OPTICAL CHARACTER RECOGNITION

A Btech Project report
Submitted in partial fulfillment of
the requirement for the B.Tech.
under Biju Patnaik University of Technology, Rourkela.

Submitted By

DEEPAK KUMAR PADHY **Regd No. 0801314150**

SIVAJI SAHOO **Regd No. 0801314156**

JULY- 2012

Under the guidance of

AJAY ANAND

APEX INSTITUTE OF TECHNOLOGY & MANAGEMENT
Pahala, Bhubaneswar, Odisha – 752101, India

CERTIFICATE

This is to certify that the project work entitled **'OPTICAL CHARACTER RECOGNITION'** is a bonafide work done by **Deepak Kumar Padhy and Sivaji Sahoo** bearing Registration No. **0801314150 and 0801314156** respectively submitted in partial fulfillment of the requirements for the award of the degree **Bachelor of Technology** in **Computer Science and Engineering** during the year 2011-12.

Mr. Ajay Anand

Assistant Professor
Guide

Dr. SATYA RANJAN PATTANAIK

B Tech. Project Coordinator

Prof R. C. Dash

PRINCIPAL

SYNOPSIS

Optical character recognition (OCR), is the translation of scanned images of typewritten, handwritten or printed characters into text format. It is widely used to convert books and documents into text files, to digitalize the documents in an office, or to publish the generated text on website. OCR makes it possible to edit the text, search for a word or phrase, store it more compactly, display or print a copy free of scanning articles. And further augmenting with text to speech conversion it reads out the text file. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

OCR recognition gets complicated by added noise, image distortion, and the various character typefaces, sizes, and fonts that a document may have. OCR systems require optimization techniques to read a specific font; early versions needed to be programmed with images of each character, and worked on one font at a time. "Intelligent" systems with a high degree of recognition accuracy for most fonts are now common. Some systems are capable of reproducing formatted output that closely approximates the original scanned page including images, columns and other non-textual components.

Acknowledgement

We would like to express my immense sense of gratitude to my guide**, Mr. Ajay Anand**, for his valuable instructions, guidance and support throughout my seminar.

We again owe my special thanks to **Dr. SatyaRanjan Pattnaik** , B.Tech. project Coordinator for giving me an opportunity to do this report.

And finally thanks to **Prof. R.C. Das**, Principal, AITM for his continued drive for better quality in everything that happens at AITM. This report is a dedicated contribution towards that greater goal.

SIVAJI SAHOO

DEEPAK KUMAR PADHY

Table of Contents

Section - 1

Introduction

Chapter – 1

Introduction and scope of thesis

Chapter - 1

Introduction and Scope of the Thesis

1.1 What is OCR?

As you read these words on your computer screen, your eyes and brain are carrying out optical character recognition without you even noticing! Your eyes are recognizing the patterns of light and dark that make up the characters (letters, numbers, and things like punctuation marks) printed on the screen and your brain is using those to figure out what I'm trying to say (sometimes by reading individual characters but mostly by scanning entire words and whole groups of words at once).

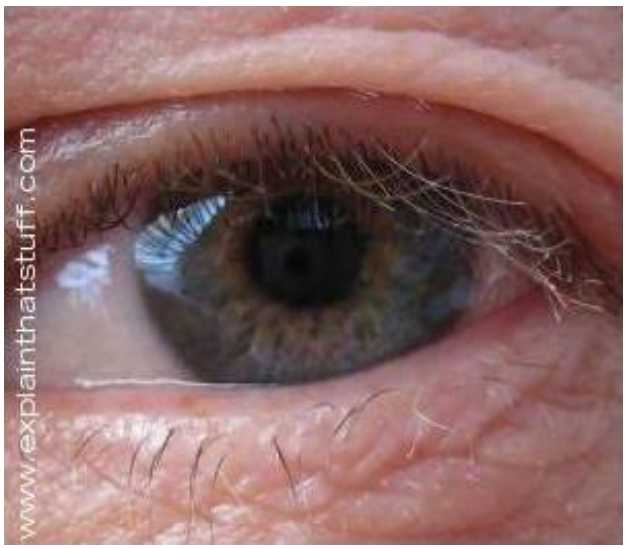


Figure 1.1. 1 Eye

Computers can do this too, but it's really hard work for them. The first problem is that a computer has no eyes, so if you want it to read something like the page of an old book, you have to present it with an image of that page generated with an optical scanner or a digital camera. The page you create this way is a graphic file (often in the form of a JPG) and, as far as a computer's concerned, there's no difference between it and a photograph of the Taj Mahal or any other graphic: it's a completely meaningless pattern of pixels (the colored dots or squares that make up any

computer graphic image). In other words, the computer has a picture of the page rather than the text itself—it can't read the words on the page like we can, just like that. OCR is the process of turning a picture of text into text itself—in other words, producing something like a TXT or DOC file from a scanned JPG of a printed or handwritten page.

1.2 How does it work?

1.2.1 Pattern recognition

If everyone wrote the letter A exactly the same way, getting a computer to recognize it would be easy. You'd just compare your scanned image with a stored version of the letter A and, if the two matched, that would be that. Kind of like Cinderella: "If the slipper fits..."

Figure 1.2.1. 2 Pattern Recognition

So how do you get everyone to write the same way? Back in the 1960s, a special font called OCR-A was developed that could be used on things like bank checks and so on. Every letter was exactly the same width (so this was an example of what's called a monospace font) and the strokes were carefully designed so each letter could easily be distinguished from all the others. Check-printers were designed so they all used that font, and OCR equipment was designed to recognize it too. By standardizing on one simple font, OCR became a relatively easy problem to solve. The only trouble is, most of what the world prints isn't written in OCR-A—and no-one uses that font for their handwriting! So the next step was to teach OCR programs to recognize letters written in a number of very common fonts (ones like Times, Helvetica, Courier, and so on). That meant they could recognize quite a lot of printed text, but there was still no guarantee they could recognize any font you might send their way.

1.2.2 Feature detection

Also known as feature extraction or intelligent character recognition (ICR), this is a much more sophisticated way of spotting characters. Suppose you're an OCR computer program presented with lots of different letters written in lots of different fonts; how do you pick out all letters As if they all look slightly different? You could use a rule like this: If you see two angled lines that meet in a point at the top, in the center, and there's a horizontal line between them about halfway down, that's a letter A. Apply that rule and you'll recognize most capital letter As, no matter what font they're written in. Instead of recognizing the complete pattern of an A, you're detecting the individual component features (angled lines, crossed lines, or whatever) from which the character is made. Most modern omnifont OCR programs (ones that can recognize printed text in any font) work by feature detection rather than pattern recognition. Some use neural networks (computer programs that automatically extract patterns in a brain-like way).

