International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Review on Prediction and Restoration of Partially Visible Letters on the Scanned Document, Printed Document and Images

Sagar Indrajit Jadhav

AISSMS Institute of Information Technology, Pune -411011, Maharashtra, India

ABSTRACT

In recent years, machine learning becomes a new research focus in the field of artificial intelligence. Machine Learning is been successfully used in the complex systems such as: machine vision, speech recognition, natural language processing, web search, recommendation system, intelligent robots, artificial intelligence etc. Machine is invented by human; it can never exceed the level of human intelligence. The Hindi character recognition has been a difficult problem in the field of character recognition. In the languages like English where number characters are smalls, it is difficult to use our traditional algorithm to identify it automatically. But thanks to the further development of machine artificial intelligence, the automatic identification of Hindi and other characters has entered the practical stage. Although many domestic and foreign software vendors have launched a rate of characters automatic identification system which has a good recognition, there is still large room for improvement. The prediction and restoration of partially visible letters on documents and images are now possible with OCR and different machine learning techniques. In many current domestic literatures, mainly papers aim at the research on automatic recognition of a small number of characters. It is difficult to be applied onto a big character set recognition object, mainly software for the undetected text recognition is also build.

Keywords:

OCR , Tesseract ,Binarization

1. Introduction:

All the images than are scanned have some partial visibility. The reasons may be due to increase in the ink, too much of darkness, ink scattered all over the paper, partially printed letters. Based on the sentence or using the meaning of the letter, the intention is to restore the actual letter using image processing technique. Also, older printed documents may have issues of missing letters or erased letters. This will lose the importance of storing the document.

In this paper, we have summarized multiple techniques for prediction and restoring the partially visible letters on the scanned document, printed document, and image.

Nomenclature

Binarization of : grey scaling the image
Tesseract Technique: Processing of image with software algorithm named tesseract

* *Corresponding author.* Tel.: +91 7020939651
E-mail address: sagar21202@gmail.com

2. Analysis of Existing Approaches.

There are many OCR based applications available that can extract text from an image but they lack the accuracy to provide the same function for handwritten text. In addition, the process of typing every word would be a time consuming and daunting task. If we look about the benefit of converting handwritten documents into online computerized documents, then it is easy to handle them all. In addition, new information can be added and modified something that cannot be done on a simple scanned text document.

2.1. System Architecture:

- Proposed Model/Architecture/Algorithm
- Image acquisition
- Pre-processing
- Text recognition
- Pattern matching
- Feature extraction
- Postprocessing

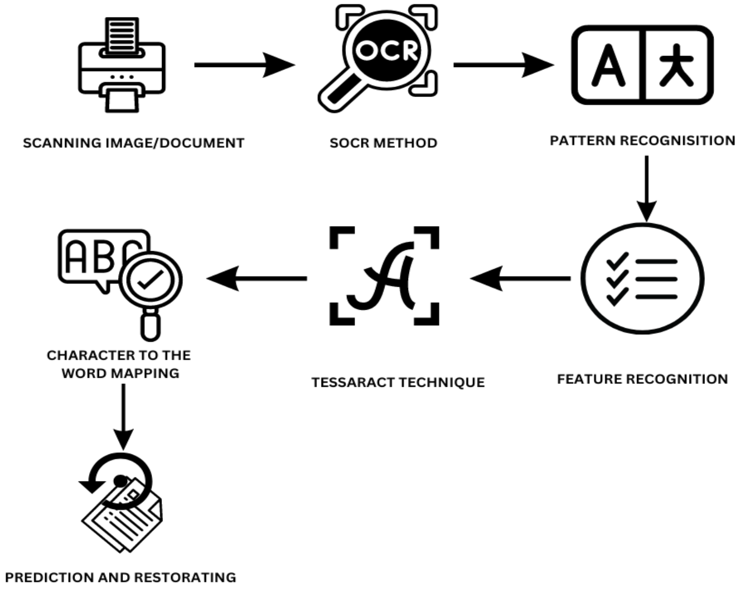


Figure-1: Steps in Text detection and restoration in natural scene images.

