken of the gap between the words (word space) and a more modest hole between the characters (character space) constituting a statement. The character hole for Telugu characters was discovered to be roughly 2/7ths of the stature of the primary character. To find the expression crevices in a line, a histogram of the amount of dark pixels in indi-vidual segments is plotted. On the off chance that the number of the ceaseless spotless segments in the projection profile is short of what 2/7ths of the tallness of a character, it is a character gap. On the off chance that it is more terrific than or equivalent to that, it is a statement gap

CHAPTER 3

THE TECHNIQUES IMPLEMENTED

3.1. Binarization

Binarization is a special case of thresholding, of which there are only two states of outputs in the resulting image, either black or white. It reduces the pressure of computational requirements of the system and may enable removal of some unwanted signals or noise. In this case we converted the image into a gray scale image. The white pixels have the value of 1 and the black pixels have value 0. The process helps us to reduce the image properties to such an extent that we don't to have to compute too much to get the desired results. The implementation becomes far easier once the technique of binarization is implemented. We also have otsu's algorithm in case we want to do an indepth study of the process, but it is not required for the scope of the project

3.2. SKEW DETECTION AND CORRECTION

If we have a skewed image, it would be impossible for us to extract the words and hence go on with the character recognition process.

The following algorithm was implemented:

1. The bottom most black pixel was found in each column of a text region of the image.
2. For each such black pixel, connected component and is determined and then its se bottom most pixel is selected only if its adequately sized. Otherwise, the bottom most pixel is rejected as noise. The next is selected and the process is repeated.
3. The pixels identified as the bottom most pixel may actually belong to the same line above the bottom most pixel. Those must be dropped from future considerations.
4. Apply hough transform on the bottom most pixel of each connected component.

5. The angle and offset from the origin for which the maximum number of pixels in step 4 are in a straight line is estimated skew angle.
6. The angle is determined by the leftmost pixels of the first connected component of that particular image and then confirmed.
7. In case any problem arises, the connected component of the bottom most line is ignored. The same process is repeated for the penultimate row.
8. The skew is corrected by tilting the image back by the negative of the skew line.

ANALYSIS

The algorithm gives almost accurate results. However it can be improved by taking more combinations of slope and offsets. The error arises due to discretization.

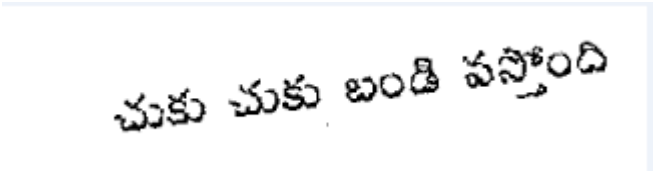

Fig 3.1 original image

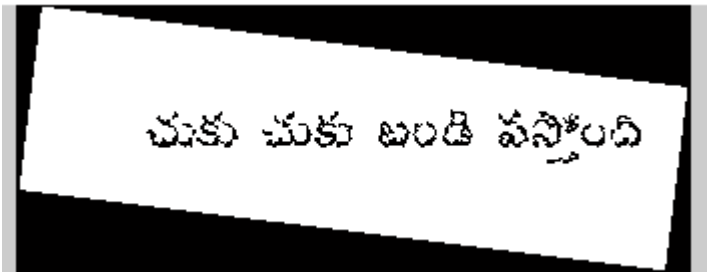

Fig 3.2 Image after skew correction

Fig 3.3 image after cropping

3.3 LINE SEGMENTATION

The algorithm applied:

The vicinity of modifier symbols above and underneath the primary character suggests that the constant columns with non-zero values in the histogram are the consequence of 3 mixtures of symbols.

1. Modifiers encountered above the main characters.
2. Main characters
3. Modifiers encountered below the main characters.

It is watched that the amount of dark columns in the histogram is more modest between the primary character and the modifier and bigger between the distinctive lines of content. The amount of nonstop columns with histogram number more than 0 show the stature of the character in that set of the column. The max of this number of the first 3 sets of columns is considered the stature of the primary character.

So, if the line gap is less than or equal to a third of the size of the main character, it is not considered a separation between two lines.