1.2.3 Block Diagram

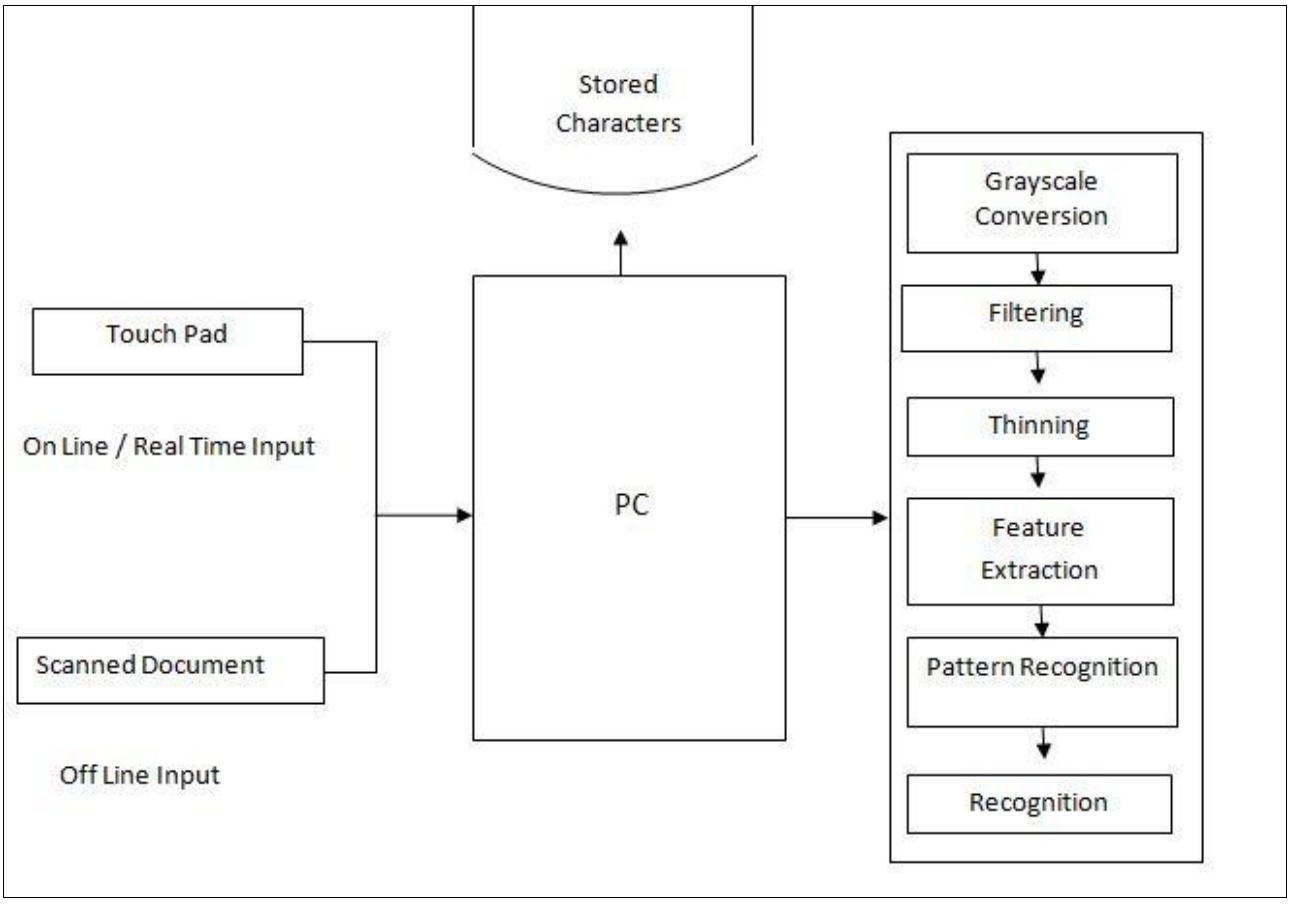


Figure 1.2.3. 3 **Block Diagram for OCR**

1.3 Application of OCR

In recent years, OCR (Optical Character Recognition) technology has been applied throughout the entire spectrum of industries, revolutionizing the document management process. OCR has enabled scanned documents to become more than just image files, turning into fully searchable documents with text content that is recognized by computers. With the help of OCR, people no longer need to manually retype important documents when entering them into electronic databases. Instead, OCR extracts relevant information and enters it automatically. The result is accurate, efficient information processing in less time.

1.3.1 Automatic Plate Recognition

It is a mass surveillance method that uses optical character recognition on images to read the license plates on vehicles. They can use existing closed-circuit television or road-rule enforcement cameras, or ones specifically designed for the task. They are used by various police forces and as a method of electronic toll collection on pay-per-use roads and cataloging the movements of traffic or individuals

Figure 1.3.1. 4 Automatic Plate Recognition

1.3.2 Legal

In the legal industry, there has also been a significant movement to digitize paper documents. In order to save space and eliminate the need to sift through boxes of paper files, documents are being scanned and entered into computer databases. OCR further simplifies the process by making documents text-searchable, so that they are easier to locate and work with once in the database. Legal professionals now have fast, easy access to a huge library of documents in electronic format, which they can find simply by typing in a few keywords.

1.3.3 OCR in other industries

OCR is widely used in many other fields, including education, finance, and government agencies. OCR has made countless texts available online, saving money for students and allowing knowledge to be shared. Invoice imaging applications are used in many businesses to keep track of financial records and prevent a backlog of payments from piling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes. As the technology continues to develop, more and more applications are found for OCR technology, including increased use of handwriting recognition. Furthermore, other technologies related to OCR, such as barcode recognition, are used daily in retail and other industries. To learn more about OCR solutions for your office, you can download a free trial of Maestro Recognition Server, CVISION's OCR toolkit, or Trapeze, our automated form-processing solution

Chapter – 2

Background Research

Chapter - 2

Background Research

2.1 Brief history of OCR

Now a days, there are software's for recognizing only the English characters. It recognizes and stores the characters in ASCII format.

Optical character recognition, usually abbreviated to OCR, is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text.

OCR is a field of research in pattern recognition, artificial intelligence and machine vision. Though academic research in the field continues, the focus on OCR has shifted to implementation of proven techniques. Optical character recognition using optical techniques such as mirrors and lenses) and digital character recognition (using scanners and computer algorithms) were originally considered separate fields. Because very few applications survive that use true optical techniques, the OCR term has now been broadened to include digital image processing as well.

In about 1965, Reader's Digest and RCA collaborated to build an OCR Document reader designed to digitize the serial numbers on Reader's Digest coupons returned from advertisements. The fonts used on the documents were printed by an RCA Drum printer using the OCR-A font. The reader was connected directly to an RCA 301 computer (one of the first solid state computers). This reader was followed by a specialized document reader installed at TWA where the reader processed Airline Ticket stock. The readers processed documents at a rate of 1,500 documents per minute, and checked each document, rejecting those it was not able to process correctly. The product became part of the RCA product line as a reader designed to

process "Turn around Documents" such as those utility and insurance bills returned with payments.

The United States Postal Service has been using OCR machines to sort mail since 1965 based on technology devised primarily by the prolific inventor Jacob Rabinow. The first use of OCR in Europe was by the British General Post Office (GPO). In 1965 it began planning an entire banking system, the National Giro, using OCR technology, a process that revolutionized bill payment systems in the UK. Canada Post has been using OCR systems since 1971.

In 1974 Ray Kurzweil started the company Kurzweil Computer Products, Inc. and led development of the first omni-font optical character recognition system — a computer program capable of recognizing text printed in any normal font. He decided that the best application of this technology would be to create a reading machine for the blind, which would allow blind people to have a computer read text to them out loud. This device required the invention of two enabling technologies — the CCD flatbed scanner and the text-to-speech synthesizer.

1992-1996 Commissioned by the U.S. Department of Energy (DOE), Information Science Research Institute (ISRI) conducted the most authoritative of the **Annual Test of OCR Accuracy** for 5 consecutive years in the mid-90s. Information Science Research Institute (ISRI) is a research and development unit of University of Nevada, Las Vegas. ISRI was established in 1990 with funding from the U.S. Department of Energy. Its mission is to foster the improvement of automated technologies for understanding machine printed documents.

One study based on recognition of 19th and early 20th century newspaper pages concluded that character-by-character OCR accuracy for commercial OCR software varied from 71% to 98%; total accuracy can only be achieved by human review. Other areas—including recognition of hand printing, cursive handwriting, and printed text in other scripts (especially those East Asian language characters which have many strokes for a single character)—are still the subject of active research.

2.2 List of OCR softwares

2.2.1 Free & Online

All these free OCR software are completely online, so you do not have to download anything to use. These free online OCR software let you just upload your scanned images, and then download the converted text. Some of these free OCR services also let you preserve your formatting.

Free Online OCR

As the name suggests, it is a free online OCR :) This is the best free OCR It lets you upload documents in multiple formats: JPG, BMP, TIFF, and PNG. Also, this is one of the few OCR software that supports uploading PDF as well. Apart from multiple input formats, it also supports multiple output format. It can convert your scanned files to Word, PDF, RTF, and TXT format. Just upload your file, choose the format to convert to, and this free OCR software will immediately do the conversion. You can then download the converted file.

i2OCR

i2OCR is another one of my favorite OCR software. This is also completely online, and does not requires you to create any account, or download anything. You can upload any scanned image, and it will extract text from image. Here are some of the features that I like in this free online OCR.

After conversion, it shows the extracted text online, and shows your uploaded image side-by-side. So, you can quickly see that the conversion was accurate. If you see any problems, you can try uploading a higher resolution image.

This is a multi-language OCR software, and supports 33 languages. Apart from English, it can read French, Spanish, Italian, German, Russian, Dutch, Swedish, Finnish, Danish, Greek, Turkish, Norwegian, Polish, Slovakian, Czech, Latvian, Lithuanian, Hungarian, Catalan, Tagalog, Serbian, Slovenian, Romanian, Ukrainian, Portuguese, Bulgarian, Indonesian, Japanese, Korean, Vietnamese, Chinese. That is quite impressive list of supported languages

You can also provide URL of an image, and it will extract text from that image.

OnlineOCR.Net

OnlineOCR is another free OCR software. It also works in the same manner as the OCR software mentioned earlier. The good part is that it can also convert scanned image to XLS format. So, if you received a scanned copy of a spreadsheet, you can convert them to XLS using this OCR software. You can also upload multiple page PDF files, and convert then to Word, TXT, or XLS format. The free version limits 15 images per hour.

BetterOCR

BetterOCR is another free OCR software that is completely online. A great feature that it provides is that you can even upload zip files. It will extract your scanned image from that, and then convert it to text. Apart from zip, it also supports PNG, JPG, GIF, TIF, and PDF format. There is no limit on number of images that you can convert with this free online OCR software.

