Marksheet Recognition Methodology-Correlation

Rushali Gurav, Trupti Joshi, V. B. Musande

Jawaharlal Nehru Engineering College,

Aurangabad, Maharashtra, India

E-mail: rushaligurav5@gmail.com, trupjosh@gmail.com, vijayamusande@gmail.com

Abstract

Now-a-days OCR becoming very popular research in image processing. Character recognition techniques associate a symbolic identity with the image of character. We are developing a system for retrieving information from scanned copy of marksheet which is later stored on a database. In a typical OCR systems input characters are digitized by an optical scanner. Each character is then situated and metameric and also the ensuing character image is fed into a pre-processor for noise reduction. Options square measure the extracted from the character for classification. The feature extraction is tough and plenty of totally different techniques exist. When classification the extracted characters square measure sorted to reconstruct the first image. Here, we have a tendency to square measure mistreatment OCR's correlation rule.

Keywords: *Feature extraction, segmentation, template matching and correlation*

INTRODUCTION

Optical character recognition (optical character reader) (OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text. It is widely used as a form of data entry from printed paper data records, whether, passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation [1, 2]. It is a common method of digitizing printed texts so that it can be electronically edited, searched, stored more compactly, displayed on-line and used in machine processes such as machine translation, text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. In Traditional systems, documents storage was done by paper work and traditional file systems, but, this type of storage is tedious due to time. Another way is to digitalize all the information manually. This is not only tedious but is causes to human errors. This is the reason why Digital Document Image Processing is emerging as the important function for any industry. Increasing demand for digitalization of data needs an automated tool for converting hard coded data into digital format. Data cannot be retrieved directly from images, it has to be done manually. An OCR can only detect text, however, preprocessing of the images is very difficult before using an OCR because an image in raw form cannot be processed by OCR. Also, after OCR detects the text from images, the data obtained should be stored in a database where it can be handled and processed easily. Currently, marksheet details are manually entered into the database. This requires lots of human efforts and is time as well. In addition to this there is a risk of human errors because it is a tedious job. Hence, using methods of image processing we try to automate the complete process of creating a student database from marksheet [3, 4].

WHAT IS OCR?

The aim of Optical Character Recognition (OCR) is to classify optical patterns to alphameric or alternative characters. Th e method of OCR involves few steps as well as segmentation, feature extraction, and classification. Every of those steps could be a field unto itself and is represented concisely here within the co ntext of a Matlab implementation of OCR. One example of OCR is shown below. A portion of a scanned image of text, taken from the source, is shown along with the matching characters from that text. The examples of OCR applications are listed given below. The most common for use OCR is that people often wish to convert text documents to digital data.

1. 1. People wish to scan in a document and have the text of that document available in a word processor.
2. Recognizing license plate numbers
3. Post Office needs to recognize zip-codes.

STEPWISE FLOW OF SYSTEM

1. The input provided is scanned marksheet image.
2. Correlation algorithm is applied onto this image.
3. The binarized image results are given as an input to OCR.
4. OCR detects text out of this binarized image.
5. This text is fetched from OCR.

6. This excel file is saved as database.

To recognize character first off, the input pictures area

unit noninheritable containing English text as AN input image. Pictures area unit then keep in

some image file like BMP, JPG etc. This image afterward passes through preprocessing, segmentation, feature extraction and classification steps.

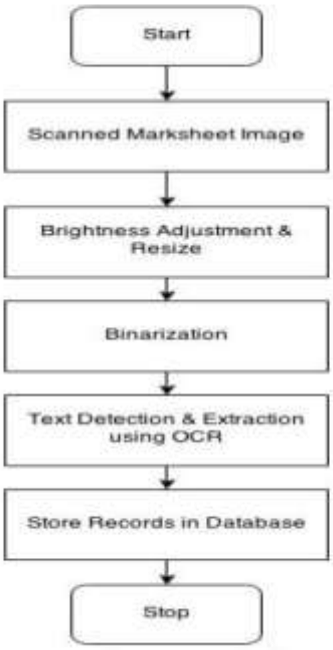

Start
Scanned Marksheet Image
Brightness Adjustment & Resize
Binarization
Text Detection & Extraction using OCR
Store Records in Database
Stop

Fig. 1: *System Flow.*

CORRELATION

This section will explain the proposed algorithm, i.e., what are different steps involve for achieving the OCR recognition. Figure 1 is showing the different steps required to do correlation based OCR recognition.