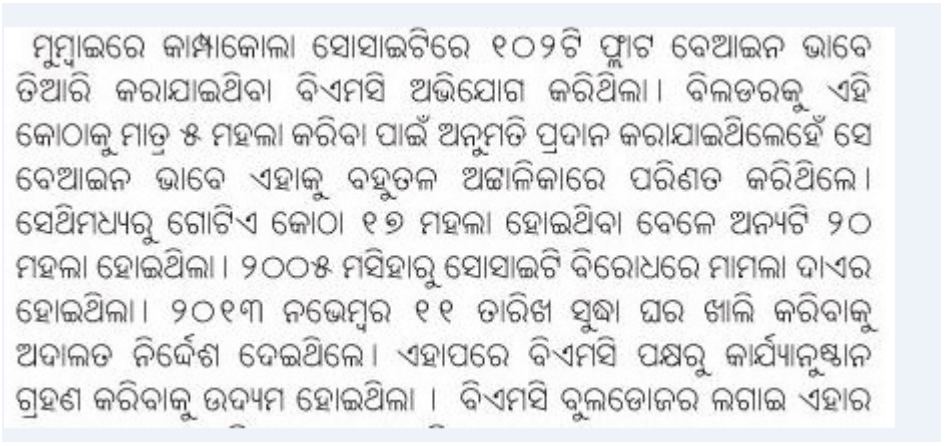
ମୁମ୍ବାଇରେ କାମ୍ପାକୋଲା ସୋସାଇଟିରେ ୧୦୨ଟି ଫ୍ଲାଟ ବେଆଇନ ଭାବେ ତିଆରି କରାଯାଇଥିବା ବିଏମସି ଅଭିଯୋଗ କରିଥିଲା। ବିଲଡରକୁ ଏହି କୋଠାକୁ ମାତ୍ର ୫ ମହଲା କରିବା ପାଇଁ ଅନୁମତି ପ୍ରଦାନ କରାଯାଇଥିଲେହେଁ ସେ ବେଆଇନ ଭାବେ ଏହାକୁ ବହୁତଳ ଅଟ୍ଟାଳିକାରେ ପରିଣତ କରିଥିଲେ। ସେଥିମଧ୍ୟରୁ ଗୋଟିଏ କୋଠା ୧୭ ମହଲା ହୋଇଥିବା ବେଳେ ଅନ୍ୟଟି ୨୦ ମହଲା ହୋଇଥିଲା। ୨୦୦୫ ମସିହାରୁ ସୋସାଇଟି ବିରୋଧରେ ମାମଲା ଦାଏର ହୋଇଥିଲା। ୨୦୧୩ ନଭେମ୍ବର ୧୧ ତାରିଖ ସୁଦ୍ଧା ଘର ଖାଲି କରିବାକୁ ଅଦାଲତ ନିର୍ଦ୍ଦେଶ ଦେଇଥିଲେ। ଏହାପରେ ବିଏମସି ପକ୍ଷରୁ କାର୍ଯ୍ୟାନୁଷ୍ଠାନ ଗ୍ରହଣ କରିବାକୁ ଉଦ୍ୟମ ହୋଇଥିଲା । ବିଏମସି ବୁଲଡୋଜର ଲଗାଇ ଏହାର

Fig.3.4 Input image for line Segmentation

RESULTS AFTER LINE SEGMENTATION

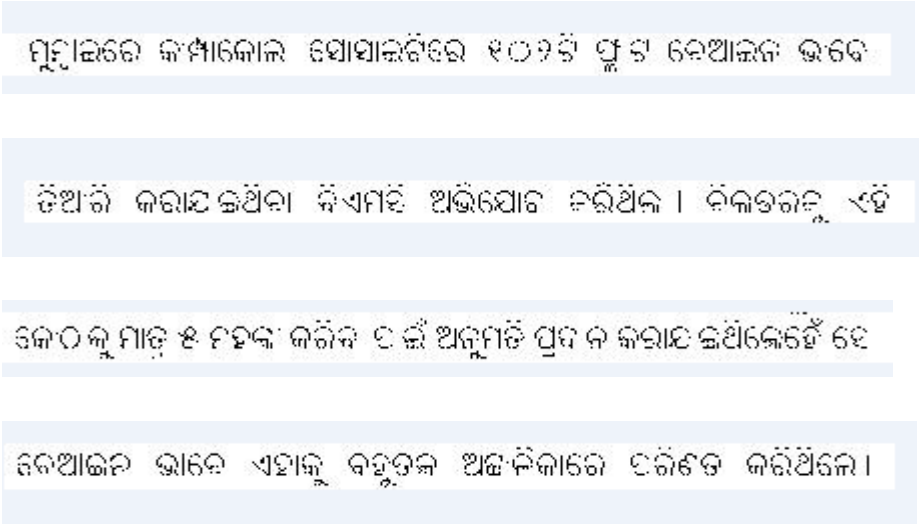

Fig 3.5 output image for line segmentation

3.4 WORD SEGMENTATION

In this step we attempt to divide the lines into words.