OCR PROCESS FLOW

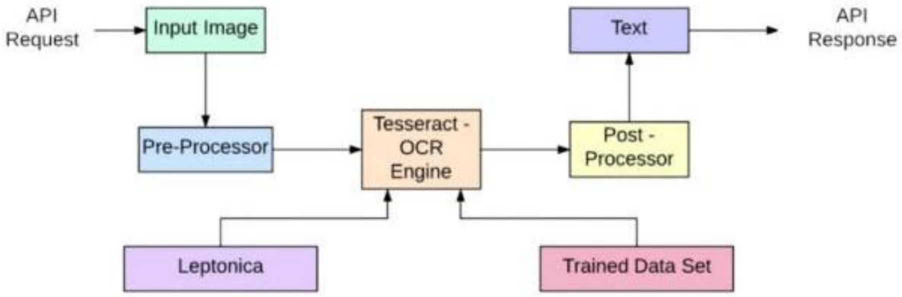


Figure-2: OCR process flow

OCR:

Techniques in OCR:

Pre-processing

OCR first pre-processes the image or document for improvement in the recognition of data. For this pre-processing, there are some techniques as follows:

DE speckle – It removes the positive and negative spots with grey scale.

De-skew –If the document is not aligned [properly while scanning, then we must make it perfectly horizontal for better recognition.

Binarization –Here grey scale is used for Binarization is used for separation of text. There are so many commercial algorithms used for this process. The quality of the character recognition stage and the careful decisions, since the quality of the binarization method employed to obtain the binary result depends on the type of the input image like scanned document, scene text image, historical degraded document etc provided.

Line removal – Here boxes and line cleaning are done efficiently.

Layout analysis -It deals with structure of data like column row etc.

Line and word detection – It checks the shapes of word or character in the scanned document.

Script recognition – Here, the type of language is detected and identified of that the right OCR can be invoked properly for precision.

Segmentation Single character which is broken into multiple pieces due to artifacts must be connected for the recognition purpose.

Post-processing

Lexical analysis improves the accuracy of OCR recognition. If there are words in document which are not present in the lexicon, then it becomes difficult for the software to recognise it. For fast recognition, software like Tesseract maintains the dictionary, this improves the accuracy of the recognition software. Output of the OCR software can be formatted like the document or it can be in text form. It depends on the software and output. It can be like a plain txt document or a well formatted pdf. It also includes the images in their proper aligns and effects. Machine learning algorithms like K nearest neighbour are used for nouns and some common words which are not grammatically correct. Grammar analysis also helps in recognising the unidentified words and those can be guessed, this also increases the accuracy.

Tesseract:

Tesseract is a free, open-source OCR engine that was developed at HP super-nova.

Tesseract is highly customizable and we can operate it using most languages and vertical text. The tesseract engine hardly written in the c and c++language.It does not have any GUI to interact with user but it provides different kind of APIs. These APIs are used by the user by importing their corresponding modules. Using tesseract, we can recognize more than 100 languages by just downloading their configurations. Basically, tesseract recognizes the text from images with high accuracy. The accuracy to recognise the text depends upon the type of input image given.

Tesseract takes different types of input images like as PDF, JPG, PNG, TIFF and after recognition, the text from images is stored in to the pdfs format.

After installing tesseract, we need to install pytesseract. Pytesseract is an OCR tool for python that also serves as a wrapper for the Tesseract-OCR Engine. It is developed by the google. The current version of pytesseract is 5.2.0 which released on 6 July 2022.

Pre-processing:

The experiment aimed at measuring out-of-the-box performance, so documents were submitted without further pre-processing using the OCR engines' default settings. Footnote7 While this is an uncommon use of Tesseract, it treats the engines equally and helps highlight the degree to which Tesseract is dependent on image pre-processing.The English corpus was submitted to all three OCR engines in a total of 42,504 document processing requests. The Arabic corpus was only submitted to Tesseract and Document AI—since tesseract does not support Arabic—for a total of 8800 processing requests.

The Tesseract processing was done in R with the package tesseract (v4.1.1). For Tesseract, it was carried out via the R package paws (v0.1.11), which provides a wrapper for the Amazon Web Services API. For Document AI, I used the R package diary (v0.8.0) to access the Document AI API v1 endpoint. The processing was done in April and May of 2021 and took an estimated net total of 150–200 h to complete. The Document AI and Tesseract APIs processed documents at a rate of approximately 10–15 s per page. Tesseract took 17 s per page for Arabic and 2 seconds per page for English on a Linux Desktop with a 12-core, 4.3 GHz CPU and 64GB RAM.

