JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANALYSIS OF OPTICAL CHARACTER RECOGNITION

Dr.P.Selvarani[1] K.Sagar[2] Subham kumar[3] Nellore Vinay[4] Phanindra5
Associate Professor [1] UG Scholar [2,3,4,5,]
Department of Computer Science and Engineering
Vel Tech High tech Engineering College, Chennai, Tamil Nadu, India.

Abstract

OCR expanded as optical character recognition, is a subset of pattern recognition systems and is an active research area of significant scientific and practical importance. His OCR technique uses an optical scanner to first digitise typed letters. Following the identification and segmentation of each character, the resulting character picture is forwarded to the preprocessor for normalization. The characteristics are then classified based on certain traits that were extracted from them. There are numerous methods for feature extraction, each with a different purpose, and feature extraction is a crucial step. This paper provides an overview of feature extraction methods for character recognition. The selection of the feature extraction method is the most important factor in achieving high recognition performance in a character recognition system. Different feature extraction methods are designed for different character representations. The most effective feature extraction technique should be chosen for a given application after a number of promising methods have been identified and empirically tested. Today, many paper documents are digitized, facilitating information processing such as searching, analysis, and conversion. This article presents his OCR usability analysis for automated data entry. Our research looked at the entire process of digitizing print. Scanning, data indexing, approval, market. The study showed the recognition of printed text is reliable and data processing is significantly faster. On the contrary, handwritten text seemed difficult for his OCR system to recognize.

Keywords: String recognition, Optical measurement, OCR classification, printed text identification, barcodes, data indexation.

I. INTRODUCTION

For many years, the field of optical character recognition (OCR) has been evolving. Digitizing a document image into individual characters is described as the procedure, developing his OCR with human-like capabilities remains a challenge despite decades of intense research. is. Remains an open topic. Because of this challenge, researchers in industry and academia are turning to optical character recognition. In recent years, the number of students. The number of research institutes and companies involved in character recognition research is increasing dramatically. This study aims to summarizes previous research at the field of OCR provide an overview. It describes various aspects of OCR[11] and provides relevant suggestions for resolving OCR problems.

Pattern identification is the association of physical objects or events with one of several predefined objects. category. and many other apps. B. – Radar signal classification and analysis, character (letters or numbers) recognition, and handwriting analysis [12] (notepad computer). various uses. These include zip codes, check reading, tablet

computers, and personal digital assistants (PDAs). Verification of signatures, form processing, and recognition. OCR is crucial for text entry (office automation), data entry (banking environments), and process automation (mail sorting), which are the fundamentals of the language that are utilised to create the different language structures. component. Alphabets are used as characters, whereas words, strings, sentences, etc. are used as structures. Recognition techniques, [4] as a subset of pattern recognition, confer specific symbolic identities offline. A printed or written image of a character. Character recognition is commonly known as optical character recognition because the recognition of optically processed characters involves magnetic processing. Character detection technology has the benefit of saving both time and labour. Provides a fast and easy alternative enter manually. The recognition of any character is the process of reading its image, Preprocess the image, extract appropriate image features, and classify characters based on what is extracted Image features and known features are stored in vectors and recognize images accordingly Similarity between the loaded image and the image database.

Many documents are now produced in paper form. Documents are copied repeatedly, Changes were made in subsequent stages of processing, and today there are many different copies of this type, but the inefficiency of the process led to the decision to digitize the document. Working with files are cheap, it does not require document storage space and can handle conventional documents. Many times each document exists in one copy, so changes and annotations appear in all documents user. Scanning, indexing (data input), and presenting the digitised materials are the three key processes in the digitization process. This white paper concentrated on usage, especially indexing. Automated system for this process the popularity of character recognition has increased significantly. It has become an important research area. Character recognition is a research area use techniques to classify your character's input according to pre-defined classes. Several algorithms have been proposed, but the choice of algorithm and classifier varies from problem domain to problem domain.

