

OPTICAL CHARACTER RECOGNITION USING CBIR

CBIR PROJECT-REVIEW

19BCE1646 SHASHIDHAR REDDY MALIGIREDDY
20MIA1013 JAYANTH GURAJADA
19BCE1526 LAXMAN T

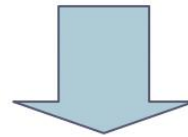
- Optical Character Recognition(OCR) is the mechanical or electrical conversion of images of typewritten or printed text into machine-encoded text. It is widely used as a form of data entry from printed paper data records, whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation. 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.

- Every optical image when converted into grey scale can be considered as a matrix with 1's and 0's as its elements. The theory behind this optical character recognition is division of the image into suitable number of pixels which represent the element of the matrix as stated above and comparing these pixels with those of pre-defined set of templates. .

- For a single character , it can be directly compared with templates whereas for a word which consists of several characters it is primarily divided into group of clusters each consisting of a single character which is compared with the given templates

An example of OCR working shown below.

of descriptive bibliographies of authors and presses. His ubiquity in the broad field of bibliographical and textual study, his seemingly complete possession of it, distinguished him from his illustrious predecessors and made him the personification of bibliographical scholarship in his time.



of descriptive bibliographies of authors and presses. His ubiquity in the broad field of bibliographical and textual study, his seemingly complete possession of it, distinguished him from his illustrious predecessors and made him the personification of bibliographical scholarship in his time.

PreProcessing

Deals with improving quality of the image for better recognition by the system. Techniques include -

- De-skew
- Despeckle
- Binarization
- Line Removal
- Zoning etc..

Character recognition


There are basic types of core OCR algorithm which may produce a ranked list of candidate character-

- 1) Matrix matching
- 2) Feature extraction

Post Processing :

OCR accuracy can be increased if the output is constrained by lexicon. Eg, all the words in the English Language can be problematic if the document contains the words that are not in the lexicon, like proper nouns.

- A few examples of OCR applications are listed here. The most common for use OCR is the first
- item; people often wish to convert text documents to some sort of digital representation.
- 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

- 
- Machine replication of human functions, like reading, is an ancient dream.
 - However, over the last five decades, machine reading has grown from a dream to reality. Optical character recognition has become one of the most successful applications of technology in the field of pattern recognition and artificial intelligence.
 - Many commercial systems for performing OCR exist for a variety of applications, although the machines are still not able to compete with human reading capabilities. Optical Character Recognition deals with the problem of recognizing optically processed characters.

- Optical recognition is performed off-line after the writing or printing
- has been completed, as opposed to on-line recognition where the computer recognizes the characters as they are
- drawn. Both hand printed and printed characters may be recognized, but the performance is directly dependent
- upon the quality of the input documents. Progress in OCR has been steady if not spectacular since its commercial
- introduction at the Reader's Digest in the mid-fifties.

Developments of OCR

- Historically character recognition system has evolved in three ages , namely the periods cited denoting as 1900-1980 (early ages) – The history of character recognition can be traced as early as 1900. When the Russian Scientist Tyering attempted to develop an aid for visually handicapped.
- The first character recognizers appeared in the middle of 1940s with the development of digital computers. The early work on the automatic recognition of characters has been concentrated either upon machine printed text or upon small set of well distinguished hand written text or symbols. The commercial character recognizers were available in 1950s.

- **1980-1990 Developments** – The studies until 1980 suffered from the lack of powerful computer hardware and data acquisition devices. However, the character recognition research was focused on basically the shape recognition techniques without using any semantic information.
- **After 1990 advancements** – The real progress on character recognition system is achieved during this period, using the new development tools and methodologies, which are empowered by continuously growing information technologies.

