Optical Character Recognition

A Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of
Bachelor of Technology
in
Computer Science & Engineering

by

Pratik Parwal
Nikhil Agarwal
Richa Pandey
Saurav Jha
Pramee Chowdhury

to the
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
MOTILAL NEHRU NATIONAL INSTITUTE OF TECHNOLOGY
ALLAHABAD
April, 2017

UNDERTAKING

I declare that the work presented in this report titled "*Optical Character Recognition*", submitted to the Computer Science and Engineering Department, Motilal Nehru National Institute of Technology, Allahabad, for the award of the ***Bachelor of Technology*** degree in ***Computer Science & Engineering***, is my original work. I have not plagiarized or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, I accept that my degree may be unconditionally withdrawn.

April, 2017
Allahabad

Pratik Parwal
Nikhil Agarwal
Richa Pandey
Saurav Jha
Pramee Chowdhury

CERTIFICATE

Certified that the work contained in the report titled "*Optical Character Recognition*", by Pratik Parwal, Nikhil Agarwal, Richa Pandey, Saurav Jha, Pramee Chowdhury has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

(Prof. Suneeta Agarwal)
Computer Science and Engineering Dept.
M.N.N.I.T, Allahabad

April, 2017

Preface

Document Image processing and Optical Character Recognition (OCR) have been frontline research area in the field of human-machine interface for the last few decades. Recognition of machine printed or hand printed document is an essential part in applications like intelligent scanning machines, text to speech converters and automatic language to language translators. The objective of document image analysis is to recognise the text and graphics components in the paper document and to extract the intended information, as human beings do. Two components of document image analysis are Textual processing and Graphical processing. Textual processing deals with the text component of the document image. The graphical processing deals with non-textual line and symbol components that make up line diagrams, delimiting straight lines between text sections and company logos etc. In the current context, we limit ourselves to the textual processing part that comprises of printed capital and small letters of English alphabet.

Acknowledgements

We are highly grateful to our project mentor Prof. Suneeta Agarwal, Department of Computer Science and Engineering, Motilal Nehru National Institute Of Technology Allahabad for her indispensable guidance, suggestions and feedback that she gave us. She has always been there to solve all our doubts. She also encouraged us in developing new ideas that made us think better to solve our problems in the most efficient way. She also provided a new vision for the practical application and utility of the project we are working. Our thanks to all other faculty members of Department of Computer Science and Engineering, Motilal Nehru National Institute Of Technology Allahabad for helping us in every possible manner.

Contents

List of Figures

Chapter 1

Introduction

Optical Character Recognition (OCR) is the process of extracting text from an image. The main purpose of an OCR is to make editable documents from existing paper documents or image files. Significant number of algorithms is required to develop an OCR and basically it works in two phases such as character and word detection. In case of a more sophisticated approach, an OCR also works on sentence detection to preserve a document's structure.

1.1 Motivation

There are numerous applications where we want to select some text from an image. Here OCR comes to our rescue . The most common for use OCR is that people often wish to convert text documents to some sort of digital representation. The applications of OCR are as follows :

1. People wish to scan in a document and have the text of that document available in a word processor.
2. Post Office needs to recognize zip-codes.
3. Defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR.
4. Assistive technology for blind and visually impaired users.

Chapter 2

Related Works

Optical character recognition for English has become one of the most successful application in pattern recognition and artificial intelligence. [Line, 1993; Pal Chaudhuri, 2004] clearly states the proposed works in this field by dividing the commercial OCR systems into following four generations depending on their versatility, robustness and efficiency:

1. **First Generation OCR Systems**: Character recognition originated as early as 1870 when Carey invented the retina scanner, which is an image transmission system using photocells. It is used as an aid to the visually handicapped by the Russian scientist Tyurin in 1900. However, the first generation machines appeared in the beginning of the 1960s with the development of the digital computers.The first generation machines are characterized by the constrained letter shapes which the OCRs can read. These symbols were specially designed for machine reading, and they did not even look natural. The first commercialized OCR of this generation was IBM 1418, which was designed to read a special IBM font, 407.

