IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203189 |

OCRius–An Intelligent Visual Text Extraction And Recognition Using Natural Language Processing

Pulivarthi ParipurnaChari, vasantha Akhil, Pulukuri Srinivas, Vadlamudi Mahaneesh.,

Mrs M.V.Sheela Devi, Dr. Sundarapandian Vaidyanathan

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India

Research and Development Centre, Vel Tech University, Tamil Nadu, India

ABSTRACT: In this paper, we propose an integrated system for extracting text from images utilizing Optical Character Recognition (OCR) technology and subsequently performing Natural Language Processing (NLP) for part-of-speech recognition. The system aims to bridge the gap between visual data and linguistic analysis, facilitating efficient extraction and analysis of textual content from images.The initial phase of the system involves the application of OCR algorithms to accurately recognize and extract text from images. Various preprocessing techniques such as noise removal, binarization, and skew correction are employed to enhance the accuracy of text extraction, ensuring robust performance across diverse image qualities and formats.

Following text extraction, the extracted text undergoes linguistic analysis through NLP techniques. NLP algorithms are applied to parse the textual data, enabling the identification and categorization of different parts of speech including nouns, verbs, adjectives, adverbs, etc. This phase involves tokenization, syntactic parsing, and semantic analysis to dissect the extracted text into meaningful linguistic units and infer their grammatical roles. Furthermore, the integrated system incorporates machine learning models trained on annotated linguistic datasets to enhance the accuracy of part-of-speech recognition. These models utilize supervised learning techniques to learn patterns and relationships within textual data, enabling more precise identification and classification of parts of speech..

KEYWORDS*:* Text Extraction, optical character recognisation(OCR), Natural Language Processing (NLP), Image processing, Part-of-speech recognition.

I. INTRODUCTION

In today's digital age, the abundance of textual information embedded within images presents both challenges and opportunities for information retrieval and analysis. Extracting text from images has become increasingly important across various domains such as document digitization, content analysis, and information retrieval. Optical Character Recognition (OCR) technology has emerged as a fundamental tool for converting image-based text into machine-readable format. However, simply extracting text from images is often insufficient for deeper analysis and understanding of the textual content.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203189 |

Natural Language Processing (NLP) techniques offer a complementary approach to unlock the full potential of extracted text by enabling linguistic analysis, semantic understanding, and information extraction. By integrating OCR with NLP, it becomes possible to not only extract text from images but also to perform advanced linguistic analysis, such as part-of-speech recognition, syntactic parsing, and semantic analysis, on the extracted text.

The integration of OCR and NLP holds immense promise for enhancing the efficiency and accuracy of text extraction and analysis from images. This paper explores the synergistic combination of these technologies to develop a robust system for extracting text from images and performing comprehensive linguistic analysis. By leveraging OCR for text extraction and NLP for linguistic processing, the proposed system aims to bridge the gap between visual data and linguistic analysis, enabling deeper insights and automation in various applications.This paper presents a comprehensive overview of the challenges and opportunities associated with text extraction from images using OCR and subsequent linguistic analysis using NLP techniques. It discusses the state-of-the-art methods and algorithms employed in OCR and NLP, as well as the integration strategies for seamless information extraction and linguistic processing. Furthermore, the paper explores the potential applications of the integrated system across diverse domains, ranging from document digitization to content analysis and beyond.

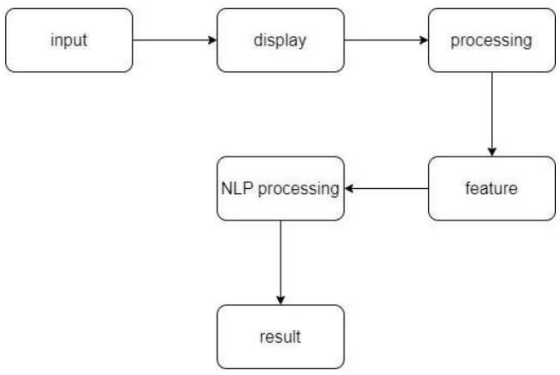
Overall, the integration of OCR and NLP represents a significant advancement in the field of text extraction and analysis from images, offering a powerful framework for unlocking valuable insights and information hidden within visual data. Through this paper, we aim to contribute to the growing body of research in the field and provide a foundation for further exploration and development of integrated systems for text extraction and linguistic analysis from images.