The algorithm used is:

1. One line of text is taken as input.
2. A vertical scan is done.
3. If in one vertical scan two or less black pixels are encountered, then scan is denoted by 0. Else the scan is denoted by the no. of black pixels.
4. The vertical projection profile is constructed.
5. If in the profile, there exists a run of at least wg consecutive 0s, then the midpoint of the run is considered as the limiting line of the word.
6. Value of wg is taken as one- third of the text line height.

The reason for taking up this algorithm is that if in case there are no black pixels, it is considered as a blank area. The spaces are always in accordance with the height of the text. The one third ratio is obtained from statistical evidence.

The boundary of the word is thus determined. It is taken as the middle of the run and not the beginning because that might lead to incoherent images.

Fig 3.6 Input odia image

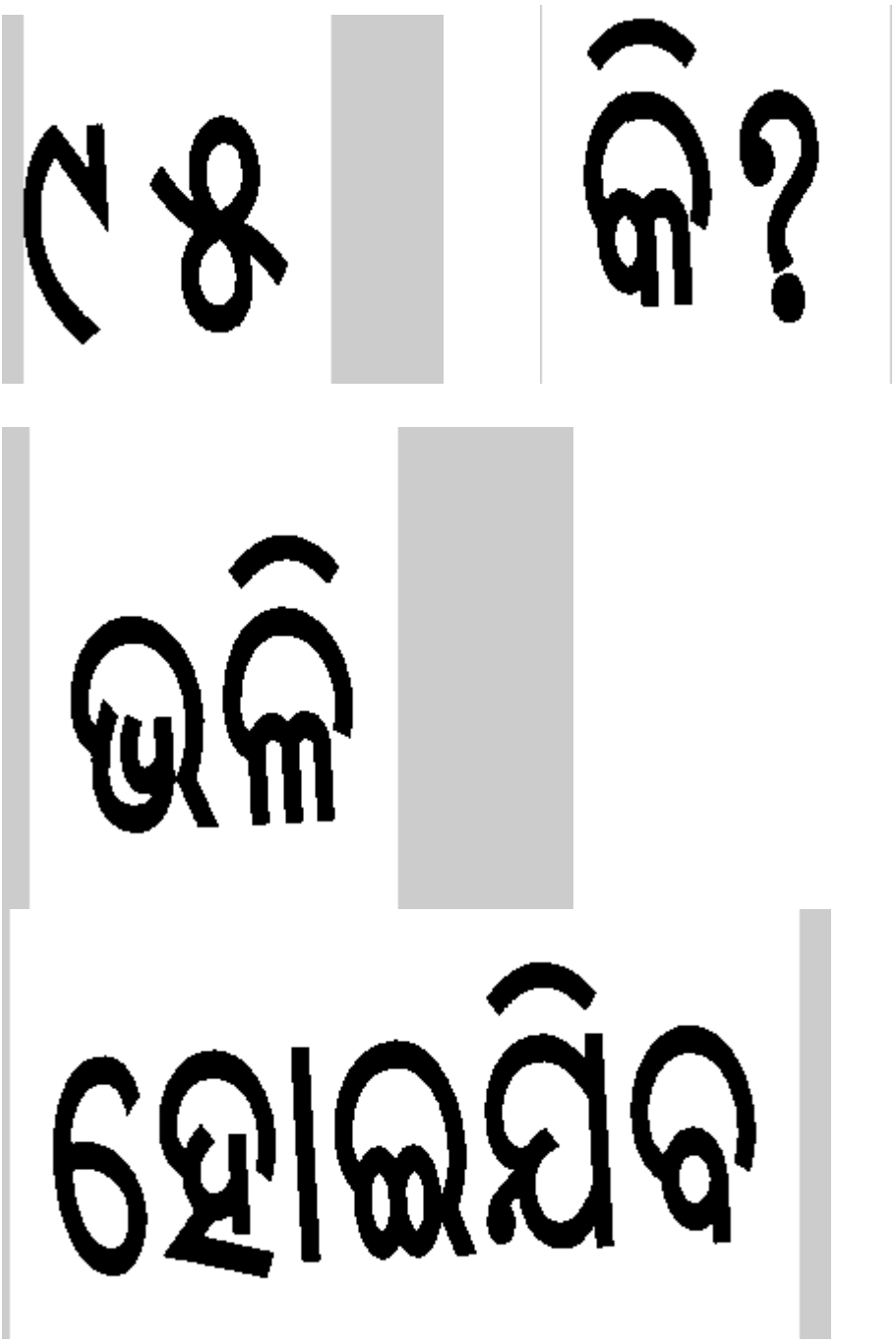

Fig 3.7 output images for word segmentation

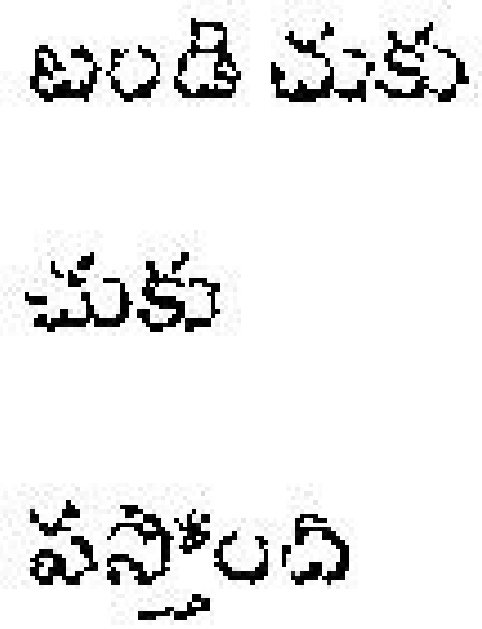

Fig 3.8 output images for telgu word segmentation

3.5 CHARACTER SEGMENTATION

The character segmentation algorithm is very similar to the word segmentation algorithm. We have employed dissection algorithm. Hence we only have to fix the space between the characters. Here we have taken the space to be two seventh of the height of the main character.

The other feature that needs to be taken care of is the height of the character. Because it contains a lot of modifiers, the height of characters vary from each other. We take the max height of the character and we define our window according to that.

After character segmentation.

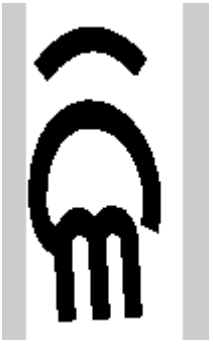

Fig 3.9 output images for odia character segmentation

CHAPTER 5

Conclusion and Future Scope