2. **Second Generation OCR Systems**: Next generation machines were able to recognize regular machine-printed and hand- printed characters. The character set was limited to numerals and a few letters and symbols. Such machines appeared in the middle of 1960s to early 1970s. The first automatic letter-sorting machine for postal code numbers from Toshiba was developed during

this period. The methods were based on the structural analysis approach. Significant efforts for standardization were also made in this period. An American standard OCR character set: OCR-A font (Figure 1) was defined, which was designed to facilitate optical recognition, although still readable to humans. A European font OCR-B (Figure 2) was also designed.

Figure 1: OCR-A font

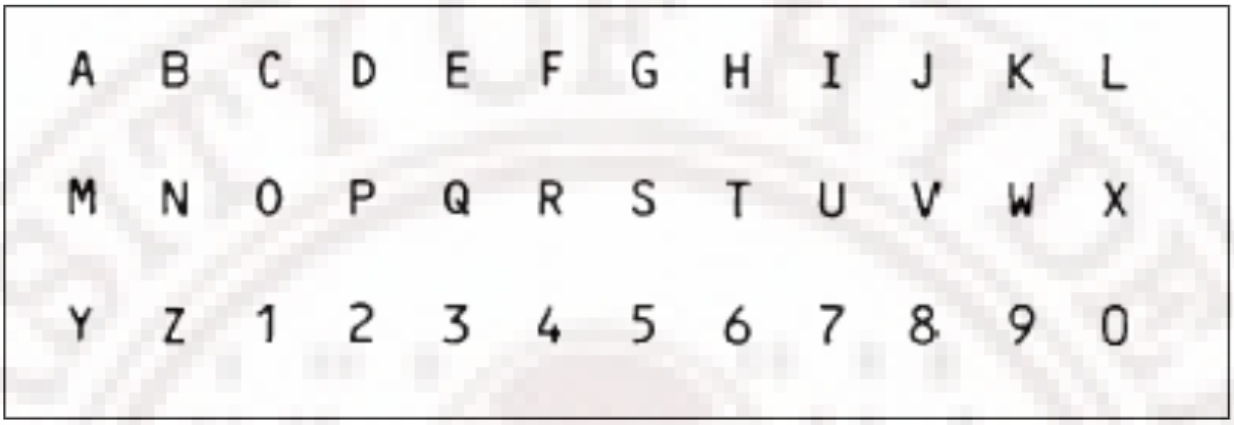


Figure 2: OCR-B font

3. **Third Generation OCR Systems**:For the third generation of OCR systems, the challenges were documents of poor quality and large printed and handwritten character sets. Low cost and high performance were also important objectives. Commercial OCR systems with such capabilities appeared during the decade 1975 to 1985.

4. **OCRs Today (Fourth generation OCR systems)**: The fourth generation can be characterized by the OCR of complex documents intermixing with text, graphics, tables and mathematical symbols, unconstrained hand- written characters, color documents, low-quality noisy documents, etc. Among the commercial products, postal address readers, and reading aids for the blind are available in the market.

   Nowadays, there is much motivation to provide computerized document analysis systems. OCR contributes to this progress by providing techniques to convert large volumes of data automatically. A large number of papers and patents advertise recognition rates as high as 99.99%; this gives the impression that automation problems seem to have been solved.

   Various methods have been proposed to increase the accuracy of optical character recognizers. In fact, at various research laboratories, the challenge is to develop robust methods that remove as much as possible the typographical and noise restrictions while maintaining rates similar to those provided by limited-font commercial machines [Belaid,1997]. Although OCR is widely used presently, its accuracy today is still far from that of a seven-year old child, let alone a moderately skilled typist [Nagy, Nartker Rice, 2000].

   Thus, current active research areas in OCR include handwriting recognition, and also the printed typewritten version of non-Roman scripts (especially those with a very large number of characters).

Chapter 3

Proposed Work

We aim at developing a comprehensive system for recognition of unconstrained printed English small and capital alphabets in an manner outlining all the issues involved.

The following are the proposals of our work :

1. To select pre processing algorithms that retains the shape of the character.
2. To work at feature extraction level and classification level and select combination that gives best results.
3. To analyze the performance using various Decision Tree classifiers and their combination for the robust recognition of the printed characters.
4. To design hybrid systems that combine several classification and/or recognition techniques along with the one we have used.
5. To analyze the efficiency of these classifiers and find the optimum one.

Chapter 4

Experimental Setup and Results Analysis

4.1 Block Diagram of an OCR system

As shown in the block diagram in figure 1, an OCR system has two phases, namely the training phase and the recognition phase. Both the phases have some common steps that have been divided by a horizontal line.