2.1 *Text recognition*

For producing possible sets of character there are mainly two OCR algorithms that are recommended.

In matrix matching image is stored pixel by pixel, it is called as image correlation or pattern recognition. The techniques like OCR success well in typed letter but not well on handwritten documents. The lines, loops, inks, loops are extracted and identified by feature extraction. The dimensionality reduction is necessary for easy and smooth process. This is done by Extraction features.

The artificial intelligence technique like computer vision is also comes in OCR technology implementation. This makes possible the recognition of handwritten text in most of newly developed software's.

KNN algorithms i.e., K nearest neighbour algorithm is used for image identification with stored glyphs.

Tesseract is also used for character recognition. The adaptive recognition makes easy to identify the next remaining characters and half of text easily.

This makes ease in identifying the blurred fonts and distorted letters in distorted images.

The latest and modern software's like Google Docs OCR, Transym and FineReader are best in this area of Text detection and extraction. Software like OCRopus and Tesseract are based on neural network. These software focuses on full line of text instead of single character.

The techniques like Iterative OCR can crop and sector the documents easily for ease in identification. The OCR software work on these sectors effectively and this improves the accuracy of recognition. Several patents were issued across the globe for such techniques.

The output result of this technique is stored in the document formats like XML and hOCR.

2.2 Image Pre-processing Methods

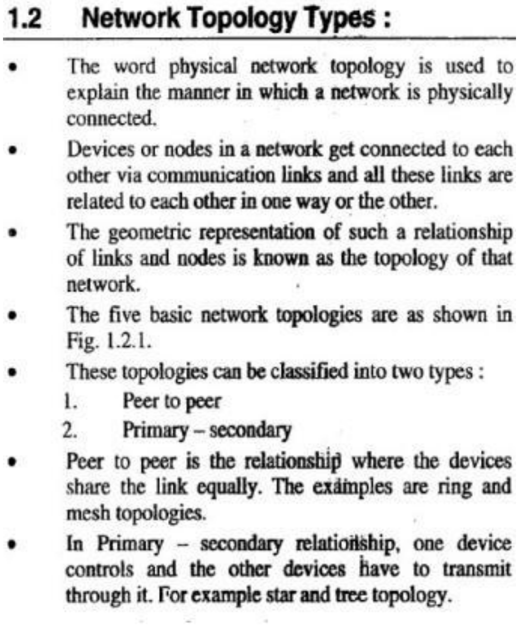
Design:

In this experiment I have taken two input images with high resolutions but in the first image I was recognise vertical text and kind of text fonts in English language. In another input image in which I was recognize the horizontal or curved text with different fonts.

Processors:

To recognise the first image which contained vertical text I have chosen OCR and for next image containing different styles and horizontal text I have used Tesseract engine.

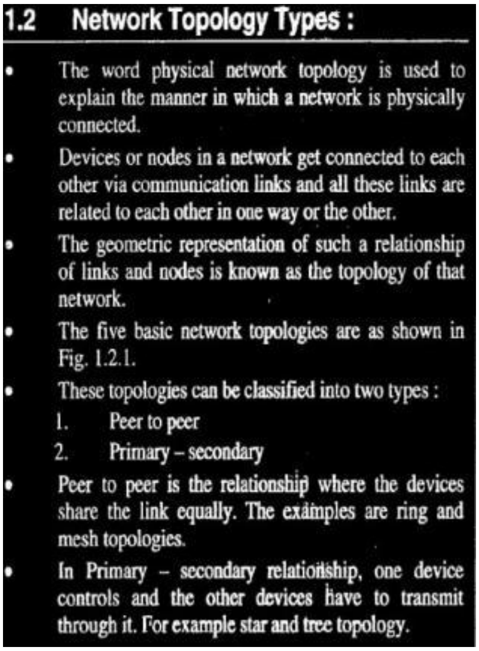
1.Original



1.2 Network Topology Types :

- The word physical network topology is used to explain the manner in which a network is physically connected.
- Devices or nodes in a network get connected to each other via communication links and all these links are related to each other in one way or the other.
- The geometric representation of such a relationship of links and nodes is known as the topology of that network.
- The five basic network topologies are as shown in Fig. 1.2.1.
- These topologies can be classified into two types :
  1. Peer to peer
  2. Primary – secondary
- Peer to peer is the relationship where the devices share the link equally. The examples are ring and mesh topologies.
- In Primary – secondary relationship, one device controls and the other devices have to transmit through it. For example star and tree topology.

