PROJECT - I

Report on

OPTICAL CHARACTER RECOGNITION

Submitted in partial fulfillment of the requirements

of the degree of

Bachelor of Engineering
(Electronics and Telecommunication Engineering)

by

Tawade Shubham Ratnoji (15ET1105)
Singh Vishal Kumar (15ET1138)
Sonawane Riddesh Rajesaheb(15ET1043)

Supervisor

Mr. Jaswantsing Rajput

D Y PATIL
RAMRAO ADIK
INSTITUTE OF
TECHNOLOGY
NAVI MUMBAI

Department of Electronics and Telecommunication Engineering
Ramrao Adik Institute of Technology,
Sector 7, Nerul , Navi Mumbai
(Affiliated to University of Mumbai)
November 2018

D Y PATIL
RAMRAO ADIK
INSTITUTE OF
TECHNOLOGY
NAVI MUMBAI

Ramrao Adik Education Society's

Ramrao Adik Institute of Technology

(Affiliated to the University of Mumbai)

Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400 706.

Certificate of Approval

This is to certify that, the Project -I report entitled

"OPTICAL CHARACTER RECOGNITION"

is a bonafide work done by

Tawade Shubham Ratnoji (15ET1105)
Singh Vishal Kumar (15ET1138)
Sonawane Riddesh Rajesaheb(15ET1043)

and is submitted in the partial fulfillment of the requirement for the degree of

Bachelor of Engineering
(Electronics and Telecommunication Engineering)
to the
University of Mumbai.

Examiner

Supervisor

Project Coordinator

Head of Department

Principal

D Y PATIL
RAMRAO ADIK
INSTITUTE OF
TECHNOLOGY
NAVI MUMBAI

Ramrao Adik Education Society's
Ramrao Adik Institute of Technology
(Affiliated to the University of Mumbai)
Dr. D. Y. Patil Vidyanagar,Sector 7, Nerul, Navi Mumbai 400 706.

Declaration

We wish to state that work embodied in this dissertation entitled "**OPTICAL CHARACTER RECOGNITION**" has been carried out under the guidance of Mr. Jaswantsing Rajput at Department of Electronics and Telecommunication Engineering, Ramrao Adik Institute of Technology during 2018-2019.

We declare that the work being presented forms our own contribution and has not been submitted for any other Degree or Diploma of any University/Institute. Wherever references have been made to previous works of others, it has been clearly indicated. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Tawade Shubham Ratnoji

Singh Vishal Kumar

Sonawane Riddesh Rajasaheb

Acknowledgments

We are very glad to thank to our project guide **Mr. Jaswantsing Rajput**, our project co-ordinators **Dr. Ashwini Naik** and **Mr. Bhushan Deore** and our **H.O.D Dr. M.D.Patil** for their encouragement and tremendous guidance. Here it is very special thank feelings for our colleagues for their support and help. We have been fortunate to have received many useful suggestions from our colleagues which have greatly improved the clarity of our report. At the end special thanks to our **Principal Dr. Ramesh Vasappanavara**. We would like to appreciate suggestions and criticisms about the report from the readers.

Abstract

In India, more than 310 million people use Devanagari script for documentation. There has been a significant improvement in the research related to the recognition of printed as well as handwritten Devanagari text in the past few years. All feature-extraction techniques as well as training, classification and matching techniques useful for the recognition are used.

In any OCR or classification system extracting discriminating feature is most important and crucial step for its success. Accuracy of such system often depends on the good feature representation. In this project, we have used histogram of individual characters and GLCM(Gray Level Co-occurence Matrix) for feature extraction. In this part, first image is preprocessed to remove noise, converted to binary image, resized to fixed size and then convert to gray scale image using mask operation, it blurs the edges of the images. Finally network weight parameters are fine tuned by supervised back propagation learning to improve the overall recognition performance.

Now-a-days there are many new methodologies required for the increasing needs in newly emerging areas, with these methodologies there are many techniques are present for the character recognition of handprint Devnagari, Bangla, Tamil, China etc. But it has not been used practically. So in this project, we have used a Minimum distance classifier technique for OCR System of printed as well as scanned newsprint Marathi script.