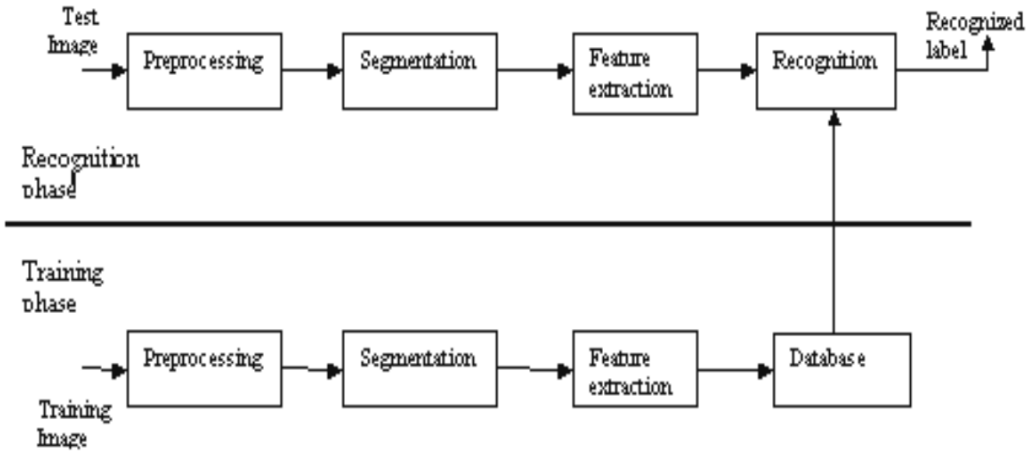
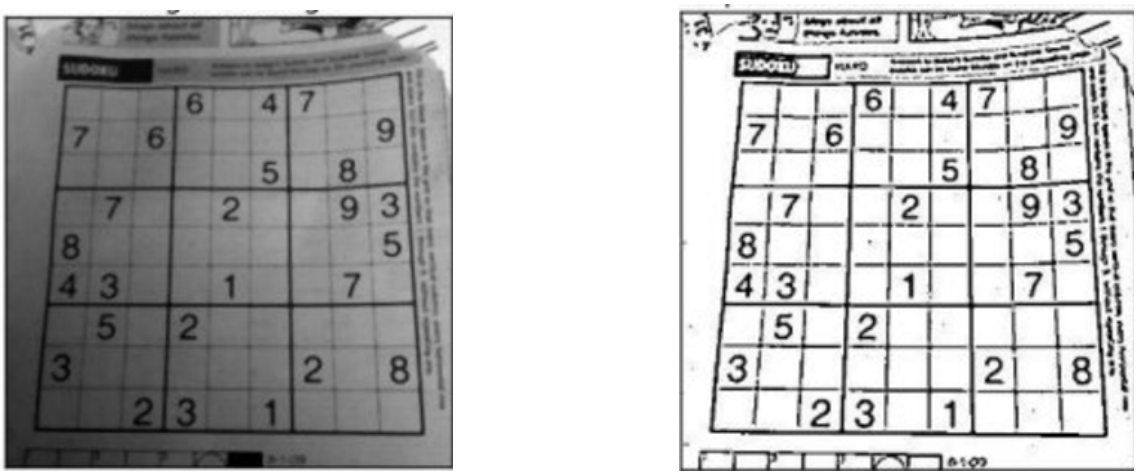


Figure 3: Block diagram of an OCR system

4.2 Image Preprocessing

4.2.1 Binarization

The complete experimental setup is based on binarized images. Different state-of-the-art binarization methods can be classified into two groups: (i) global binarization (like Otsu) and (ii) local binarization (like Sauvola). Global binarization estimates a single threshold for the complete image, whereas local binarization calculates a threshold for each individual pixel based on the neighborhood information. In general, local binarization works better than global binarization under different types of document image degradations like non-uniform shading or blurring etc. However, local binarization methods are slower than global binarization methods. In our project , we have used local binarization technique of Adaptive Gaussian thresholding for better results . The input is taken as a grayscale image. The corresponding output binarization pixels are denoted $b_i \in \{0, 1\}$, where 0 refers to black and 1 refers to white.