FreeOCR

Free OCR is last free OCR software in our list. It supports all the common formats that we mentioned above, like, JPG, GIF, TIFF BMP or PDF. The reason it is last in our list is because:

- If you upload a PDF file, it converts only first page
- There is a size limit of 2MB per uploaded file
- Only 10 images can be converted per hour.

2.2.2 Paid softwares

For Business

Omnipage Professional 18

- Automatically batch convert files.
- Monitor, recognize, and convert files from incoming e-mail.
- Archive documents directly into Microsoft SharePoint.
- Includes PaperPort 11 and PDF Create.

OmniPage Enterprise 18

Turn high volumes of paper and digital documents into files you can edit, search and share in the format of your choice.

For Individuals

Omnipage 18 Standard

- The most accurate conversion in 123 languages.
- Superior formatting control.
- Complete recognition of forms, text, tables, graphics, and images.

For Developers

OmniPage Capture Software Developers Kit

- Everything you need for scanning, OCR, ICR, OMR, PDF, and document conversion.
- Integrated PDF toolkit including searchable PDF and patented PDF-MRC
- Supports Windows, Linux, Macintosh, mobile and embedded OCR development.

2.3 Overview of technologies used in OCR software

2.3.1 Introduction to Digital Image

Pixels

A digital image is usually a rectangular grid comprised of individual pixels *(picture element or PEL)*. A good analogy might be a tile mosaic, with the smallest element in the mosaic being the individual tiles *(each of which is one color or shade)*. Each pixel in a digital image has a bit-depth value, which informs the computer which color *(or shade of gray)* the pixel will display *(the greater the bit-depth value, the more colors/grays to choose from)*. The combined effect of all the individually colored pixels creates the image.

Resolution

The number of pixels in an image is often used as a way to describe the image's resolution. The word resolution has a specific technical meaning to microscope users, namely the ability to distinguish between two closely adjacent objects at a given magnification. In the context of digital images, the word resolution usually refers to how frequently an object was sampled.

The same object sampled at three different pixel densities

3x6 13x25 25x50

Figure 2.3.1. 5 Resolution

Image resolution is often confused with the resolution of the output device *(computer monitor or printer)*. Output devices typically express their resolution in dots/inch *(DPI)*. Digital imaging software programs *(e.g., Adobe Photoshop™)* often set their scale factors based on the monitor resolution *(72 DPI)*, however, this setting is really only useful for images that will ultimately be displayed on a monitor *(WWW pages)*

Color

The most commonly used color model is RGB *(Red, Green, Blue)* for on-screen color. RGB is an additive color model, the three different phosphors on the monitor screen are excited at different intensities *(usually an 8-bit range for each color, for 256 intensities per color, for a total of 16.7 million possible color combinations)* and based on the mix of the three intensities the eye perceives a color.

2.3.2 Image Enhancement Techniques

Image enhancement is basically improving the interpretability or perception of information in images for human viewers and providing `better' input for other automated image processing techniques. The principal objective of image

enhancement is to modify attributes of an image to make it more suitable for a given task and a specific observer. During this process, one or more attributes of the image are modified. The choice of attributes and the way they are modified are specific to a given task. Moreover, observer-specific factors, such as the

human visual system and the observer's experience, will introduce a great deal of subjectivity into the choice of image enhancement methods. There exist many techniques that can enhance a digital image without spoiling it. The enhancement methods can broadly be divided in to the following two categories:

1. Spatial Domain Methods
2. Frequency Domain Methods

In spatial domain techniques, we directly deal with the image pixels. The pixel values are manipulated to achieve desired enhancement. In frequency domain methods, the image is first transferred in to frequency domain. It means that, the Fourier Transform of the image is computed first. All the enhancement operations are performed on the Fourier transform of the

image and then the Inverse Fourier transform is performed to get the resultant image. These enhancement operations are performed in order to modify the image brightness, contrast or the distribution of the grey levels. As a consequence the pixel value (intensities) of the output image will be modified according to the transformation function applied in the input values.

2.3.3 Pattern Recognition and Clustering routines

Pattern recognition is the assignment of a label to a given input value. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of *classes*. Pattern recognition algorithms generally aim to provide a reasonable answer for all possible inputs and to do "fuzzy" matching of inputs. This is opposed to *pattern matching* algorithms, which look for exact matches in the input with pre-existing patterns.

What is Clustering?

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.
A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”.
A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.
We can show this with a simple graphical example:

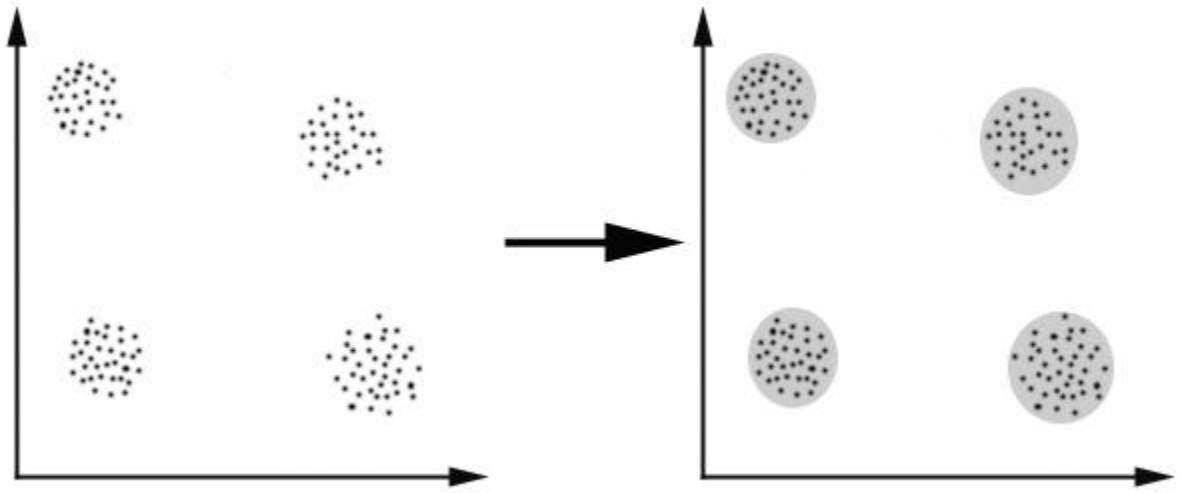


Figure 2.3.3. 6 Clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called *distance-based clustering*. Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

The Goals of Clustering

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.

For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*).

2.3.4 Thresholding Method

During the thresholding process, individual pixels in an image are marked as "object" pixels if their value is greater than some threshold value (assuming an object to be brighter than the background) and as "background" pixels otherwise. This convention is known as threshold above. Variants include threshold below, which is opposite of threshold above; threshold inside, where a pixel is labeled "object" if its value is between two thresholds; and threshold outside, which is the opposite of threshold inside Typically, an object pixel is given a value of "1" while a background pixel is given a value of "0." Finally, a binary image is created by coloring each pixel white or black, depending on a pixel's labels.

Threshold selection

The key parameter in the thresholding process is the choice of the threshold value (or *values*, as mentioned earlier). Several different methods for choosing a threshold exist; users can manually choose a threshold value, or a thresholding algorithm can compute a value automatically, which is known as *automatic thresholding*. A simple method would be to choose the mean or

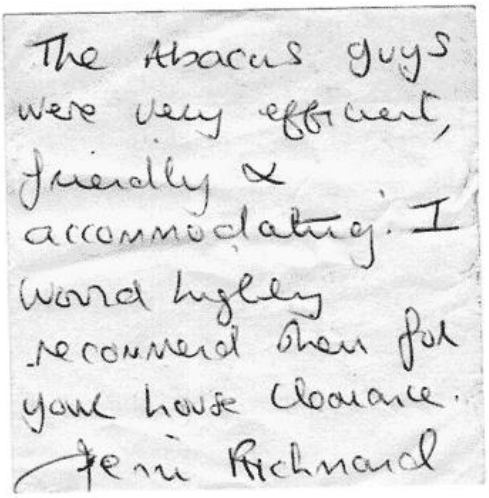


Figure 2.3.4. 1 Source Image

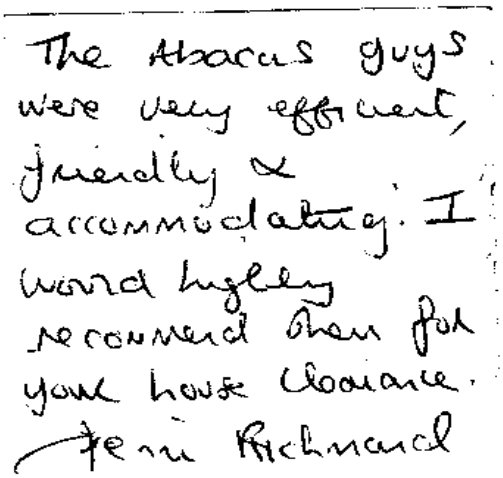


Figure 2.3.4. 2 Thresholding

median value, the rationale being that if the object pixels are brighter than the background, they should also be brighter than the average. In a noiseless image with uniform background and object values, the mean or median will work well as the threshold, however, this will generally not be the case. A more sophisticated approach

might be to create a histogram of the image pixel intensities and use the valley point as the threshold. The histogram approach assumes that there is some average values for both the background and object pixels, but that the actual pixel values have some variation around these average values. However, this may be computationally expensive, and image histograms may not have clearly defined valley points, often making the selection of an accurate threshold difficult.