Step 1: The image is cropped to fit the text.

Step 2: Once trimmed the image, next step is to separate each line.

Step 3: Once attained unconnectedly each line of the image, continue to extract one letter of the image matrix and continue until all word get separated.

Step 4: Classification-The main process is used for the classification was the two-dimensional correlation. This process gives a value of the correspondence between two matrices (images).

Step 5: Output file.

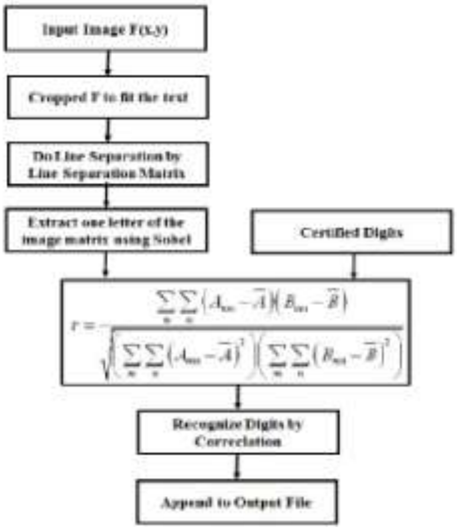

Input Image F(x,y)
Cropped F to fit the text
Do Line Separation by Line Separation Matrix
Extract one letter of the image matrix using Sobel
Certified Digits
$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}}$
Recognize Digits by Correlation
Append to Output File

Fig. 2: *Flow Chart.*

DESIGN AND IMPLEMENTATION STEPS

Database Creation

Initially, we have created information of all character pictures having upper-case letters, lower-case letters and numeral digits of English scripts from A-Z, a-z and

0-9 of pixels 42×26.

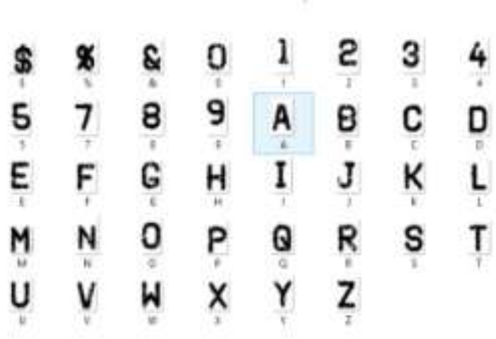


Fig. 3: *Dataset.*

Data Acquisition

Through the scanning process a digital image of the original document is captured. Scanned images are then stored in some picture file such as BMP, JPG etc.

RGB to Gray Conversion

In the pre-processing 1st stage is to convert the input RGB image into gray scale image.

Fig. 4: *RGB Converted into Binary Images.*

Feature Extraction

After character segmentation, features from each segmented character are extracted which is in the form of Matrix as shown in Figure.

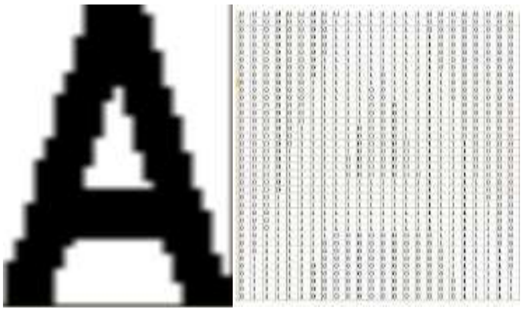


Fig. 5: *Character Extraction in Form of Matrix.*

Binarization

Binarization is process of converting gray scale image into binary image.

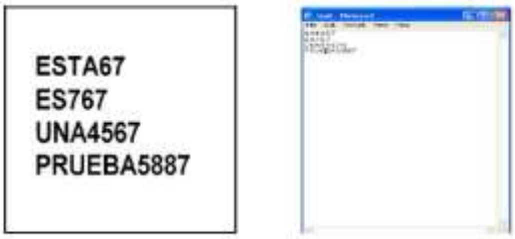
Segmentation

It is an operation that seeks to decompose a picture of sequence of characters into sub pictures of individual symbols. Character segmentation may be a key demand that determines the utility of typical Character Recognition systems. It includes line, word and character segmentation. Completely different strategies used may be classified supported the sort of text and strategy being followed like recognition-based segmentation.