II. RELATED WORKS

Although character recognition is not a new issue, the origins of the issue may be found in older systems. Computer invention. His early OCR systems were machines, not computers. I could understand the characters, but it was very slow and inaccurate. 1951 M. Shepard invents his GISMO of reading and robotics. This can be considered his first OCR work in modern times. GISMO can read both notes and words on printed pages separately. However, you can Recognizes 23 characters. The machine also has the ability to copy typed pages. In 1954, Rainbow developed a machine that allowed him to read each capital letter of English as it was typed minute. Early OCR systems from the business were criticised for having faults and having sluggish recognition rates. There was a lot of study done on the issue in the 1960s and 1970s, but government organizations and major businesses like banks, newspapers, and airlines only continued to expand the field. Because of the complexity involved in cognition he felt the need to use three standardized OCR fonts. ANSI and EMCA therefore developed his OCRA and OCRB. 1970 provided relatively acceptable acceptance rates. In the last thirty years. A lot of research has been done on OCR. This led to document image analysis (DIA), multilingual, handwriting, and omni font OCR. Despite this extensive research effort, The ability of machines to read text reliably is well below that of humans. Therefore, OCR research is currently being conducted to improve the accuracy and speed of OCR for various types of printed/handwritten documents. An environment without borders. There was no open source or paid software available for such difficult languages as Sindhi or Urdu.

This post provided an overview of various OCR techniques. OCR requires several steps, including capture, pre-processing, segmentation and feature extraction, classification, and post-processing [10] ; it is not an atomic operation. This document describes each step in detail. Usage As a future work, these methods can be combined to develop an efficient OCR system. OCR systems may also be utilized for a wide range of real-time practical applications, including license plate recognition, smarts, libraries, and many more. Character identification in languages like Arabic, Sindhi, and Urdu remains a problem despite much study in the topic. Future study will provide an overview of his OCR methods for various languages. Systems for multilingual character recognition are other crucial research fields. Finally, the use of OCR Real systems is still an area of active research. Character Recognition Through Optics a prerequisite for programmers' that take input from typewriters or handwritten documents. Recognition of written text produces successful outcomes. Nearly all of the information was accurate. Few of them

are known. The field had problems, but the scanning procedure rendered it unreadable or corrupted. According to this ranking, his LBP with SVM produces the greatest outcomes, as seen by his 96.5 Curacy. Our research showed Data manually rewritten from forms by experienced users has fewer errors than data Recognized by OCR/systems. In this whitepaper, we reviewed and explained feature extraction. a method of classifying characters for optical character recognition. This topic has been the subject of much inquiry. The accuracy of feature extraction and categorization approaches is still being improved, though. The numerous feature extraction and categorization techniques, however, are highly powerful. Easy-to-implement template-matching method Improved algorithm simplicity and flexibility when changing recognized classes. Is Recognition is most powerful in monotypes and his uniform one-column page and also saves time. No sample training is required, but the template can only recognize the same font size and rotation. Neural networks are capable of abstractly recognizing characters. damaged text and scanned documents

III PROPOSED METHODOLOGY

There are various methods of the character identification which can be classified into the following groups:
- Pattern systems
- Structural systems
- Feature systems
- Neural network systems.

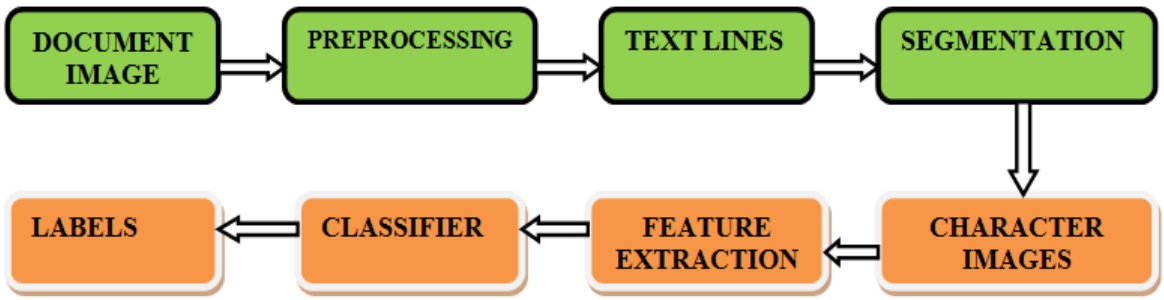

Figure 1: FLOW DIAGRAM OF OCR

PRE-PROCESSING

A crucial part of an OCR system is preprocessing. Preprocessing [6] in this method comprises of binarization and size which are described in the following.