2.3.5 Edge Detection

Edge detection is a fundamental tool in image processing, machine vision and computer vision, particularly in the areas of feature detection and feature extraction, which aim at identifying points in a digital image at which the image brightness changes sharply or, more formally, has discontinuities.

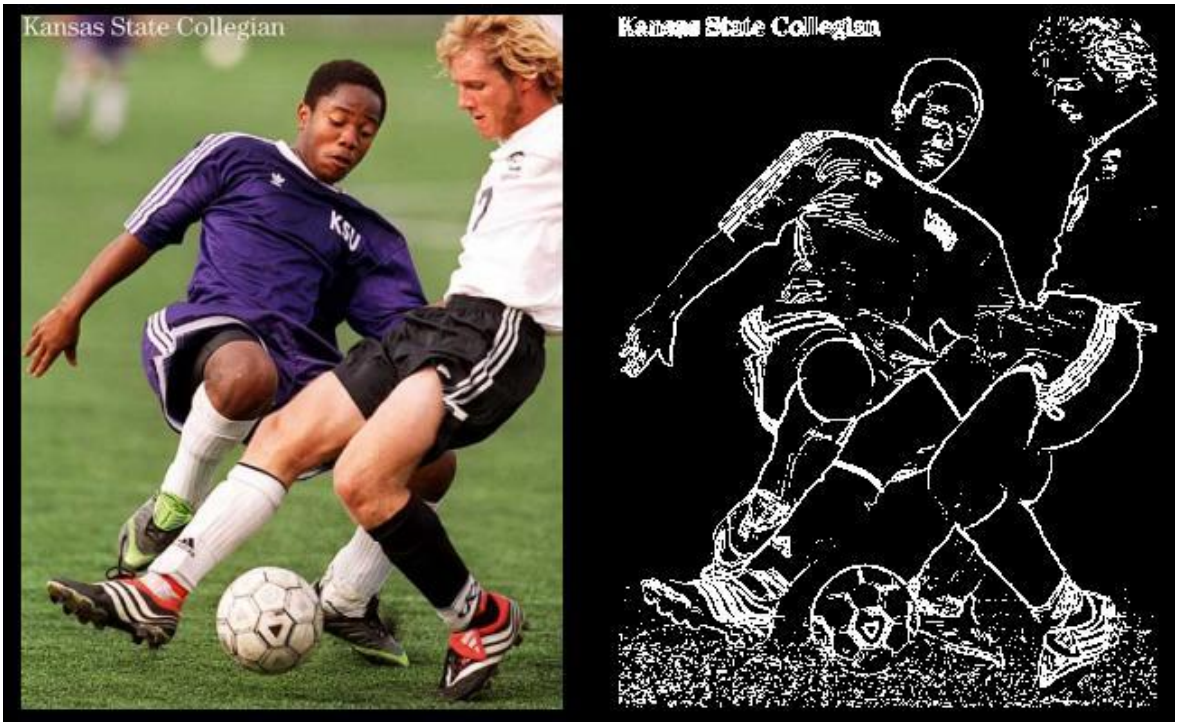


Figure 2.3.5 Edge Detection

Motivation

The purpose of detecting sharp changes in image brightness is to capture important events and changes in properties of the world. It can be shown that under rather general assumptions for an image formation model, discontinuities in image brightness are likely to correspond to:

- discontinuities in depth,
- discontinuities in surface orientation,
- changes in material properties and
- variations in scene illumination.

In the ideal case, the result of applying an edge detector to an image may lead to a set of connected curves that indicate the boundaries of objects, the boundaries of surface markings as well as curves that correspond to discontinuities in surface orientation. Thus, applying an edge detection algorithm to an image may significantly reduce the amount of data to be processed and may therefore filter out information that may be regarded as less relevant, while preserving the important structural properties of an image. If the edge detection step is successful, the subsequent task of interpreting the information contents in the original image may therefore be substantially simplified. However, it is not always possible to obtain such ideal edges from real life images of moderate complexity.

Chapter – 3

Problem Statement

Chapter - 3

Problem Statement

Constraint: Oriya documents only

Constraint: No composite characters

3.1 Overview of technologies used in OCR software

3.1.1 Introduction to Digital Image

A digital image is a numeric representation (normally binary) of a two-dimensional image. Depending on whether the image resolution is fixed, it may be of vector or raster type. Without qualifications, the term "digital image" usually refers to raster images also called bitmap images.

3.1.2 Pattern Recognition and Clustering routines

In machine learning, pattern recognition is the assignment of a label to a given input value. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of classes (for example, determine whether a given email is "spam" or "non-spam"). However, pattern recognition is a more general problem that encompasses other types of output as well. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values (for example, part of speech tagging, which assigns a part of speech to each word in an input sentence); and parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence.

Chapter – 4

Analysis

Chapter - 4

Analysis

4.1 Analysis of English characters

Unique Features of English Characters :

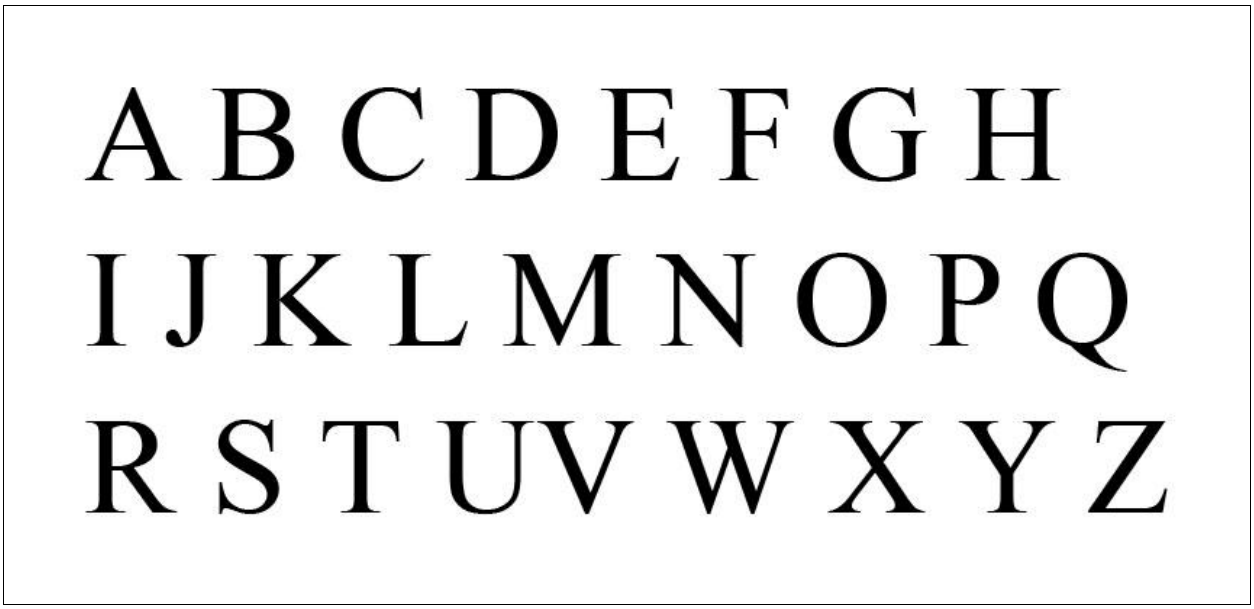


Figure 4.1 English Characters

The English language consists of 26 characters only. It has basically two sets i.e. Capital Letters and small Letters. In the view of Character Recognition, the English characters are easy to extract from a document since they don't have any identifiers. Even the English characters are very simple and plain. That is they have fixed type of patterns and curves. For example, for the character O and Q, they both have a circle curve. And further considering the characters L and I, they simply have a vertical straight line. So using these features of curve, we can extract characters by using the pattern matching algorithms.

4.2 Difference between English and Oriya characters

Generally, The English language has 26 characters and the oriya language has 45 characters. Moreover, the English characters are very simple and plain. On the other hand, for oriya characters the complexity of the characters vary with different set of words. The English characters has a very few number of patters that needs to be recognized, but oriya has a wide range of features like different curves and different patterns. In addition, there are different identifiers which make the recognition process a bit complex.