(a) Original Image (b) Image after Binarization

Figure 4: Binarization

4.2.2 Segmentation

Segmentation is the process of extracting objects of interest from an image. In document analysis, the first task towards text recognition is to segment the textual regions from graphics, maps and other figures. For detecting individual character in the text we follow following segmentation techniques :

1) *Line Segmentation:* To separate text lines, the horizontal projection profile of the text document image is calculated. The horizontal projection profile (HPP) is a histogram of a number of ON pixels along every row of the image. When the projection profiles are plotted, we can see peaks and valleys in the plot. White space between the text lines is used to segment the text lines.

2) *Character Segmentation:* The spacing between the characters is used for character segmentation. Generally in printed English scripts, spacing between the characters of a word remains uniform throughout the text. The spacing between the characters is found by taking the Vertical Projection Profile (VPP) of an input text line. Vertical Projection profile is the sum of ON pixels along every column of the image within the line.

4.2.3 Normalization

The segmented characters are normalized to a predefined size before proceeding to the next phase . This allows all font sizes to be converted to a fixed size (24 x 24) to improve the generality of the ocr in recognising different fonts sizes .

4.2.4 Thinning

Thinning is an image preprocessing operation performed to make the image crisper by reducing the binary-valued image regions to lines that approximate the skeletons of the region. Thinning cleans the image so that only reduced amount of data needs to be processed in the next image processing stage. Shape analysis could be done easily.

Our method of extracting the skeleton of a picture consists of removing all contour points of the picture except those points that belong to the skeleton . In order to preserve the connectivity of the skeleton we divide each iteration into two subiterations .

1. In the first iteration , the south east boundary points and north west corner points are removed . In this iteration , contour point P_1 is deleted from the digital pattern if it satisfies the following condition :

   (a) $2 \leq B(P_1) \leq 6$

   (b) $A(P) \leq= 1$

   (c) $P_2 * P_4 * P_6 = 0$

   (d) $P_4 * P_6 * P_8 = 0$

   where A(P_1) is the number of 01 patterns in the ordered set $P_2, P_3 ... P_8, P_9$ that are the eight neighbours of P_1.
   B(P_1) is the number of non zero neighbours of P_1.

2. In the second iteration , the north west boundary points and south east corner points are removed.
   Ony conditions (c) and (d) of first iteration are changed as follows :

   (a) $P_2 * P_4 * P_8 = 0$

   (b) $P_2 * P_6 * P_8 = 0$

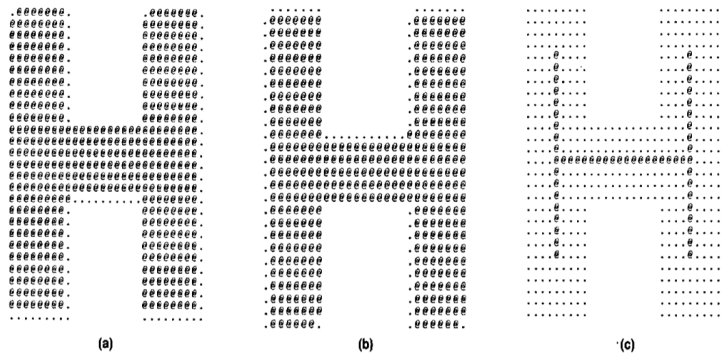


Figure 5: Outcome of thinning

4.3 Feature Extraction

The heart of any optical character recognition system is the formation of feature vector to be used in the recognition stage. Feature extraction can be considered as finding a set of parameters (features) that define the shape of the underlying character as precisely and uniquely as possible. The term *feature selection* refers to algorithms that select the best subset of the input feature set. Methods that create new features based on transformations, or combination of original features are called *feature extraction algorithms*. The features are to be selected in such a way that they help in discriminating between characters. Selection of feature extraction methods is probably the single most important factor in achieving high performance in recognition . A large number of feature extraction methods are reported in literature; but the methods selected depend on the given application.

The features used in our OCR are :

1. **Zoning**: The frame containing the character is divided into four non overlapping zones and the count of the black pixels in each zone is analyzed and form the features. Thus, for each character we get a signature array of length 4 calculated from each zones.