Template Matching and Correlation

These techniques are different from the others in that no features are actually extracted. Instead the matrix containing the image of the input character is directly matched with a set of prototype characters representing each possible class. The distance between the pattern and each prototype is computed, and the class of the prototype giving the best match is assigned to the pattern. The technique is simple and easy to implement in hardware and has been used in many commercial OCR machines.

Following is sample input and output in text file is shown.



ESTA67
ES767
UNA4567
PRUEBA5887

Fig. ***6:*** *Text Image Sample and its Output.*

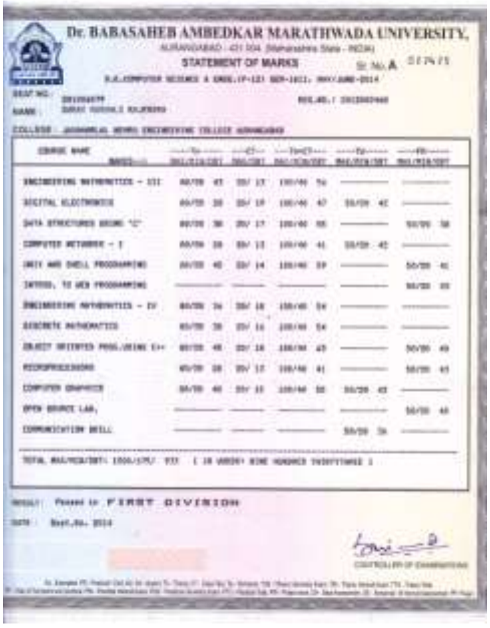
ACCURACY OF OCR

Accuracy of an OCR system depends on the quality of input document. Sometimes the output from OCR systems is often quite "noisy". Post processing is done on the text to correct the noise. The average time taken to recognize 20 words is 350 ms and that of 100 words is 500 ms. The accuracy of the OCR system also depends on the camera used to capture the raw image of the document. Various factors affecting the quality are: Focus of the camera, resolution of the picture, amount of noise present etc. Correlation algorithm achieved an average accuracy of 93%. To estimate the OCR accuracy, the OCR output can then be compared to the text of the original document, called the ground truth.

CONCLUSION

Document image processing is burgeoning as a critical domain in computer engineering. These system revolutionaries the conventional approach by automating document image processing. It minimizes the need to do any manual work. This system can be further extended for various other documents having fixed formats and bold faced text

EXPERIMENTAL RESULTS



Dr. BABASAHEB AMBEDKAR MARATHWADA UNIVERSITY,
AURANGABAD - 431 004 (Maharashtra State - INDIA)
STATEMENT OF MARKS
Sr. No. A [illegible]
S.E.COMPUTER SCIENCE & ENGG.(P-12) [illegible] MAY/JUNE-2014

SEAT NO.: [illegible]
NAME : [illegible]
COLLEGE : JAWAHARLAL NEHRU ENGINEERING COLLEGE AURANGABAD