4.3 Classification of Oriya Characters

4.3.1 Phonetic classification: Vowels, Consonants, composite-characters

- The direction of writing Oriya script is from left to right in horizontally.
- In Oriya script, there is no distinction between Upper Case and Lower Case characters.
- Oriya script has its own specified composition rules for combining vowels, consonants and modifiers.

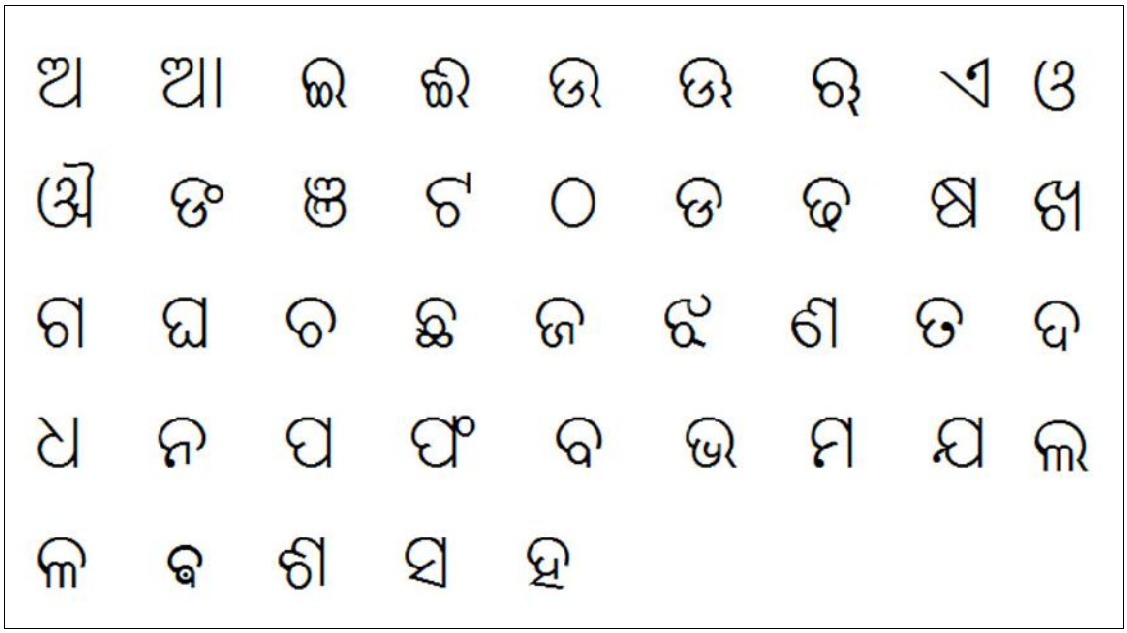


Figure 4.3.1. 1 Consonants of the Oriya Language

- Modifiers are attached to the top, bottom, left or right side of the other characters.

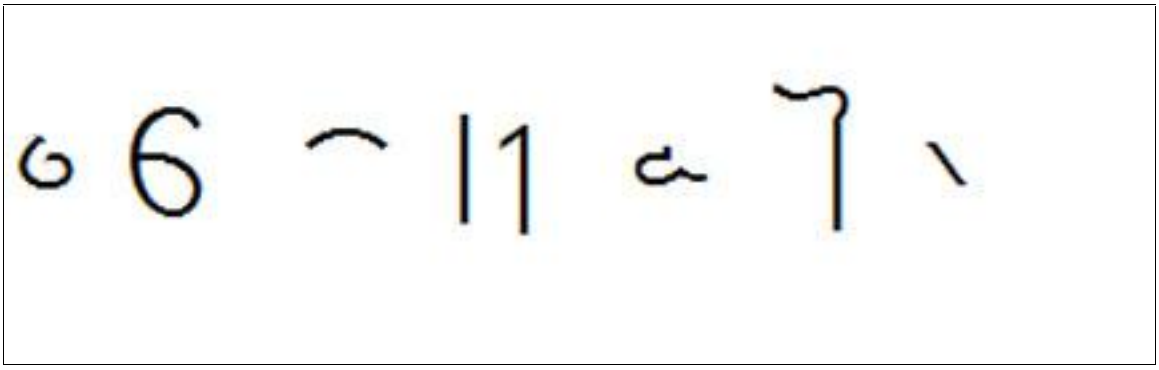


Figure 4.3.1. 2 **Vowels of the Oriya Language**

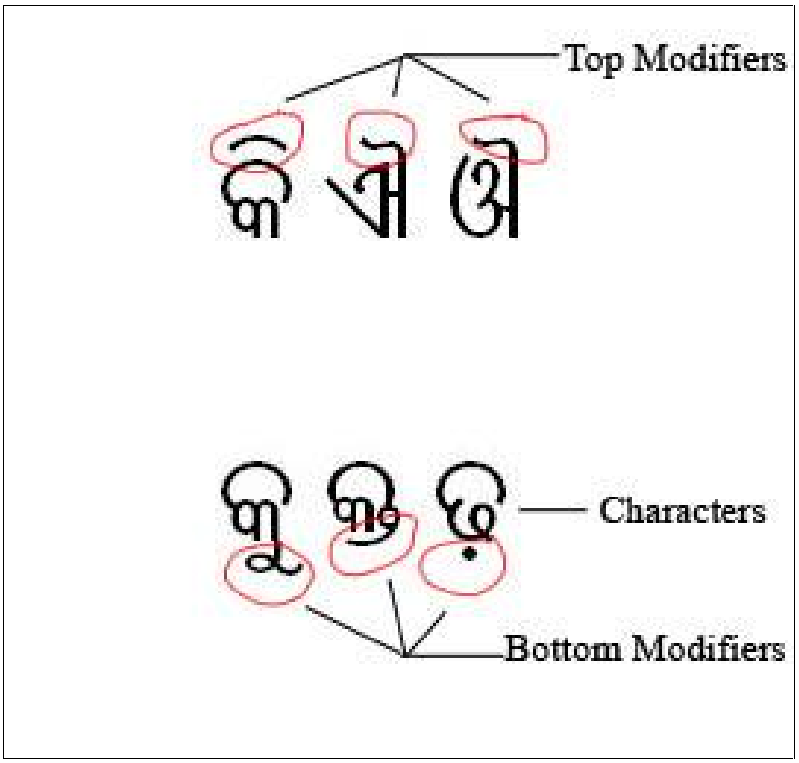


Figure 4.3.1. 7 **Identifiers Usage**

- Top Modifiers are placed above the characters and the bottom modifiers are placed below the characters

- To Indicate the end of sentence or a phrase, use of a single vertical line called “Purnacheda” is used

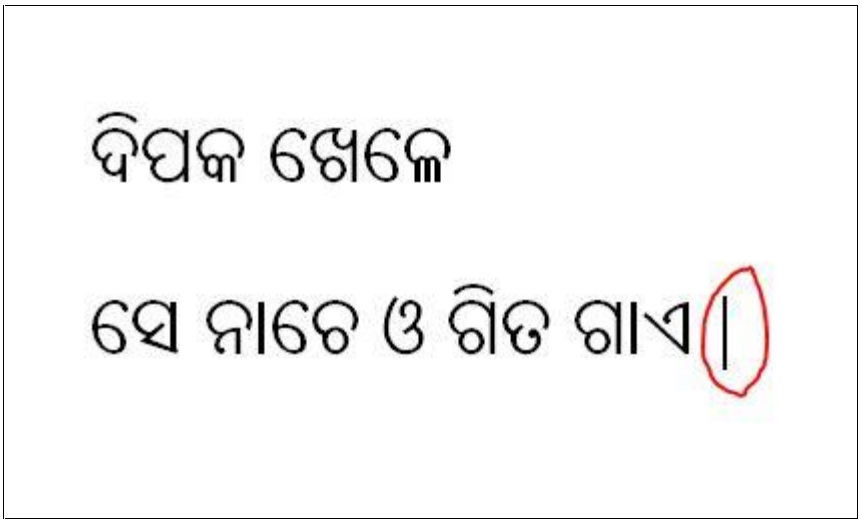
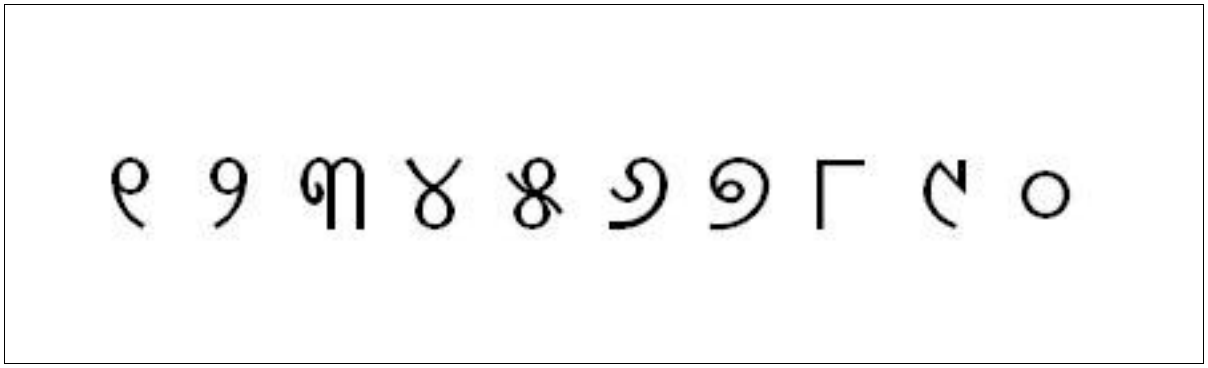