Binarization

Grayscale or color images are frequently desirable to represent as binary images in order to decrease storage needs and to boost processing performance. The name of this procedure is digitization.

Size Normalization

The character is adjusted to take into account size variations. Normalization is the process of getting a character to fit into an array of a certain size.. The size of this array is determined through a process of trial and error to determine the value that gives the best results. Processing and mapping of characters of any size and shape.

SEGMENTATION

Character position in the image is determined during segmentation and image size is normalized to character template dimensions . Both internal and external segments are possible. The separation of different writing units, such as sentences, paragraphs, or words, is known as external segmentation. An image of a sequence of characters is segmented internally into smaller images of each individual character.

FEATURE EXTRACTION METHODS

Most recognition systems need a more compact and characteristic representation in order to reduce extra complexity and improve algorithm accuracy.
For simplicity, an extraction is made of a set of features for each class that are invariant to differences in characteristics, yet help distinguish it from other classes.

MATCHING OF TEMPLATES

Matching refers to a group of similarity-based techniques where the separation between the feature vectors that describe the extracted characters is measured and a calculation is made for each class's description. Although other measurement standards may be employed, the Euclidean distance is the most typical. When the classes are well separated—that is, This minimum distance classifier performs [4] better if the distance between means is significant for each class spread. The correlation approach is used when whole characters are used as classification input and no features are extracted (template matching)[6] . Here, the separation between the character image and each character class' database image is calculated.

Zoning

There are various overlapping or non-overlapping zones within the character's frame. Analysis[5] of the concentration of the points or certain traits in various regions represents the form. For instance, contour direction features calculate the histograms of chain codes within the rectangular and diagonal zones that make up the image array to measure the direction of the character's contour.

Moments

The term "moment" in this context refers to a few of the properties that can be computed from the photos. Moments of various orders are employed in patterns due to their inclusion in statistics. In this instance, a character's moments at various points are used as a feature.

These are the character recognition techniques that are most frequently utilized. Moments, like central moments and Zernike moments, create a condensed version of the original document image that makes object recognition size, translation, and rotation independent Moments are regarded as series expansion representations since they allow for a complete reconstruction of the original image.

Chain coding

Freeman chain coding is recommended in as a method for identifying loops and curves in a character. A boundary is represented by a connected series of straight lines in chain codes specified length and direction line segments. Each segment's direction is coded using a numbering system.
By clockwise tracing an object's boundary and giving each segment connecting two pixels a direction, it is possible to create a chain code. Chain of Freeman Essentially, coding is created by mapping a character's strokes into a 2-D parameter space made up of codes.

IV. METHODS FOR RECOGNITION AND CLASSIFICATION

Neural Networks (NN's)

A massively parallel interconnected computing architecture is what is known as a NN [7]of flexible "neural" processing units. It can process computations more quickly than traditional methods due to its parallel nature. Due to its adaptability, it can learn the features of the input signal and adjust to changes in the data. Several nodes make up a NN. One node in the network feeds its output to another, and the network's overall outcome is determined by the intricate relationships between all of the nodes. It can be demonstrated that most NN architectures are similar to statistical pattern recognition techniques despite the disparate underlying concepts.

Statistical classifiers [8] that are ideal. Support Vector Machines, Principal Component Analysis (PCA), and Kernel Principal Component Analysis are the most crucial Kernel approaches. e.g. (KPCA). A group of supervised learning techniques[9] called support vector machines (SVM) can be used for categorization. Data is typically split into training and testing sets when performing a classification task. The goal of SVM is to create a model that foretells the test data's target values. There are several varieties of SVM kernel functions, including linear, polynomial, Gaussian radial basis function (RBF), and sigmoid.

Character recognition has no standardized test sets, and an OCR system's performance is heavily dependent. Since the input quality is low, comparing and evaluating various systems is made challenging. However, recognition rates are frequently given and are frequently displayed as the proportion of characters that were correctly identified. This, however, says nothing about the mistakes made. Consequently, when evaluating an OCR system.

Recognition rate
Percentage of recognized characters that were properly categorized.

Rejection rate
Number of characters that the system failed to recognize. Her OCR system flags rejected characters, making it simple to find them and manually repair them.

Error rate
Percentage of incorrectly classified characters. Misclassified characters [2][3] are not recognized by the system and you must manually review the recognized text to detect and correct these errors.