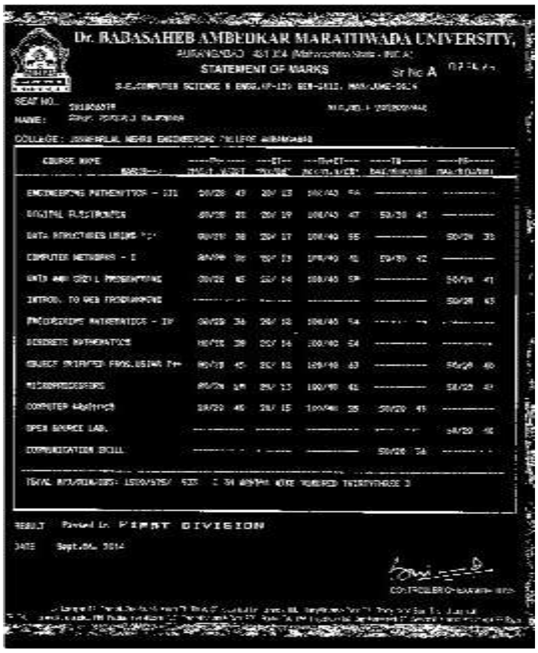
COURSE NAME	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
ENGINEERING MATHEMATICS - III	[illegible]	[illegible]	[illegible]	—	—
DIGITAL ELECTRONICS	[illegible]	[illegible]	[illegible]	[illegible]	—
DATA STRUCTURES USING "C"	[illegible]	[illegible]	[illegible]	—	[illegible]
COMPUTER NETWORK - I	[illegible]	[illegible]	[illegible]	[illegible]	—
UNIX AND SHELL PROGRAMMING	[illegible]	[illegible]	[illegible]	—	[illegible]
INTROD. TO WEB PROGRAMMING	—	—	—	—	[illegible]
ENGINEERING MATHEMATICS - IV	[illegible]	[illegible]	[illegible]	—	—
DISCRETE MATHEMATICS	[illegible]	[illegible]	[illegible]	—	—
OBJECT ORIENTED PROG.USING C++	[illegible]	[illegible]	[illegible]	—	[illegible]
MICROPROCESSORS	[illegible]	[illegible]	[illegible]	—	[illegible]
COMPUTER GRAPHICS	[illegible]	[illegible]	[illegible]	[illegible]	—
OPEN SOURCE LAB.	—	—	—	—	[illegible]
COMMUNICATION SKILL	—	—	—	[illegible]	—

TOTAL [illegible] ([illegible])

RESULT: Passed in FIRST DIVISION
DATE : Sept.[illegible] 2014

CONTROLLER OF EXAMINATIONS

Fig. ***7:*** *Original Image.*



Dr. BABASAHEB AMBEDKAR MARATHWADA UNIVERSITY,
AURANGABAD 431 004 (Maharashtra State - INDIA)
STATEMENT OF MARKS
Sr. No. A [illegible]
S.E.COMPUTER SCIENCE & ENGG.(F-12) SEM-[illegible], MAY/JUNE-2014

SEAT NO. [illegible]
NAME: [illegible]
COLLEGE: JAWAHARLAL NEHRU ENGINEERING COLLEGE AURANGABAD

COURSE NAME	TH	CT	TH+CT	TW	PR
ENGINEERING MATHEMATICS - III	80/28 43	20/ 13	100/40 56	----	----
DIGITAL ELECTRONICS	80/28 31	20/ 19	100/40 47	50/20 42	----
DATA STRUCTURES USING "C"	80/28 38	20/ 17	100/40 55	----	50/20 33
COMPUTER NETWORKS - I	80/28 [illegible]	20/ 13	100/40 41	50/20 42	----
[illegible]	80/28 45	20/ 14	100/40 59	----	50/20 41
INTROD. TO WEB PROGRAMMING	----	----	----	----	50/20 43
ENGINEERING MATHEMATICS - IV	80/28 36	20/ 18	100/40 54	----	----
DISCRETE MATHEMATICS	80/28 38	20/ 16	100/40 54	----	----
OBJECT ORIENTED PROG.USING C++	80/28 45	20/ 18	100/40 63	----	50/20 40
MICROPROCESSORS	80/28 28	20/ 13	100/40 41	----	50/20 42
COMPUTER GRAPHICS	80/28 40	20/ 15	100/40 55	50/20 41	----
OPEN SOURCE LAB.	----	----	----	----	50/20 40
COMMUNICATION SKILL	----	----	----	50/20 34	----

TOTAL [illegible]: 1500/[illegible]/ 533 [illegible]

RESULT Passed in FIRST DIVISION
DATE Sept.06, 2014

CONTROLLER OF EXAMINATIONS

Fig. 8: *Gray Image.*

REFERENCES

1. Sukhpreet Singh, Talwandi Sabo. Improve optical character recognition using templates and correlation. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*. 2014; 3(9).
2. Richa Muke, Sharvari Patil, Janhavi Acharya, Snkirti Shiravale. Marksheet image processing. 2014; 3(3).
3. Sravan Ch, Shivanku Mahna, Nirbhay Kashap. Optical character recognition on handheld devies. 2015; 115(22).
4. G.Vamvakas, B.Gatos, N. Stamatopoulos, S.J.Perantonis. A complete optical character recognition methodology for historical documents.