Figure 4.3.1. 4 Usage of Purnacheda

- Oriya script has a native set of symbols for numerals



4.3.2 Classification based on geometric features and character width

Large circular features

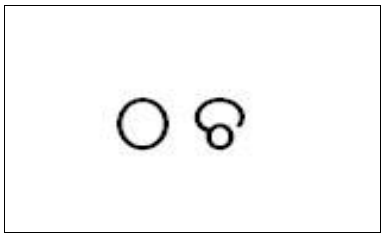


Figure 4.3.2. 1 Large Circular Features

There are certain characters which have some particular features. Like the letters ‘tha’ and ‘cha’ have circular features and those are the most significant features of the characters.

Dominant linear features

Apart from the characters which possess circular features, there are other characters which also posseess some other characteristics or features. Such as the characters ‘kha’ and ’sa’ have dominant vertical line like features. And this is the most prominent and visible features of these characters.

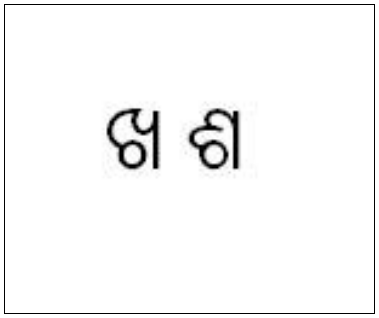


Figure 4.3.2. 2 Dominant Features

Mixed features

Apart from the above mentioned two categories of characters, there are some new and different kind of characters, which has a mixture of two or more characters, so as to represent a word or pronunciation more effectively. Like there can be a single character ‘sa’ and below that another oriya characters ‘pa’ can be joint so that it forms the word ‘spa’. This further increases the complexity for the OCR to recognize the characters from the document.

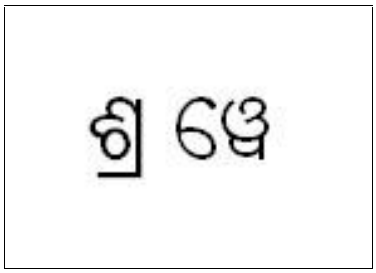


Figure 4.3.2. 8 Mixed Features

Section-2

The OCR Process

Chapter – 5

Main CR Process

Chapter - 5

Main CR Process

5.1 Preprocessing

5.1.1 Removing Noise

The input image that is used to recognition of characters is being processed. First is the removal of noise. Noise is removed by using the threshold process. IN the threshold process, we generally, scan each pixels and count for a specific or thershehold intensity value, and the value greater than that is accepted and the values less than that are rejected or are not selected. Thereby noise is removed because the noise are always the pixels with lesser intensity or doesn't have significant intensity in the input image

5.1.2 Increasing Contrast

The input image after removal of the noise, is further processed for the increasing of the contrast of the image. Basically, we scan each and every pixels and retrieveing the intensity values , we simply add some more values/numbers so as to increase the intensity values higher than that of the old values. Thereby, the image which is relatively darker than normal appears to be brighter and with good contrast feature is obtained. And hence, we can have a crisp and clear image

5.2 Segmentation

Preliminary Segmentation of Words

- When horizontal projection is computed, the row containing maximum number of black pixel is considered to be the header line.
- A vertical projection is made of the given image starting from the top of the image to the bottom row of the word image box. The columns with no black pixels are treated as boundaries. Thus, the character/symbol box can be separated.

5.2.1 Line Detection

Figure 5.2.1. 9 Line count

The input image needs to undergo through a segmentation process, so that the features can be extracted. In the first process, the line detection algorithm is used. This algorithm normally, calculates the number of dark intensities per scan line and the values per row are plotted and a graph is obtained(in the above fig.) The areas where the highest intensity peak is obtained and is stored in a vector. After that there is a process to check the number of transitions in the graph, containing set of black pixels and white pixels. The image starts with a white pixel scanline and after scanning further set of black pixels are obtained too. Then the transitions obtained are done a XOR operation so that if the two input values are of same intensity i.e. either black or white, then the transitions are discarded and if different intensities are discovered then the transitions are stored. And after processing the transition values, we obtain the number of lines.

INPUT		OUTPUT
A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

5.2.2 Word Detection

For word detection, the same algorithm for the line detection is used except for the fact that they are implemented in a vertical direction. That is, the line detection loop scans along the rows but word detection goes along the columns. If there are white pixels in between certain characters then the characters within the two white pixel areas are considerred to be a word. In addition to that, we normally calculate the average character width to avoid some typographic errors. The diagram below demonstrates the fact. In this figure, there is a

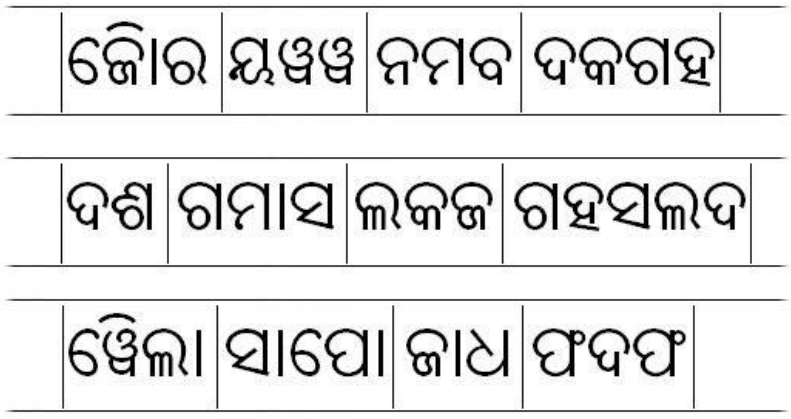


Figure 5.2.2. 1 Word Detection

typographic error, that is a space is entered by mistake. And if this image is used for recognition then the program may detect as three words. But due to average character width which is calculated previously. It then checks for these type of errors that creep in due to the typing. So it analyses that the given space is not to be considered as a separate word, rather it's 2 words.

5.3 Feature detection

5.3.1 Scaling to standard Size

Scaling

Image scaling is the process of resizing a digital image. Scaling is a non-trivial process that involves a trade-off between efficiency, smoothness and sharpness. As the size of an image is increased, so the pixels which comprise the image become increasingly visible, making the image appear "soft". Conversely, reducing an image will tend to enhance its smoothness and apparent sharpness.

Apart from fitting a smaller display area, image size is most commonly decreased (or subsampled or downsampled) in order to produce thumbnails. Enlarging an image (upsampling or interpolating) is generally common for making smaller imagery fit a bigger screen in fullscreen mode, for example. In "zooming" an image, it is not possible to discover any more information in the image than already exists, and image quality inevitably suffers. However, there are several methods of increasing the number of pixels that an image contains, which evens out the appearance of the original pixels.