Contents

Abstract **iv**

List of Figures **vi**

1 Introduction to Devanagari Script **1**

2 Optical Character Recognition (OCR) **2**

2.1 Definition of OCR 2

2.2 History of OCR 3

2.3 Need of OCR 3

2.4 OCR for devanagari script 3

3 Working **4**

3.1 Scanning 4

3.2 Pre-Prcoessing 5

3.2.1 Compression 5

3.2.2 Binarization 6

3.3 Image Segmentation 7

3.4 Feature Extraction 7

3.4.1 Histogram of individual characters 8

3.4.2 GLCM(Gray Level Co-occurence Matrix) 9

3.5 Training and Recognition 10

3.6 Post-processing 10

4 Result **12**

5 Future Scope **13**

6 Conclusion **14**

Bibliography **15**

List of Figures

1.1 Consonants in devanagari script 1

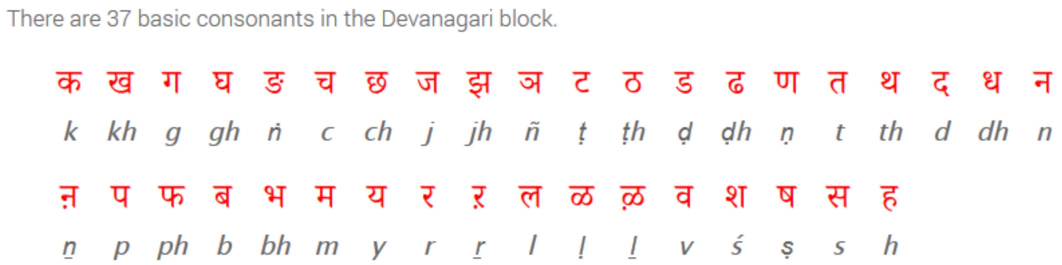
2.1 OCR 2

3.1 Scanning in OCR 4
3.2 Binarization 6
3.3 Image Segmentation 7
3.4 Histogram of individual characters 8
3.5 Features calculated from GLCM 9

Chapter 1

Introduction to Devanagari Script

Devanagari is a Northern Brahmic script related to many other South Asian scripts including Gujarati, Bengali, and Gurmukhi, and, more distantly, to a number of South-East Asian scripts including Thai, Balinese, and Baybayin. The script is used for over 120 spoken Indo-Aryan languages, including Hindi, Nepali, Marathi, Maithili, Awadhi, Newari and Bhojpuri. It is also used for writing Classical Sanskrit texts. Generally the orthography of the script reflects the pronunciation of the language.Hindi is normally spoken using a combination of around 52 sounds, ten vowels, 40 consonants,nasalisation and a kind of aspiration. These sounds are represented in the Devanagari script by 13 characters traditionally regarded as vowels and 40 consonants.



There are 37 basic consonants in the Devanagari block.

क	ख	ग	घ	ङ	च	छ	ज	झ	ञ	ट	ठ	ड	ढ	ण	त	थ	द	ध	न
k	*kh*	*g*	*gh*	*ṅ*	*c*	*ch*	*j*	*jh*	*ñ*	*ṭ*	*ṭh*	*ḍ*	*ḍh*	*ṇ*	*t*	*th*	*d*	*dh*	*n*

ऩ	प	फ	ब	भ	म	य	र	ऱ	ल	ळ	ऴ	व	श	ष	स	ह
ṉ	*p*	*ph*	*b*	*bh*	*m*	*y*	*r*	*ṟ*	*l*	*ḷ*	*ḻ*	*v*	*ś*	*ṣ*	*s*	*h*

Figure 1.1: Consonants in devanagari script

Guideline for writting in devanagari script

i Devanagari characters hang from a horizontal line (called the h ead stroke) written at the top of the character. Unlike English letters which are written up from a line below them.

ii The body of the Devanagari characters should occupy about two thirds of the space between the lines.

iii In general the first stroke, or strokes, in a character are written from the left to the right and are then followed by any down strokes and finally the head stroke is added. Note that in some characters the head stroke is broken.

Chapter 2

Optical Character Recognition (OCR)

2.1 Definition of OCR

OCR Stands for Optical Character Recognition. OCR application is able to recognize and extract text information out of scanned document, such as PDF, TIFF, or other document image files. A PDF Converter with OCR ability can converts scanned PDF document into editable text.