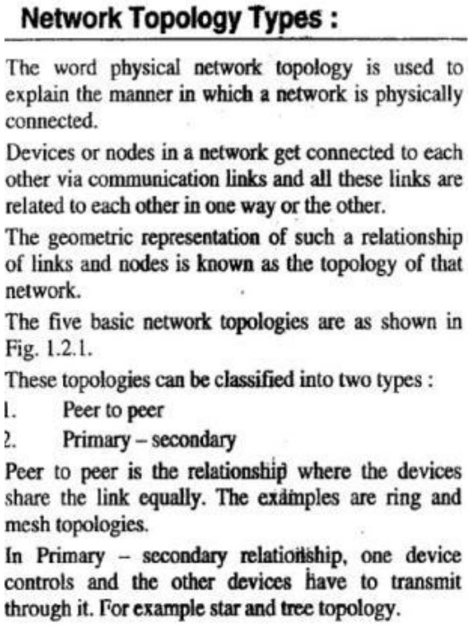
2. Inverted Image



1.2 Network Topology Types :

- The word physical network topology is used to explain the manner in which a network is physically connected.
- Devices or nodes in a network get connected to each other via communication links and all these links are related to each other in one way or the other.
- The geometric representation of such a relationship of links and nodes is known as the topology of that network.
- The five basic network topologies are as shown in Fig. 1.2.1.
- These topologies can be classified into two types :
  1. Peer to peer
  2. Primary – secondary
- Peer to peer is the relationship where the devices share the link equally. The examples are ring and mesh topologies.
- In Primary – secondary relationship, one device controls and the other devices have to transmit through it. For example star and tree topology.

3.Binarization



Network Topology Types :

The word physical network topology is used to explain the manner in which a network is physically connected.

Devices or nodes in a network get connected to each other via communication links and all these links are related to each other in one way or the other.

The geometric representation of such a relationship of links and nodes is known as the topology of that network.

The five basic network topologies are as shown in Fig. 1.2.1.

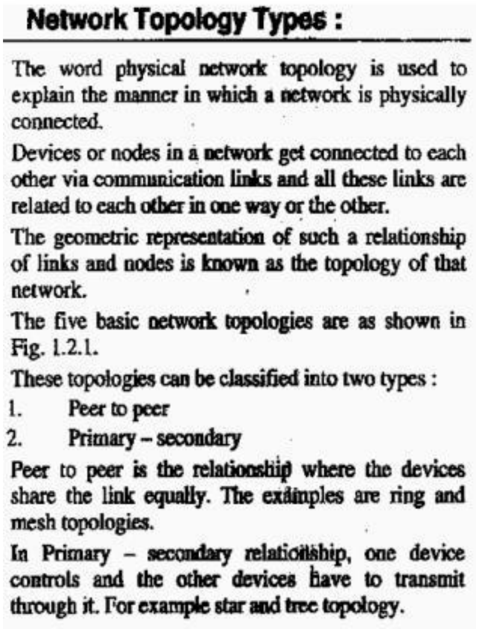
These topologies can be classified into two types :

1. Peer to peer
2. Primary – secondary

Peer to peer is the relationship where the devices share the link equally. The examples are ring and mesh topologies.

In Primary – secondary relationship, one device controls and the other devices have to transmit through it. For example star and tree topology.

4. Thresholding Image.



Network Topology Types :

The word physical network topology is used to explain the manner in which a network is physically connected.

Devices or nodes in a network get connected to each other via communication links and all these links are related to each other in one way or the other.

The geometric representation of such a relationship of links and nodes is known as the topology of that network.

The five basic network topologies are as shown in Fig. 1.2.1.