Scale Up

If the image from which the character is to be extracted is of small size or the font size is very small to be able to detect. Then using the scaling algorithm, we scale up the images to a fixed dimensions. The input data is first resized a height of 128 pixels to satisfy the procedure,

regardless of whether the image is of single character or a word. It should be noted that no skew correction was done, ensuring that the scanning process is of high quality.Basically, while scaling up the images, a particular pixel contriubtes to a multiple no of pixels in the output image. If there is a single black intensity in the input image, then the corresponding black pixel is contributed to a number of pixels in the output image. The figure below demonstrates the application

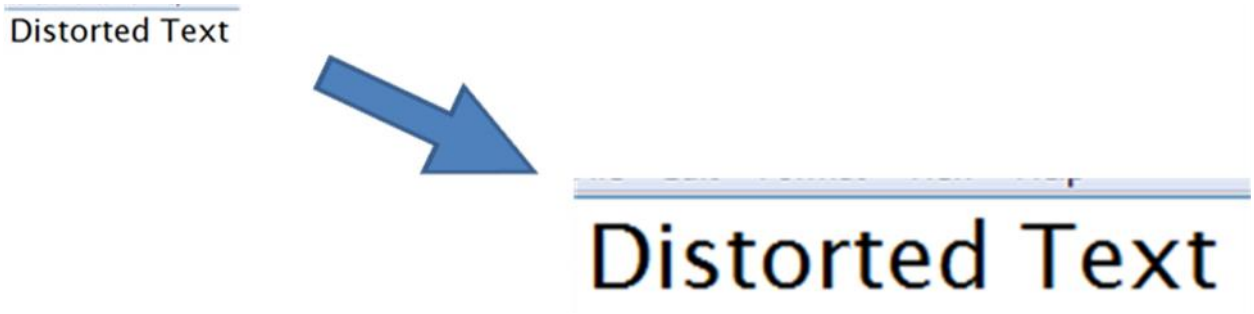


Figure 5.3.1. 1 Scaling down

Scale down :

If the image from which the character is to be extracted is of large size or the font size is very large be able to detect. Then using the scaling algorithm, we scale down the images to a fixed dimensions. The input data is first resized a height of 128 pixels to satisfy the procedure, regardless of whether the image is of single character or a word. It should be noted that no skew correction was done, ensuring that the scanning process is of high quality. Basically, while scaling down the images, a multiple pixels contriubtes to a single pixel in the output image. If there is are multiple black intensities in the input image, then the corresponding black pixels are contributed to a single pixel in the output image. The figure below demonstrates the application

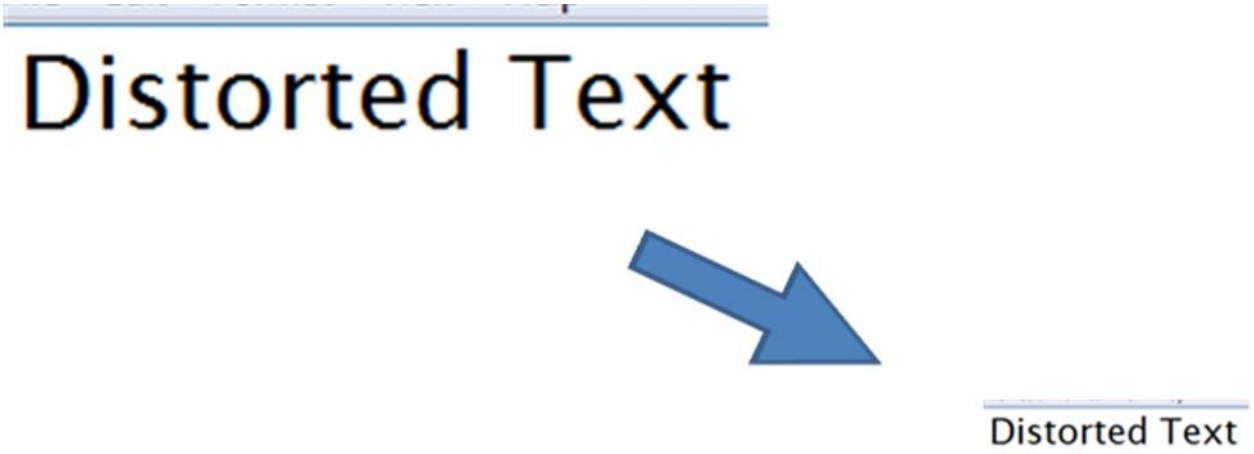


Figure 5.3.1. 2 scale of scale down

5.3.2 Template Matching

After the processing of the images, the words are extracted out of it. And the extracted words are generally mapped on to the the templates that are present in the databases. The process is that each pixel of the word is mapped and the maximum mapping percentage is calculated.

$$fitness = \sum_{r=0}^{h_T} \sum_{c=0}^{w_T} (T_{r,c} - 127)(I_{r,c+i} - 127) \; for \; i = 0 \ldots (w_I - w_T)$$

where T represents template and I represents Image
For each template, get i and fitness corresponding to heighest fitness.

Suppose in the input image it consists of the character O and the template matching is done with Q, then if there is matching of a single black intensity in both the input image and from the database, then calculated value has the maximum value. Similarily, if there are white intensity pixels are matched that even results in maximum output value is generated. Thereby ensuring us that a particular character has a maximum matching.

5.3.3 Inference

5.3.4 Printing Output

After the character Recognition process is over, then an output is generated as a text file. The set of characters that are present in the printed document in oriya language are detected and are then with maximum matching, one by one characters are updated In the text file. And finally output file is obtained with a .txt extension and that can be opened up by the Notepad software and can be easily manipulated or can be edited further to correct the errors manually, or make any necessary changes, if required.

Chapter – 6

Database generation process

Chapter - 6

Database generation process

6.1 Producing character map for standard Oriya Font

Before the character recognition process begins, we have to create a database containing all the set of characters, so that it would be very handy to map the characters from the image to that from our database. Further database of different set of languages can be created, which would enhance to create the CR process for another language.

So for this process, we collect or write all the characters of the Oriya Lanuage in a document with a standard font size so that it is visible properly to the naked eyes and then take a snapshot of it, and save it in PNG format, so that we can manipulate the file using our own library function that we have developed.

We need all the characters because for the matching from the input image, it has to match with each and every character of the Oriya Language.

6.2 Extracting high quality images for individual glyphs

Once after the snapshot of the image is obtained, the next part of the process is to extract each and every character from the snapshot containing all characters of the Oriya Language. The cropping algorithm is used with different dimensions and size and accordingly the characters are extracted. And also the different glyphs are also extracted.

6.3 Standardizing bitmaps of glyphs to be used as templates

The next process is to standardize the bitmaps of glyphs that is to be used as templates to a fixed dimensions that is 128 pixels by 128 pixels, which enables the proper matching of the characters that is scanned from the input image. Basically, the templates are of good resolutions, so that

while matching accurate results are obtained. Moreover, necessary scaling are done on requirements.

Section-3

Testing and Inferences

Chapter – 7

Test Cases and Results

Chapter - 7

Test Cases and Results

We tested our project with different set of images taking those as our samples for recognition of characters from the documents and obtaining the output for the same. The sample image that we took is below :

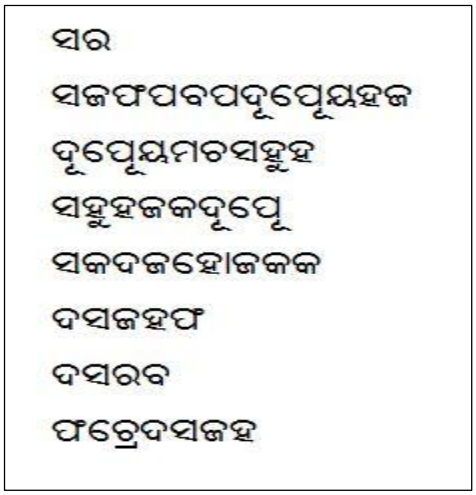
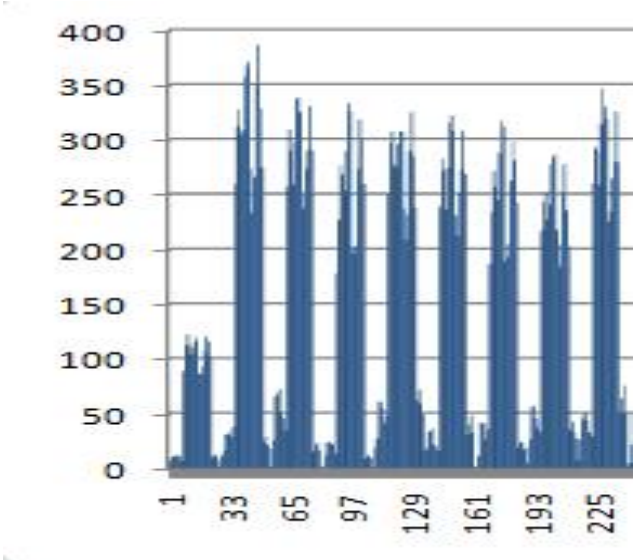


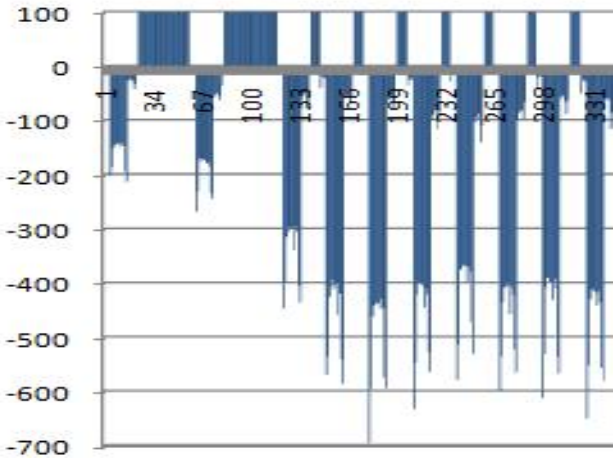
Figure 7. 1 Sample Text

The first process that executes is Line detection, where it detects the number of lines present in the image, by plotting a graph of maximum intensities along a single row. And this is done for the total pixels contained in the whole image. And the other graph is the extended version of the first graph, i.e. we plot the lower intensities (black pixels values) along the negative axis of the Y axis, and the spaces are plotted in the positive direction of Y axis so as to distinguish or to find the difference between the black and white pixels or in other words, the words and the character spaces respectively.