Optical character recognition is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into machine-encoded text. It is extensively used as a tool for data entry from some sort of original paper data source, whether documents, sales receipts, mail, or any number of printed records. It is a general method of digitizing printed texts so that they can be electronically searched, stored more efficently, displayed on-line, and used in machine processes such as machine translation, text-to-speech and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision.

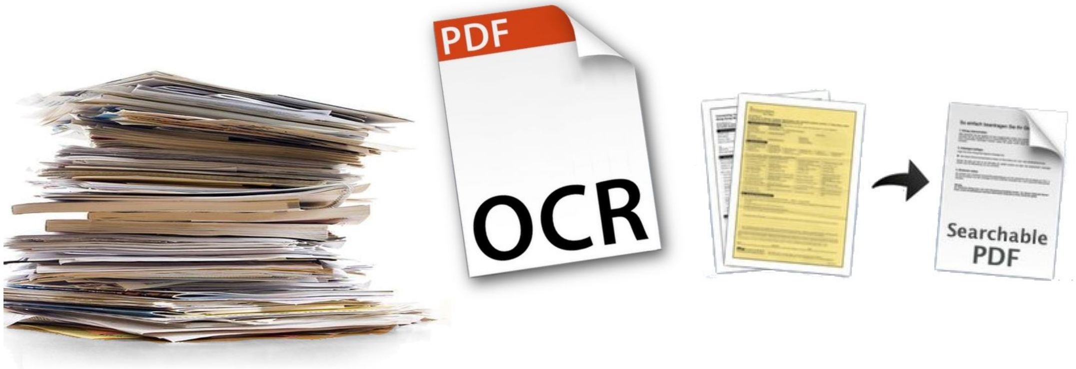


Figure 2.1: OCR

2.2 History of OCR

The very first device with OCR requirements was called as "Optophone".It was developed by Dr.Edmund Fournier d'Able of Birmingham University in the year 1913 which used Selenium photosensor to detect block print and convert it into an audible output which could be interpreted by a blind person.Since then we have made huge progress in OCR technologies.

2.3 Need of OCR

Scanning a document and saving it in an image file format (.tiff, .jpeg, etc.) is similar to taking a picture of a document with a camera or making a copy of it by using a copy machine, but instead you are using a scanner. If you would like to edit the scanned image in your word processor, you will have to perform OCR on the image to convert the image into editable text. OCR is a technology that converts documents that you can read into documents that your computer can read. During the conversion, the document is analyzed, and characters and words are saved as editable text.

2.4 OCR for devanagari script

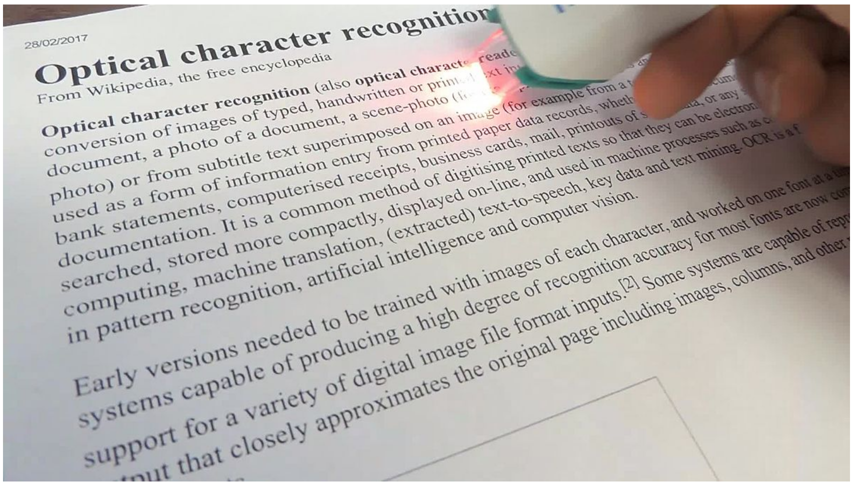
In this project, we have utilized a scheme to develop complete OCR system for particular fonts and sizes of Devnagari characters so that we can use this system for digitizing devanagari document. We have implemented steps of the OCR system like preprocessing, segmentation, feature extraction and classification.