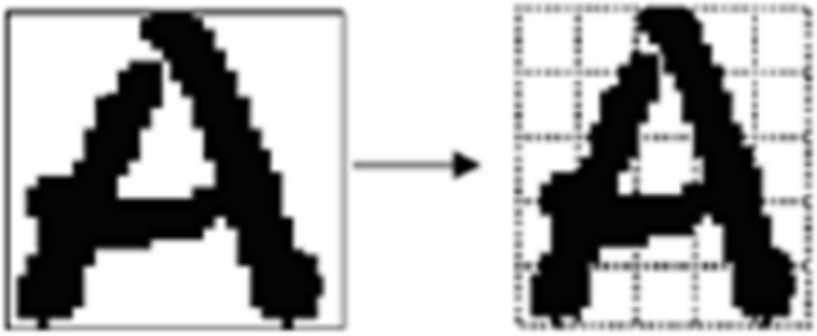


Figure 6: An image before and after Zoning

2. **Crossing**: Crossing counts the number of times the character shape is crossed by vectors along certain directions through the character image. The directions considered by us are the horizontal, vertical and the two diagonal lines. This technique is often used by commercial systems because it can be performed at high speed and requires low complexity. This contributes to the addition of 50 features in our feature set.

3. **Intersection/Junctions and end points in character** : An intersection, also referred to as a junction, is a location where the chain code goes in more than a single direction in an 8-connected neighborhood. Thinned and scaled character image is divided into 4 segments each of size 6 6 pixels wide. For each segment the number of open end points and junctions are calculated. Intersection point is defined as a pixel point which has more than two neighboring pixels in 8-connectivity while an open end has exactly one neighbor pixel. Intersection points are unique for a character in different segment. Thus the number of 8 features within the 4 constituent segments of the character image are collected, out of which first 4 feature represents the number of open ends and rest 4 features represents number of junction points within a segment. These features are observed after image thinning and scaling, as without thinning of the character image there will be multiple open end points and multiple junction points within a segment. For thinning, standard algorithm is used.

4. **Moments** : Moment normalization strives to make the process of recognizing an object in an image size translation and rotation independent. Here the moments of black points about a chosen centre, for example the centre of gravity, or a chosen coordinate system, are used as features. The use of moments results in further contribution of 5 more features to our feature set.

4.4 Image Classification

This is the final stage in recognition of characters. Here labels are assigned to character images based on their features. The relationships between features are also expressed and individual characters are recognized and output them in machine editable form.
Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called **overfitting**. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a **test set**. We performed a random split of our data set into a train set of size 75% and a test set of size 25% for each iteration of the classifier algorithms.

There are numerous machine learning algorithms available for classification problems. The ones used in our OCR are mainly based on decision trees. Decision trees based algorithms for classification have the following advantages :

1. *Speed*: They are very fast to build and test.

2. *Non-linearity*: They work exceedingly well with highly non-linear data.

3. *Visualization*: In some use cases, visualizing the tree might be important. This can't be done in complex algorithms addressing non-linear needs like ensemble methods.

4. *Flexibility*: Unlike other decision-making tools that require comprehensive quantitative data, decision trees remain flexible to handle items with a mixture

of real-valued and categorical features, and items with some missing features. Once constructed, they classify new items quickly.

Following are the two decision trees based classifiers used by us:

1. **Random Forest Classifier**
   Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

2. **Extra Trees Classifier**
   Adding one further step of randomization yields extremely randomized trees, or ExtraTrees. These are trained using bagging and the random subspace method, like in an ordinary random forest, but additionally the top-down splitting in the tree learner is randomized. Instead of computing the locally optimal feature/split combination, for each feature under consideration, a random value is selected for the split. This value is selected from the feature's empirical range (in the tree's training set).

As any other classification algorithm, decision trees based algorithms have their own drawbacks. Some of them are:

1. *Interpretation*: Decision trees never appear to give the right answer, instead they give many possible answers. Changing the root node of the tree to start with a different variable gives a different tree. So, it is difficult to decide which one is correct!

2. *Uncorrelated variables*: Decision trees do not work best if there are a lot of uncorrelated variables. Decision trees work by finding the interactions between the variables. If we have a situation where there are no interactions between variables, linear approaches might be the best.

3. *High variance and unstable* : As a result of the greedy strategy applied by decision trees, variance in finding the right starting point of the tree can greatly impact the final result, i.e small changes early on can have big impacts later.