In this thesis I have implemented the different stages of preprocessing on the odia documents. We have also tested the results with the telgu script. It shows us that the algorithm can be implemented without any hassles. The characters this extracted can be used for the next stages of character recognition i.e feature extraction and classification. The characters thus extracted are in a way that the modified characters are presented in a division of its separate constituents. So the researcher furthering the scope of this project will have lesser difficulties in the modified characters. Also the line segmentation technique used can be implemented for only printed documents. However because we have used the skew correction method even in the other module, the skew correction works for both hand written as well as printed odia documents.

Bibliography

[1] Maiyre Ibrayim, Askar Hamdulla and Dilmurat Tursun *Dynamic Programming Method for Segmentation of Online Cursive Uyghur Handwritten Words into Basic Recognizable Journal Of Software, VOL. 8, NO. 10, OCTOBER 2013*

[2] A. Cheung, M. Bennamoun, N.W. Bergmann. *An Arabic optical character recognition system using recognition-based segmentation*

[3] Syed Saqib Bukhari, Faisal Shafait, Thomas M. Breue, *Coupled snakelets for curled text-line segmentation from warped document images,* DOI 10.1007/s10032-011-0176-2.

[4] Rong Cheng and Yanping Bai, *A Novel Approach for License Plate Slant Correction, Character Segmentation and Chinese Character Recognition,* International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.7, No.1 (2014), pp.353-364

[5] C. Vasantha Lakshmi Æ C. Patvardhan, *An optical character recognition system for printed Telugu text*, Pattern Anal Applic (2004) 7: 190–204, DOI 10.1007/s10044-004-0217-2

[6] Ismail Bouazizi, Fahd Bouriss, Yassine Salih-Alj *,Arabic Reading Machine for Visually Impaired People using TTS and OCR,* DOI 10.1109/ISMS.2013.49

[7] Lang li et al, *Character Segmentation and Retrieval for Learning Support System of Japanese Historical Books,* http://dx.doi.org/10.1145/2501115.2501129

[8] Mashasi Koga et al, *Gabor Feature Extraction for Character Recognition: Comparison with Gradient Feature,* Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05) 1520-5263/05