Chapter 3

Working

3.1 Scanning

The first component in OCR is optical scanning. Through scanning process digital image of original document is captured. In OCR optical scanners are used which consist of transport mechanism and sensing device that converts light intensity into grey levels. Printed documents consist of black print on white background. When performing OCR multilevel image is converted into bi-level black and white image. This process known as thresholding is performed on scanner to save memory space and computational effort. The thresholding process is important as the results of recognition are totally dependent on quality of bi-level image.



28/02/2017

Optical character recognition

From Wikipedia, the free encyclopedia

Optical character recognition (also **optical charact** reade
conversion of images of typed, handwritten or print
document, a photo of a document, a scene-photo (f
photo) or from subtitle text superimposed on an image (for example from a t
used as a form of information entry from printed paper data records, wheth
bank statements, computerised receipts, business cards, mail, printouts of s
documentation. It is a common method of digitising printed texts so that they can be electron
searched, stored more compactly, displayed on-line, and used in machine processes such as c
computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a f
in pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font at a
systems capable of producing a high degree of recognition accuracy for most fonts are now com
support for a variety of digital image file format inputs.[2] Some systems are capable of rep
put that closely approximates the original page including images, columns, and other

Figure 3.1: Scanning in OCR

A fixed threshold is used where gray levels below this threshold are black and levels above are white. For high contrast document with uniform background a pre-chosen fixed threshold can be sufficient. However, documents encountered in practice have rather large range. In these cases more sophisticated methods for thresholding are required to obtain good results. The best thresholding methods vary threshold adapting to local properties of document such as contrast and brightness. However, such methods usually depend on multilevel scanning of document which requires more memory and computational capacity.

3.2 Pre-Prcoessing

The third OCR component is pre-processing. The raw data depending on the data acquisition type is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. The image resulting from scanning process may contain certain amount of noise. Depending on the scanner resolution and the inherent thresholding, the characters may be smeared or broken. Some of these defects which may cause poor recognition rates and are eliminated through pre-processor by smoothing digitized characters. Smoothing implies both filling and thinning. Filling eliminates small breaks, gaps and holes in digitized characters while thinning reduces width of line. The most common technique for smoothing moves a window across binary image of character and applies certain rules to the contents of window. Pre-processing also includes normalization alongwith smoothing. The normalization is applied to obtain characters of uniform size, slant and rotation. The correct rotation is found through its angle. For rotated pages and lines of text, variants of Hough transform are commonly used for detecting skew.

Some will argue that image pre-processing is not a good idea, since it distorts or changes the true nature of the raw data. However, intelligent use of image pre-processing can provide benefits and solve problems that ultimately lead to better feature detection. We applied common methods for image enhancements and corrections that will affect feature analysis downstream in the vision pipeline in both favorable and unfavorable ways, depending on how the methods are employed.

Pre-processing stage consists of compression and binarization steps.

1. Compression
2. Binarization

3.2.1 Compression

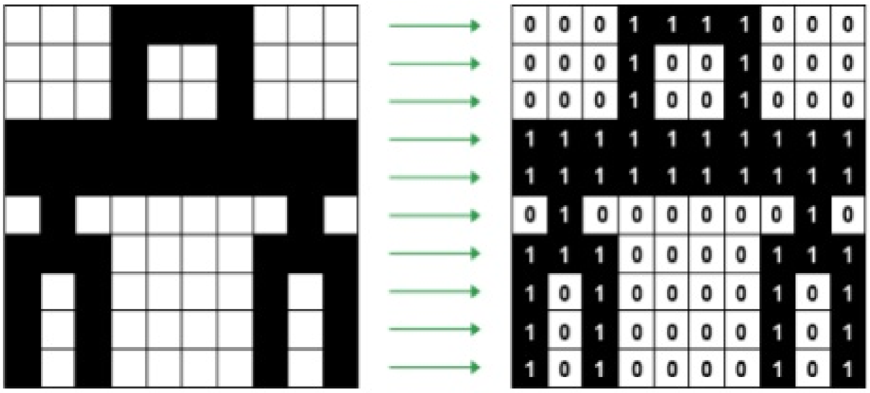
Scanners, digital cameras and other image capture devices usually produce high resolution images of very high file sizes, and these need to be compressed into smaller sizes for efficent storage and retrieval. These devices create the image files as tiff or raw format containing a lot of redundant data, as well as repeated headers. Compressing scanned images can remove this redundant data.