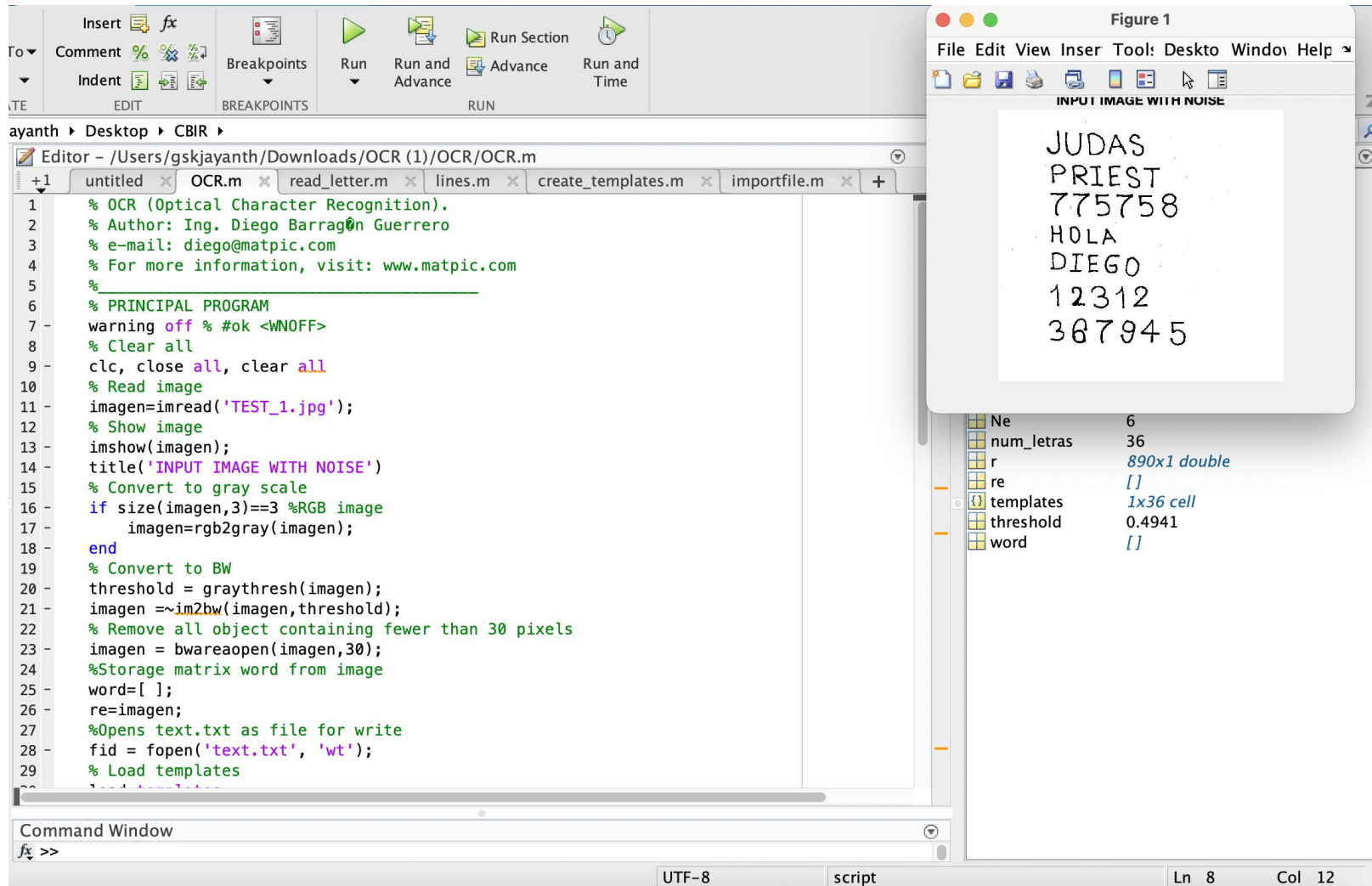
- In the early **1990's**, Image processing and Pattern recognition techniques are efficiently combined with the Artificial Intelligence methodologies.
- Nowadays in addition to the more powerful computers and more accurate electronic equipments such as scanners, cameras and electronic tablets, we have efficient, modern use of methodologies such as neural network
- Character Recognition is one of the vital tasks in Pattern Recognition.

- Indian Character Recognition Not many attempts have been made on the character recognition of Indian
- character sets. However, some major works are reported on Devanagari. Some attempts are also reported on Tamil,
- Kannada, Gujarathi, Bengali, Malayalam and Telugu.

- The popularity and use of Character Recognition is increasing day by day with the advent of new, fast and efficient hardware and software. But automatic character recognition of Indian languages is still in preliminary stage and hence there is a need of lot of research to address the various issues and their complexities. There are many factors such as noise, various font sizes, broken lines or characters, quality of the image, problems in segmentation that influence recognition process.

- India is a multi lingual country; so many more efficient and real-time text recognizers are required. A good text recognizer has many commercial and practical applications. Hence there is a need to develop a very good character recognition system which must achieve highest accuracy.

MATLAB CODE AND OUTPUT



[1] J.T. Tou and R.C. Gonzalez, Pattern Recognition Principles, Addison-Wesley Publishing

- Company, Inc., Reading, Massachusetts, 1974
- https://www.academia.edu/59323047/Recognition_of_Handwritten_Textual_Annotations_using_Tesseract_Open_Source_OCR_Engine_for_information_Just_In_Time_iJIT
- https://www.academia.edu/58585873/Automatic_Number_Plate_Recognition_System
- https://www.academia.edu/58653673/Design_of_a_Neural_Network_Based_Optical_Character_Recognition_System_for_Musical_Notes?source=swp_share

- [1] G. Nagy, S. Seth, and M. Viswanathan, “A Prototype Document Image Analysis System for Technical Journals,” *Computer*, vol. 25, no. 7, pp. 10-22, July 1992.
- [2] Fujisawa, Yasuaki Nankano, and Kiyomichi Kurino “Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis” *Proceedings of The IEEE*. Vol. 80. No. 7. July 1992.
- [3] L. O’Gorman, “The Document Spectrum for Page Layout Analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.15,no 11,pp.1162-1173,Nov 1993.

- [4] H.S. Baird, H. Bunke, P. Wang and H.S. Baird “Background Structure in Document Images,” Document Image Analysis, eds., pp. 17-34, World Scientific, 1994.
- [5] Sylvester and S. Seth, “A Trainable, Single-Pass Algorithm for Column Segmentation,” Proc. Int’l Conf. Document Analysis and Recognition, pp. 615- 618, Aug. 1995.
- [6] I. Guyon, R.M. Haralick, J.J. Hull and I.T. Phillips, “Data Sets for OCR and Document Image Understanding Research,” Handbook of Character Recognition and Document Image Analysis, H. Bunke and P. Wang, eds., pp. 779-799, World Scientific, 1997.