V. EVALUATION METRICS

CHARACTER ERROR RATE EQUATION	NORMALIZED CER EQUATION	*W*ORD ERROR RATE EQUATION
CER = S+D+I / N	***CER NORMALIZED = S+D+I / S+D+I+C***	**WER = S*W*+D*W*+ I*W* / N*W***
where: • **S** = Number of **S**ubstitutions • **D** = Number of **D**eletions • **I** = Number of **I**nsertions • **N** = **N**umber of characters in reference text.	*Where* *C*= **Number of Characters**	**Where** WER represents the number of **word** substitutions, deletions, or insertions needed to transform one **sentence** into another.

VI. CONCLUSION

This article gave a summary of feature extraction and classification methods used in optical systems for character recognition. This topic has been the subject of much inquiry. So far, everything is going smoothly. Method accuracy for feature extraction and categorization. However, the many feature extraction and classification techniques discussed here are quite useful and beneficial for beginning researchers. A template matching method with a simple and beautiful algorithm that is straightforward to implement Flexibly responds to changes in recognition target

classes. His perception is the strongest of the monotypesIt's a unified single-column page, takes less time, doesn't require sample training, but The template can only recognize characters of the similar size and rotation. For scanned documents and damaged text, neural networks' capacity to recognize characters through abstraction is excellent.

REFERENCES

[1] Q.-D. Nguyen, D.-A. Le, N.-M. Phan, N.-T. Phan and P. Kromer, "Statistical post-processing approaches for OCR texts", *Proc. Int. Joint Conf. Adv. Comput. Intell.*, pp. 457-467, 2022.

[2] Q.-D. Nguyen, D.-A. Le, N.-M. Phan and I. Zelinka, "OCR error correction using correction patterns and self-organizing migrating algorithm", *Pattern Anal. Appl.*, vol. 24, pp. 701-721, Nov. 2021.

[3] G. T. Bazzo, G. A. Lorentz, D. Suarez Vargas and V. P. Moreira, "Assessing the impact of OCR errors in information retrieval", *Proc. 42nd Eur. Conf. IR Res. (ECIR)*, pp. 102-109, Apr. 2020.

[4] A.D. Le, H. T. Nguyen and M. Nakagawa, "An end-to-end recognition system for unconstrained Vietnamese handwriting", *Social Netw. Comput. Sci.*, vol. 1, no. 1, pp. 1-8, Jan. 2020.

[5] A. Hamdi, A. Jean-Caurant, N. Sidere, M. Coustaty and A. Doucet, "An analysis of the performance of named entity recognition over OCRed documents", *Proc. ACM/IEEE Joint Conf. Digit. Libraries (JCDL)*, pp. 333-334, Jun. 2019.

[6] M. Mieskes and S. Schmunk, "OCR quality and NLP preprocessing", *Proc. Workshop Widening NLP*, pp. 102-105, 2019

[7] H. T. Nguyen, C. T. Nguyen, P. T. Bao and M. Nakagawa, "A database of unconstrained Vietnamese online handwriting and recognition experiments by recurrent neural networks", *Pattern Recognit.*, vol. 78, pp. 291-306, Jun. 2018.

[8] Mei, A. Islam, A. Moh'd, Y. Wu and E. Milios, "Statistical learning for OCR error correction", *Inf. Process. Manage.*, vol. 54, no. 6, pp. 874-887, Nov. 2018.

[9] C. Amrhein and S. Clematide, "Supervised OCR error detection and correction using statistical and neural machine translation methods", *J. Lang. Technol. Comput. Linguistics*, vol. 33, no. 1, pp. 49-76, Jul. 2018.

[10] G. Khirbat, "OCR post-processing text correction using simulated annealing (OPTeCA)", *Proc. Australas. Lang. Technol. Assoc. Workshop*, pp. 119-123, 2017.

[11] Arica, & Yarman-Vural, (2001). An overview of character recognition focused on off- line handwriting. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 31(2), 216-233.

[12] Y. Bassil and M. Alwani, "OCR post-processing error correction algorithm using Google's online spelling suggestion", *J. Emerg. Trends Comput. Inf. Sci.*, vol. 3, no. 1, pp. 90-99, 2012.