Several techniques exist for compressing scanned images. Tiff images can easily be converted into jpg, svg, gif, bmp or other such file formats that allow for up to 60 percent compression, depending on the resolution of the original scanned image.

Scanned image usually contains a lot of data that is repeated; for example, an image showing a piece of the sky will contain a large area that is uniformly blue. A compression algorithm, such as the one used in a jog file, identifies this repetition and sets up a table containing one example from each repeated data so that it can reconstruct the image from these examples. In addition, it also scans for patterns and extraneous data, so that it can discard any extra data and still contain all the essential information. In this way, it can compress our scanned images with negligible loss of quality. This allows the user to compress scanned images with minimal loss of quality and important data.

3.2.2 Binarization

The presence of background images or textures is not the only factor that can impair recognition quality. Low recognition quality brings also the low contrast of the original document and the changing brightness of the background. For such documents the Adaptive Binarization procedure is used. It measures the brightness of the background and the saturation of the black areas along the line in order to find optimal binarization parameters for each separate line's fragment. As a result, the lines and words will be correctly detected and higher recognition accuracy will be reached.



0	0	0	1	1	1	1	0	0	0
0	0	0	1	0	0	1	0	0	0
0	0	0	1	0	0	1	0	0	0
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
0	1	0	0	0	0	0	0	1	0
1	1	1	0	0	0	0	1	1	1
1	0	1	0	0	0	0	1	0	1
1	0	1	0	0	0	0	1	0	1
1	0	1	0	0	0	0	1	0	1

Figure 3.2: Binarization

With the help of bilateral filter applied on complex images multiple times, the input image is then converted to grey scale. A copy of the resultant grey scale image is inverted to handle various challenges with respect to font and background color. We will apply thresholding techniques like median blur followed by Gaussian blur on both the images and then we will add a black color border of one pixel width to both images and use flood fill technique to flood the pixels with white pixels. Now, will calculate black pixels in both the images and discard the one with less black pixels. The image is thus converted to a pure binary image that can be accessed as grey scale image with only two values 0 for black and 255 for white.

3.3 Image Segmentation

First, the captured document images are pre-processed for the perspective correction and noise removal.Then, the final image is converted to grayscale and binarized using Otsu segmentation method for further processing. Furthermore, looking at the mean horizontal run length of both black and white pixels the proper segmentation of foreground objects is checked. For example, for the document images having dark background and light foreground, the output of the binarization is reversed i.e. black background (represented as 0's) and white foreground (represented as 1's).

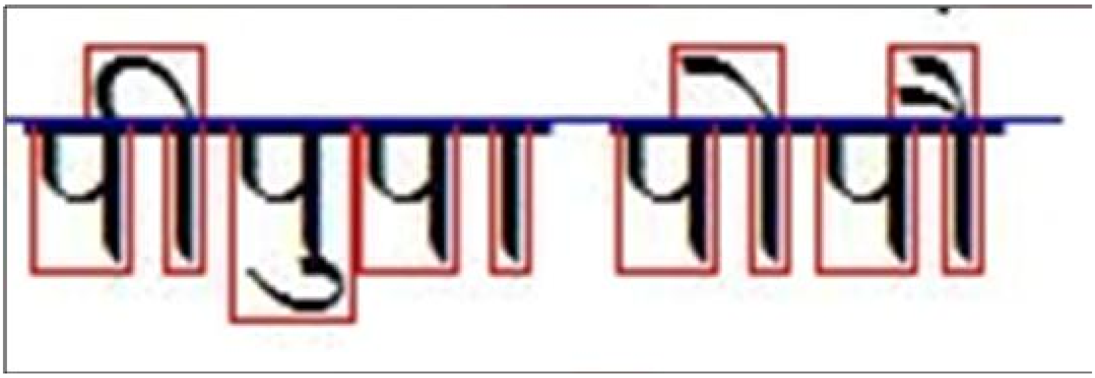


Figure 3.3: Image Segmentation

Segmentation of lines and words: The preliminary segmentation consists of the following steps:

1. We compute the horizontal projection of the document image box. Create one vector in which all the columns in row are white pixels. And from that number of rows line are separated from text.
2. We compute the horizontal projection of the document image box. The row containing maximum number of black pixels is considered to be the header line and remove it.
3. Separate character/symbol boxes of the image below the header line: To do this, we make vertical projection of the image starting from header line position to the bottom row of the word image box. The columns that have no black pixels are treated as boundaries for extracting image boxes corresponding to characters.
4. We compute the vertical projection of the image, starting from the top row of the image to the header Line Position. The columns that have no black pixels are used as delimiters for extracting top modifier symbol boxes.

3.4 Feature Extraction

The next OCR component is feature extraction. The objective of feature extraction is to capture essential characteristics of symbols. Feature extraction is accepted as one of the most difficult problems of pattern recognition. The most straight forward way of describing character is by actual raster image. Another approach is to extract certain features that characterize symbols but leaves the unimportant attributes. The techniques for extraction of such features are divided into three groups' viz. distribution of points transformations and series expansions and structural analysis.

The different groups of features are evaluated according to their noise sensitivity, deformation, ease of implementation and use. The criteria used in this evaluation are: robustness in terms of noise, distortions, style variation, translation and rotation and practical usage in terms of recognition speed, implementation complexity and independence. Some of the commonly used feature extraction techniques are template matching and correlation, transformations, distribution of points and structural analysis.

There are many features are extracted for the recognition of Marathi characters. For that consider features as follows-

1. Histogram of individual characters.
2. GLCM (Gray level co-occurrence matrix).

3.4.1 Histogram of individual characters

The histogram-based features used in this work are first order statistics that include mean and variance. Let z be a random variable denoting image gray levels and p(zi), = 0,1,2,3,.......L-1, be the corresponding histogram, where L is the number of distinct gray levels. The features are calculated using the above-mentioned histogram.

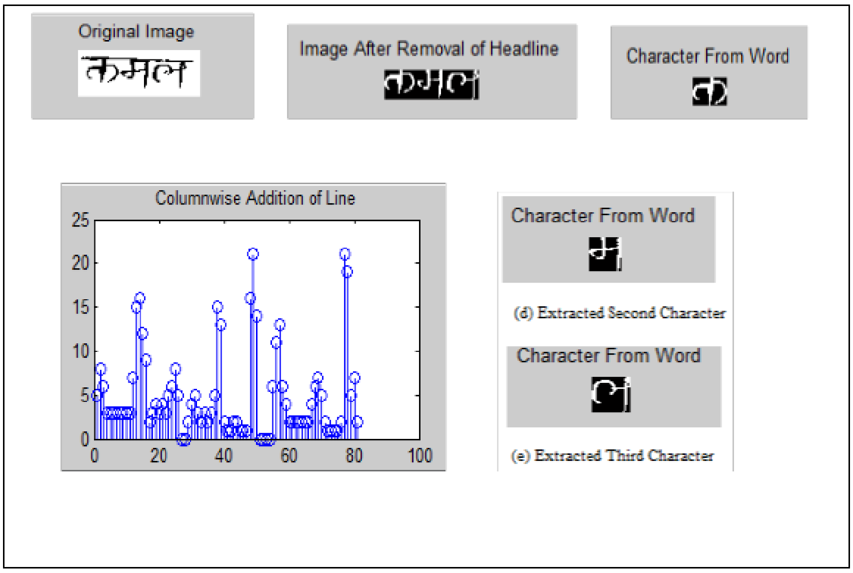

Original Image
कमल
Image After Removal of Headline
Character From Word
Columnwise Addition of Line
25
20
15
10
5
0
0
20
40
60
80
100
Character From Word
(d) Extracted Second Character
Character From Word
(e) Extracted Third Character

Figure 3.4: Histogram of individual characters

1. Mean

   The mean gives the average gray level of each region and it is useful only as a rough idea of intensity not really texture.