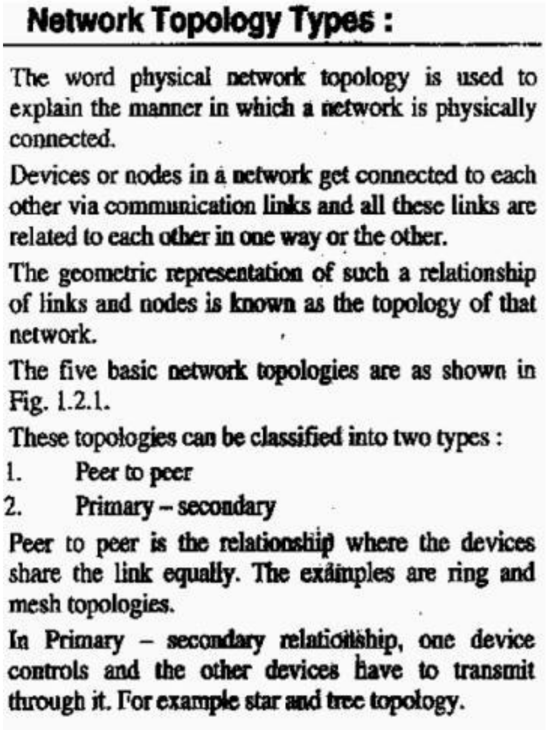
These topologies can be classified into two types :

1. Peer to peer
2. Primary – secondary

Peer to peer is the relationship where the devices share the link equally. The examples are ring and mesh topologies.

In Primary – secondary relationship, one device controls and the other devices have to transmit through it. For example star and tree topology.

5.Noise Removal



Network Topology Types :

The word physical network topology is used to explain the manner in which a network is physically connected.

Devices or nodes in a network get connected to each other via communication links and all these links are related to each other in one way or the other.

The geometric representation of such a relationship of links and nodes is known as the topology of that network.

The five basic network topologies are as shown in Fig. 1.2.1.

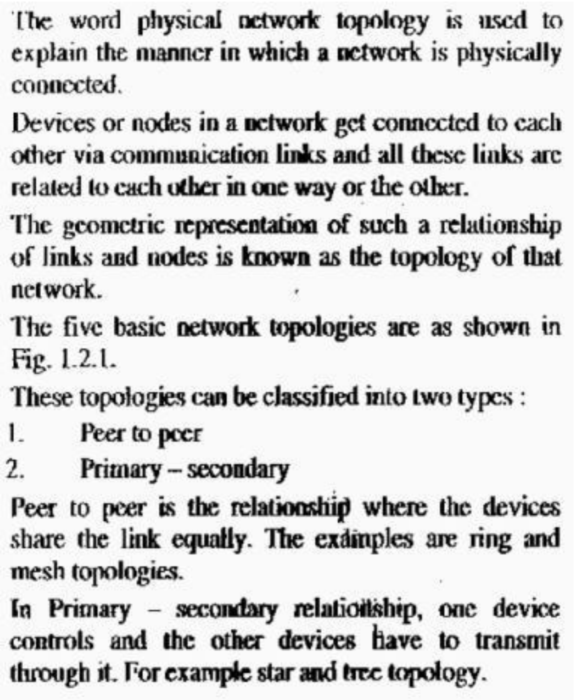
These topologies can be classified into two types :

1. Peer to peer
2. Primary – secondary

Peer to peer is the relationship where the devices share the link equally. The examples are ring and mesh topologies.

In Primary – secondary relationship, one device controls and the other devices have to transmit through it. For example star and tree topology.

6. Dilated



The word physical network topology is used to explain the manner in which a network is physically connected.

Devices or nodes in a network get connected to each other via communication links and all these links are related to each other in one way or the other.

The geometric representation of such a relationship of links and nodes is known as the topology of that network.

The five basic network topologies are as shown in Fig. 1.2.1.