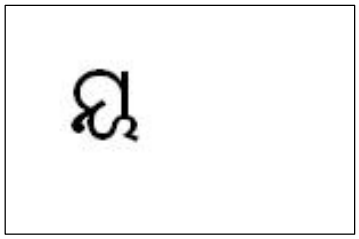


The figure shows that the intensities scanned along the rows are the number of the lines in the image.

The image represents the number of spaces in the positive direction and the number of the words along the negative direction



After obtaining the plots, we then crop the images/characters so that the cropped version of the text can be used to perform the template matching to get accurate results.



The cropped version is obtained as follows :

The next process involves template matching:

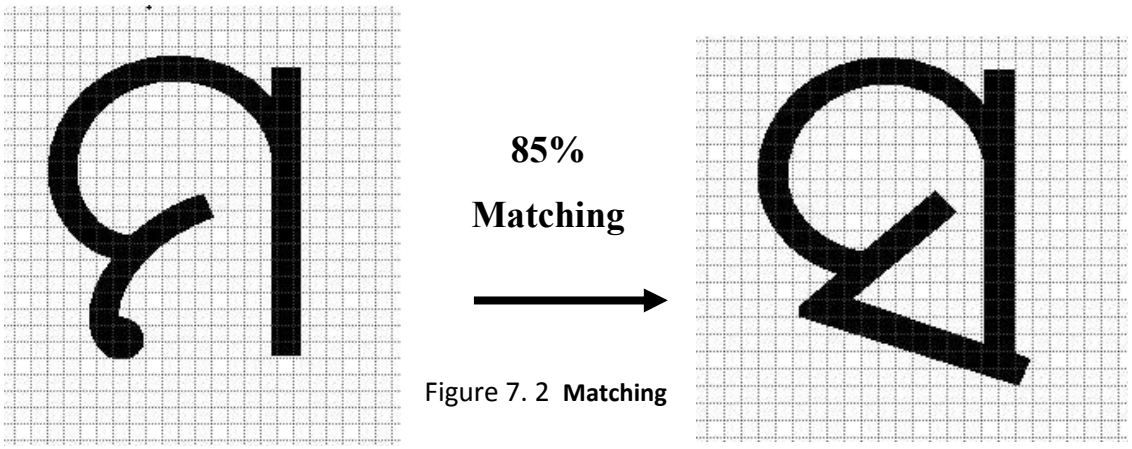


Figure 7. 2 **Matching**

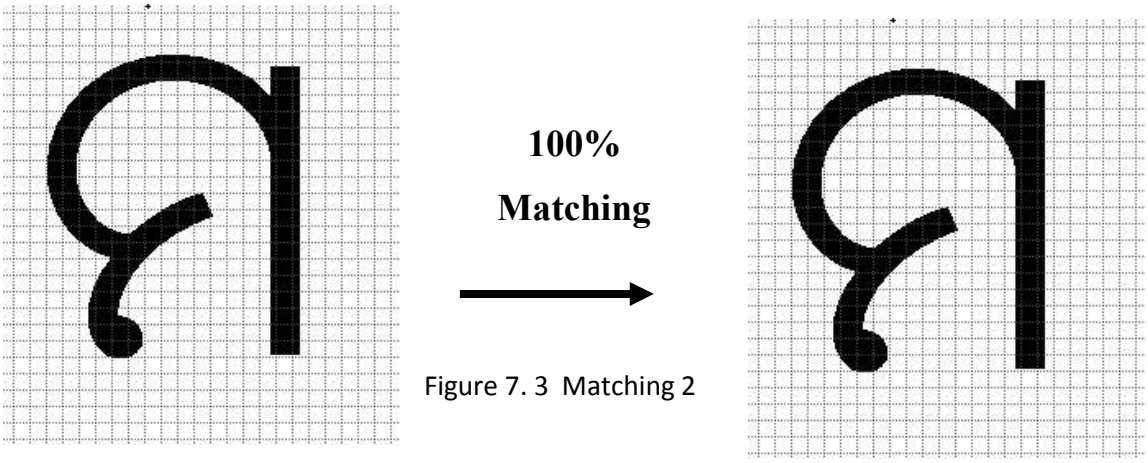


Figure 7. 3 Matching 2

The next step is the template matching. In this process, an image/character extracted from the scanned document is matched with the template of all the oriya characters and the maximum matching percentage is obtained. The character which matches with maximum percentage is said to be that letter from the database, and if there is lesser percentage of matching, then it is moved on to next character for matching and if it matches cent percentage then it stops , and confirms for the particular oriya letter.

The whole image is processed and characters are extracted and matched as stated above and finally the obtained characters are finally written in a .txt file, which can be modified for further purposes according to the need.

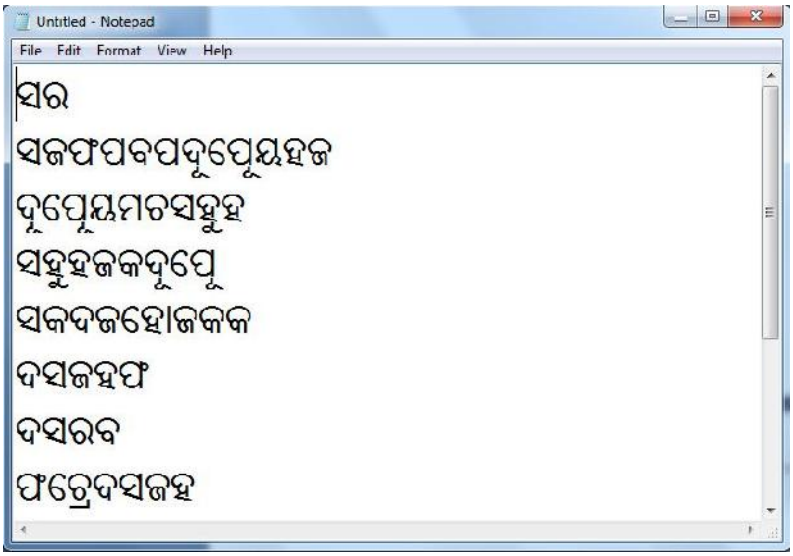


Figure 7. 4 **Output of the image**

Chapter – 8

Conclusion

Chapter - 8

Conclusion

The internship program was a great opportunity for us to experience the workings of software industry and bring our four years of academics to practice. The three months period of internship was very useful for learning the industry specific practices and standards. We are very confident that this experience will be a great boost and help for our career ahead.

The internship period was a great learning experience with very helpful project supervisor. Their evaluation and guiding helped us a lot to learn and understand new techniques and methodologies. Overall, it was a great experience and will greatly benefit us in our future.

Chapter – 9

Future Works

Chapter - 9

Future Works

Our project includes only the recognition of the characters from the printed oriya documents only. The future works are that we will develop similar approach to recognize the characters from a hand written oriya documents. For that we will use curve detection algorithms and many other feature extraction process, which will detect the specific curves, or distinct elongated features or circles. And the recognition would be such supreme that even if someone has just scribbled something, our Character Recognition Process will recognize accurately each and everything. And furthermore, we will use a dictionary embedded to it, so that after characters are detected it'll check for the existence of words from the dictionary itself and make it error free and thereby making it more robust and powerful.

OPTICAL CHARACTER RECOGNITION

BIBLIOGRAPHY

Websites

1. http://www.nuance.com/for-business/by-product/omnipage/index.htmSimple OCR FAQ - dictionaries (http://www.simpleocr.com/Info.asp#Dictionary)
2. http://office.microsoft.com/en-us/help/about-optical-character-recognition-ocr-HP003081255.aspx
3. http://www.dataid.com/aboutocr.htm
4. http://www.ehow.com/how-does_4963233_ocr-work.html
5. http://www.explainthatstuff.com/how-ocr-works.html

BOOKS

1. Holley, Rose (April 2009). "*How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs*". D-Lib Magazine. Retrieved 5 January 2011
2. Tappert, Charles C., et al (1990-08). *The State of the Art in On-line Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 12 No 8, August 1990, pp 787-ff*. Retrieved 2008-10-03