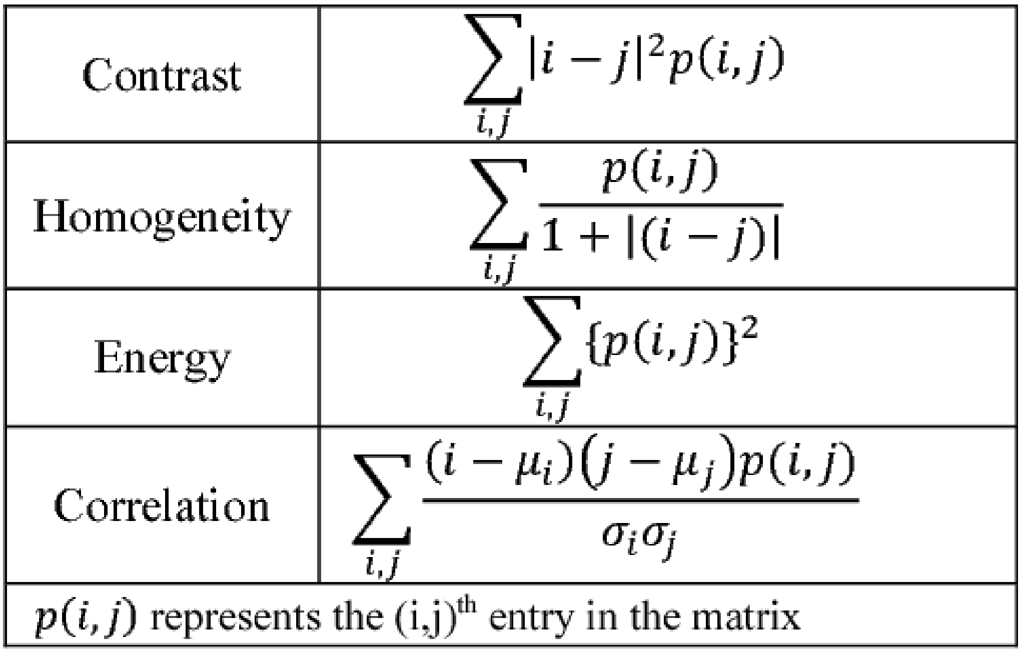
2. Variance

The variance gives the amount of gray level fluctuations from the mean gray level value.

3.4.2 GLCM(Gray Level Co-occurence Matrix)

Measures of texture computed using histograms suffer from the limitation that they carry no information regarding the relative position of the pixels with respect to each other. One way to bring this type of information into the texture analysis process is to consider not only the distribution of the intensities but also the positions of pixels with equal or nearly equal intensity values. One such type of feature extraction is from gray level co-occurrence matrices.

The second-order gray level probability distribution of a texture image can be calculated by considering the gray levels of pixels in pairs at a time. A second-order probability is often called a GLC probability. For a given displacement vector D5 at Dx Dy, the joint probability of a pixel at location (x, y) having a gray level i, and the pixel at location (x1Dx, y1Dy) having a gray level j.In other words it is a second- order joint probability P (i, j) of the intensity values of two pixels (i and j), a distance d apart along a given direction, which is the probability that j and i have the same intensity. This joint probability takes the form of a square array Pd with row and column dimensions equal to the number of discrete gray levels (intensities) in the image being examined. If an intensity image were entirely flat (i.e. cont ained no texture), the resulting GLCM would be completely diagonal. As the image texture increases, the off-diagonal values in GLCM become larger. The various features that can be calculated from the co -occurrence matrices (C) are inertia (contrast), absolute value, inverse difference, energy, and entropy.



Contrast	$\sum_{i,j} \lvert i-j \rvert^2 p(i,j)$
Homogeneity	$\sum_{i,j} \frac{p(i,j)}{1+\lvert(i-j)\rvert}$
Energy	$\sum_{i,j} \{p(i,j)\}^2$
Correlation	$\sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j}$
$p(i,j)$ represents the $(i,j)^{th}$ entry in the matrix	

Figure 3.5: Features calculated from GLCM

1. Contrast
   Contrast is the element difference moment of order 2, which has a relatively low value when the high values of C are near the main diagonal.