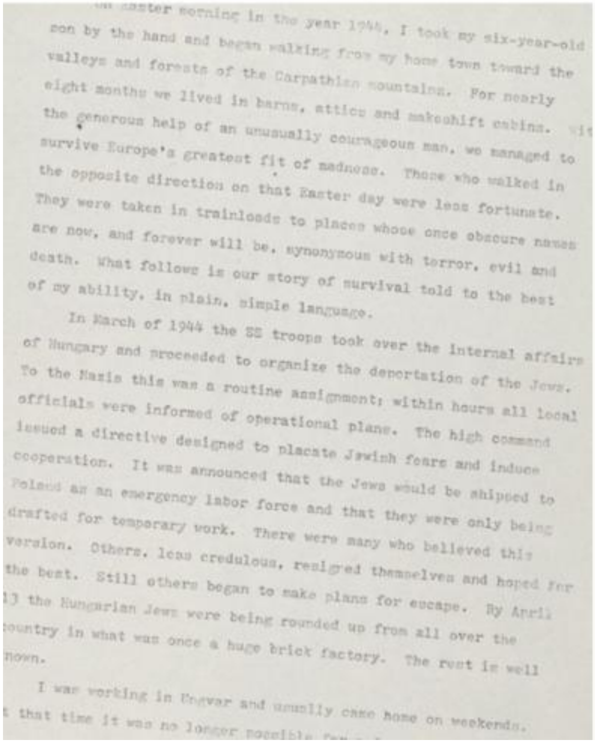
These topologies can be classified into two types :

1. Peer to peer
2. Primary – secondary

Peer to peer is the relationship where the devices share the link equally. The examples are ring and mesh topologies.

In Primary – secondary relationship, one device controls and the other devices have to transmit through it. For example star and tree topology.

6.Horizontal Image.

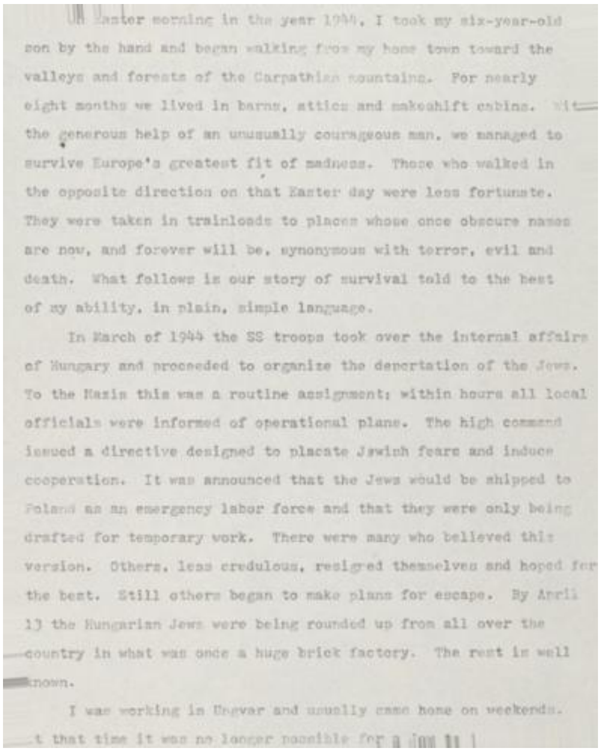


Easter morning in the year 1944, I took my six-year-old son by the hand and began walking from my home town toward the valleys and forests of the Carpathian mountains. For nearly eight months we lived in barns, attics and makeshift cabins. With the generous help of an unusually courageous man, we managed to survive Europe's greatest fit of madness. Those who walked in the opposite direction on that Easter day were less fortunate. They were taken in trainloads to places whose once obscure names are now, and forever will be, synonymous with terror, evil and death. What follows is our story of survival told to the best of my ability, in plain, simple language.

In March of 1944 the SS troops took over the internal affairs of Hungary and proceeded to organize the deportation of the Jews. To the Nazis this was a routine assignment; within hours all local officials were informed of operational plans. The high command issued a directive designed to placate Jewish fears and induce cooperation. It was announced that the Jews would be shipped to Poland as an emergency labor force and that they were only being drafted for temporary work. There were many who believed this version. Others, less credulous, resigned themselves and hoped for the best. Still others began to make plans for escape. By April 13 the Hungarian Jews were being rounded up from all over the country in what was once a huge brick factory. The rest is well known.

I was working in Ungvar and usually came home on weekends. t that time it was no longer possible fo

7.Vertical Image.