In order to account for these drawbacks, we decided to use the **Multinomial Logistic Regression** based approach as well. Logistic regression works by multiplying each input by a coefficient, summing them up, and adding a constant. Here, the output is simply a log of the odds ratio (i.e. log(probability of the event happening)/log(probability of the event not happening)).We also experimented with the **Support Vector Machine** based approach but it didn't perform so well.
Finally, we used a voting classifier based ensemble learning method that applies Voting/Majority Rule classification for unfitted estimators. The underlying idea of typical ensemble methods is to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.
While recognising characters, we also find out the labels that the character is likely to be recognised to, in decreasing order of probability.

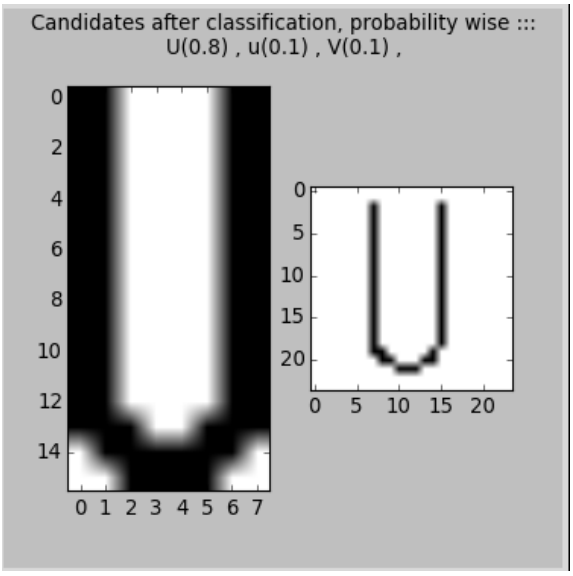


Figure 7: Probability wise chassification of character 'U'

Chapter 5

Results

The OCR was tested on multiple test images . Through training and testing along with a cross validation of 10 folds, the OCR achieved an accuracy of around 80 - 82%.The accuracy depended on the machine learning algorithm. The decision trees based algorithms outperformed the logistic regression by a small factor. The voting classifier accuracy appeared to be slightly greater than that of the constituent learning algorithms alone.

Extra Trees Classifier	82.58%
Random Forest Classifier	78.56%
Logistic Regression	80.54%
Voting Classifier	85.02%

Table 1: Accuracy of each Classifier

5.1 Limitations and Challenges

1. Segmentation of characters if they overlap each other is difficult.
2. The OCR works only for lower and upper case English alphabets as of now.
3. Cursive and slanted letters are difficult to recognize.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Overall experience of taking up this project was truly amazing. In a short span of time we learnt the technical aspects of image processing and machine learning. We learnt about what could be the possible challenges that could arise in character recognition and solutions to some of them. We realised that not all machine learning algorithms work well in all places. Understanding the domain aspect and working as a team to always supporting each other have been the most rewarding experiences.

6.2 Future Scope

An OCR is of great importance in today's world. So it can be improved a lot in the long run. The OCR can be extended to work for digits too. As of now cursive characters cannot be recognized correctly. This can be worked upon in future. We also plan to use neural networks to make the performance of the OCR even better.

References

[1] T. Y. Zhang and C. Y. Suen *A Fast Parallel Algorithm for Thinning Digital Patterns* Communications of the ACM, March 1984, Vol. 27, Number 3 (1984)

[2] Gaurav Kumar, Pradeep Kumar Bhatia *A Detailed Review of Feature Extraction in Image Processing Systems* Fourth International Conference on Advanced Computing & Communication Technologies, Rohtak, 2014, pp. 5-12. doi: 10.1109/ACCT.2014.74 (2014)

[3] M. Abdul Rahiman, M. S. Rajasree *A Detailed Study and Analysis of OCR Research in South Indian Scripts* Proc. of International Conference on Advances in Recent Technologies in Communication and Computing, pp. 31-38. (2009)

[4] Archana S. Sawant, D. G. Chougule *Text Pre-processing and Text Segmentation for OCR* 2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), Pages: 1 - 5, DOI: 10.1109/EESCO.2015.7253643 (2015)

[5] Aparna K G and A G Ramakrishnan *A Complete Tamil Optical Character Recognition System* In Lopresti D., Hu J., Kashi R. (eds) Document Analysis Systems V. DAS 2002. Lecture Notes in Computer Science, vol 2423. Springer, Berlin, Heidelberg (2002)