2. Correlation
   In correlation calculation we measures the joint probability occurrence of the specified pixel pairs.

3. Energy
   It Provides the sum of squared elements in the GLCM. Also known as uniformity or the angular second moment.

4. Homogeniety
   It Measures the closeness of the distribution of elements in GLCM.

3.5 Training and Recognition

The seventh OCR component is training and recognition. OCR systems extensively use the methodologies of pattern recognition which assigns an unknown sample into a predefined class. The OCR are investigated in four general approaches of pattern recognition as suggested in (a) template matching (b) statistical techniques (c) structural techniques and (d) ANNs. These approaches are neither necessarily independent nor disjointed from each other. Occasionally, an OCR technique in one approach can also be considered to be a member of other approaches. In all of the above approaches, OCR techniques use either holistic or analytic strategies for the training and recognition stages. The holistic strategy employs top down approaches for recognizing the full character eliminating the segmentation problem. The price for this computational saving is to constrain the problem of OCR to limited vocabulary. Also, due to the complexity introduced by the representation of a single character or stroke the recognition accuracy is decreased. On the other hand, the analytic strategies employ bottom up approach starting from stroke or character level and going toward producing a meaningful text. The explicit or implicit segmentation algorithms are required for this strategy, not only adding extra complexity to the problem but also introducing segmentation error to the system. However, with the cooperation of segmentation stage, the problem is reduced to the recognition of simple isolated characters or strokes, which can be handled for unlimited vocabulary with high recognition rates.

3.6 Post-processing

he eighth OCR component is post-processing. Some of the commonly used post-processing activities include grouping and error detection and correction. In rouping symbols in text are associated with strings. The result of plain symbol recognition in text is a set of individual symbols. However, these symbols do not usually contain enough information. These individual symbols are associated with each other making up words and numbers. The grouping of symbols into strings is based on symbols' location in document. The symbols which are sufficiently close are grouped together. For fonts with fixed pitch grouping process is easy as position of each character is known. For typeset characters distance between characters are variable. The distance between words are significantly large than distance between characters and grouping is therefore possible. The problems occur for handwritten characters when text is skewed. Until grouping each character is treated separately, the context in which each character appears has not been exploited. However, in advanced optical text recognition problems, system consisting only of single character recognition is not sufficient.

Chapter 4

Result

Chapter 5

Future Scope

All through the years, the methods of OCR systems have improved from primitive schemes suitable only for reading stylized printed numerals to more complex and sophisticated techniques for the recognition of a great variety of typeset fonts and also hand printed characters. The new methods for character recognition continue appear with development of computer technology and decrease in computational restrictions.However, the greatest potential lies in exploiting existing methods by hybridizing technologies and making more use of context. The integration of segmentation and contextual analysis improves recognition of joined and split characters.Also higher level contextual analysis which looks at semantics of entire sentences are useful.Generally there is a potential in using context to larger extent than what is done today.

In addition, a combination of multiple independent feature sets and classifiers where weakness of one method is compensated by the strength of another improves recognition of individual characters.The research frontiers within character recognition continue to move towards recognition of sophisticated cursive script that is handwritten connected or calligraphic characters.Some promising techniques within this area deal with recognition of entire words instead of individual characters.

Chapter 6

Conclusion

Character recognition techniques associate a symbolic identity with the image of character.In this project an overview of various techniques of OCR has been presented.An OCR comprises various phases such as acquisition, pre- processing, segmentation, feature extraction, classification and post-processing.Each of the steps is discussed in detail in this project.Using a combination of these techniques, an efficient OCR system can be developed as a future work.

The OCR system can also be used in different practical applications such as numberplate recognition, smart libraries and various other real-time applications.The implementation of Optical character recognition technology can be efficiently used to speed up translation of image based documents that are currently easy to discover, search and process.

Bibliography

1. https://ieeexplore.ieee.org/document/953898
2. https://ieeexplore.ieee.org/document/6993174
3. A. Choudhary, R. Rishi and S. Ahlawat, “Offline handwritten character recognition using features extracted from binarization technique”, Elsevier AASRI Conference on Intelligent Systems and Control Procedia, vol. 4,pp. 306-312, 2013.
4. M. Gupta, N. Jacobson and E. Garcia, “OCR binarization and image preprocessing for searching historical documents”, Elsevier Pattern Recognition, vol. 17,no. 3,pp. 1-6,2015.