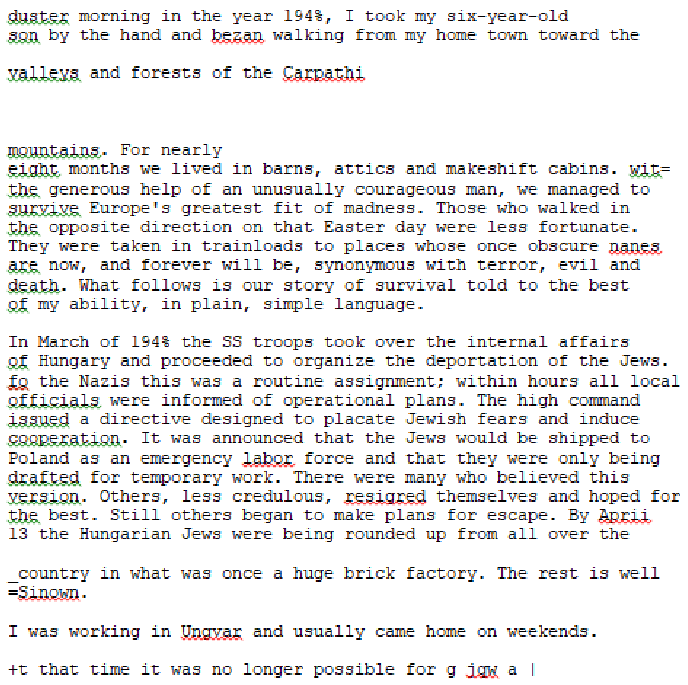


Easter morning in the year 1944, I took my six-year-old son by the hand and began walking from my home town toward the valleys and forests of the Carpathian mountains. For nearly eight months we lived in barns, attics and makeshift cabins. Wit the generous help of an unusually courageous man, we managed to survive Europe's greatest fit of madness. Those who walked in the opposite direction on that Easter day were less fortunate. They were taken in trainloads to places whose once obscure names are now, and forever will be, synonymous with terror, evil and death. What follows is our story of survival told to the best of my ability, in plain, simple language.

In March of 1944 the SS troops took over the internal affairs of Hungary and proceeded to organize the deportation of the Jews. To the Nazis this was a routine assignment; within hours all local officials were informed of operational plans. The high command issued a directive designed to placate Jewish fears and induce cooperation. It was announced that the Jews would be shipped to Poland as an emergency labor force and that they were only being drafted for temporary work. There were many who believed this version. Others, less credulous, resigned themselves and hoped for the best. Still others began to make plans for escape. By April 13 the Hungarian Jews were being rounded up from all over the country in what was once a huge brick factory. The rest is well known.

I was working in Ungvar and usually came home on weekends. t that time it was no longer possible for a

Table 1: Summary of the Current Text Detection Methods



```
duster morning in the year 194%, I took my six-year-old
son by the hand and bezan walking from my home town toward the

valleys and forests of the Carpathi


mountains. For nearly
eight months we lived in barns, attics and makeshift cabins. wit=
the generous help of an unusually courageous man, we managed to
survive Europe's greatest fit of madness. Those who walked in
the opposite direction on that Easter day were less fortunate.
They were taken in trainloads to places whose once obscure nanes
are now, and forever will be, synonymous with terror, evil and
death. What follows is our story of survival told to the best
of my ability, in plain, simple language.

In March of 194% the SS troops took over the internal affairs
of Hungary and proceeded to organize the deportation of the Jews.
fo the Nazis this was a routine assignment; within hours all local
officials were informed of operational plans. The high command
issued a directive designed to placate Jewish fears and induce
cooperation. It was announced that the Jews would be shipped to
Poland as an emergency labor force and that they were only being
drafted for temporary work. There were many who believed this
version. Others, less credulous, resigred themselves and hoped for
the best. Still others began to make plans for escape. By Aprii
13 the Hungarian Jews were being rounded up from all over the

_country in what was once a huge brick factory. The rest is well
=Sinown.

I was working in Ungvar and usually came home on weekends.

+t that time it was no longer possible for g jgw a |
```

3 *Referenced articles*

Research Article (Author/Year)	Paper Name	Methods /Techniques	Link
Rehan Tariq Nov,2015	A Review on Text Detection Techniques	1. OCR based recognition 2. Block segmentation	https://www.researchgate.net/publication/299998062_A_Review_on_Text_Detection_Techniques
Artur Haponik CEO & Co-Founder jully23,2021	Text Extraction from Images Using Machine Learning	1. Data acquisition 2.Scanned Image Files. 3.Preprocessing 4.Noise Reduction 5. Normalization.	https://addepto.com/blog/text-extraction-from-images-using-machine-learning/#
GidiShperber Oct 22, 2018	Text Extraction and Detection from Images using Machine Learning Techniques	c computer vision techniques lized deep learning approaches rd deep learning approach	https://towardsdatascience.com/a-gentle-introduction-to-ocr-ee1469a201aa

Avoid hyphenation at the end of a line. Symbols denoting vectors and matrices should be indicated in bold type. Scalar variable names should normally be expressed using italics. Weights and measures should be expressed in SI units. All non-standard abbreviations or symbols must be defined when first mentioned, or a glossary provided.

4. Conclusion

Demand of human–computer interaction has elevated this field. The researchers of this field are putting in great efforts for developing character recognition systems which can be practically implemented to aid society. The techniques like OCR and Tesseract discussed in this paper are well with a common aim of improvement in recognition rates. From the survey, it can be concluded that implemented recognition techniques correctly recognizes characters of good quality, but fails in recognizing overlapped, similar shaped, or ambiguous characters. This seminar is undertaken to explain how the partially visible data can be extracted from damaged document or image and evaluate. This study has found that generally algorithms like OCR, tesseract etc software's are used efficiently for the purpose.

Acknowledgements

The success of any project or a research paper depends largely on the encouragement and guidelines of many others. This Research Seminar would not have been possible without their support. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project.

First and foremost, we wish to record our sincere gratitude to the Management of this college and to our Respected Principal Dr.P.B. Mane, for his constant support and encouragement in the preparation of this report and for the availability of library facilities needed to prepare this report. Our sincere thanks to Dr.M.A. Thalor, Head, Department of Information Technology, AISSMS IOIT, Pune who was abundantly helpful and offered invaluable assistance, support and guidance.

We express our sincere gratitude to our Guide, Smt. J.C. Pasalkar for guiding us in investigations of this Seminar and in carrying out experimental work. Our numerous discussions were extremely helpful. We are highly indebted to him for his guidance and constant supervision as well as for providing necessary information regarding the Seminar & also for his support in completing the Seminar. We hold him in esteem for guidance, encouragement and inspiration received from him. I thank all for continuous strength we were bestowed for completion of this Seminar.

REFERENCES

1. Lopresti, D.: 'Optical character recognition errors and their effects on natural language processing', Int. J. Doc. Anal. Recognition., 2009, 12, (3), pp. 141–151[2]
2. Pal, U., Chaudhuri, B. B.: 'Printed Devnagari script OCR system', Vivek,1997, 10, (1), pp. 12–24[3] Bajaj, R., Dey, L., Chaudhury, S.: 'Devnagari numeral recognition by combining decision of multiple connectionist classifiers', Sadhana, 2002, 27, (1), pp. 59–72[4]
3. Pal, U., Sharma, N., Wakabayashi, T., et al.: 'Off-line handwritten character recognition of Devanagari script'. Int. Conf. Document Analysis and Recognition, Parana, Brazil, Sept 2007, pp. 496–500[5]
4. Sharma, N., Pal, U., Kimura, F., et al.: 'Recognition of off-line handwritten Devanagari characters using quadratic classifier'. Proc. Indian Conf. Computer Vision, Graphics and Image Processing, Madurai, India, Dec 2006, pp. 805–816[6]
5. Hanmandlu, M., Murthy, O. V. R., Madasu, V. K.: 'Fuzzy model-based recognition of handwritten Hindi characters. Biennial Conf. Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications, Glenelg, Australia, December 2007, pp. 454–461[7]
6. Chaudhuri, B.B., Pal, U.: 'Skew angle detection of digitized Indian script documents', IEEE Trans. Pattern Anal. Mach. Intell., 1997, 19, (2), pp. 182–186[8]
7. Sharma, P. K., Dhingra, K. D., Sanyal, S.: 'A rule-based approach for skew correction and removal of insignificant data from scanned text documents of Devanagari script'. IEEE Conf. Signal-Image Technologies and Internet-Based System, Shanghai, China, December 2007, pp. 899–903[9]
8. Chaudhuri, A., Mahdavia, K., Badelia, P., et al.: 'Optical character recognition systems for Hindi language', in 'Studies in fuzziness and soft computing' (Springer, Cham, 2017, 1st edn.), pp. 193–216[10]
9. Pittman, J. A.: 'Handwriting recognition: tablet PC text input', J. Comput.,2007, 40, pp. 49–54