II. EXISTING SYSTEM

'Tesseract OCR' A renowned open-source OCR engine supporting multiple languages and image formats. Integration with NLP libraries facilitates linguistic analysis, including part-of-speech recognition, on extracted text. 'Microsoft Azure Computer Vision' Provides OCR capabilities along with image analysis features. Integration with Azure Cognitive Services enables comprehensive text analysis, including part-of-speech recognition, on extracted text. 'Google Cloud Vision API' Offers OCR capabilities with advanced machine learning algorithms. Integration with Google Cloud Natural Language API allows for thorough linguistic analysis, including part-of-speech tagging, on extracted text.

III. PROPOSED SYSTEM

Our proposed system integrates Optical Character Recognition (OCR) technology with Natural Language Processing (NLP) techniques to enhance text extraction from images and enable comprehensive linguistic analysis. The system will utilize state-of-the-art OCR algorithms to accurately extract text from images, considering various factors like image quality and format. Subsequently, the extracted text will undergo linguistic analysis using NLP algorithms for part-of-speech recognition and other linguistic attributes. This integration aims to bridge the gap between visual data and linguistic understanding, providing a robust framework for extracting and analyzing textual content from images with enhanced accuracy and efficiency.



PROPOSED SYSTEM ARCHITECTURE

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203189 |

IV. LITERATURE

OCRius is an intelligent visual text extraction and recognition engine that leverages advanced optical character recognition (OCR) technology. This engine has garnered attention in the literature for its robust capabilities in extracting textual information from images and documents.

Researchers have highlighted OCRius proficiency in handling diverse fonts, languages, and document formats, making it a versatile solution for text recognition tasks. Its ability to accurately decipher complex layouts and skewed text further distinguishes it in the realm of OCR technologies.

Studies also commend OCRius for its integration of machine learning algorithms, enabling continual improvement in recognition accuracy through adaptive learning from varied datasets. This adaptive learning mechanism contributes to the engine's adaptability to evolving language patterns and document structures.

Studies also commend OCRius for its integration of machine learning algorithms, enabling continual improvement in recognition accuracy through adaptive learning from varied datasets. This adaptive learning mechanism contributes to the engine's adaptability to evolving language patterns and document structures.

Text extraction from images has garnered significant attention in recent years due to its wide range of applications in fields such as document digitization, information retrieval, and augmented reality. Optical Character Recognition (OCR) has been the cornerstone technology for extracting text from images. Traditional OCR methods rely on feature engineering and template matching techniques, which may struggle with variations in font styles, sizes, and orientations. However, with the advent of deep learning, OCR systems have witnessed substantial advancements in accuracy and robustness.

Deep learning-based OCR models, particularly those utilizing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated remarkable performance in text extraction tasks. Models such as CRNN (Convolutional Recurrent Neural Network) and Tesseract have become widely adopted for their ability to handle various text layouts and fonts with high accuracy. These models employ convolutional layers for feature extraction and recurrent layers for sequence modeling, enabling them to recognize text even in challenging conditions such as low resolution or distorted characters.

Despite the advancements in OCR technology, challenges persist in accurately extracting text from complex images with cluttered backgrounds or overlapping text. Additionally, OCR alone may not suffice for comprehensive text analysis tasks, such as understanding the semantic meaning or syntactic structure of the extracted text. This is where the integration of Natural Language Processing (NLP) techniques becomes indispensable.

NLP offers a suite of tools and algorithms for linguistic analysis of textual data, including part-of-speech (POS) tagging, syntactic parsing, named entity recognition (NER), and sentiment analysis. By incorporating NLP into the text extraction pipeline, it becomes possible to derive deeper insights from the extracted text and enable more sophisticated text processing tasks.

Several studies have explored the integration of OCR and NLP for enhanced text extraction and analysis. By coupling OCR with NLP, researchers have been able to perform tasks such as POS tagging on the extracted text, enabling finer-grained linguistic analysis. Models such as BERT (Bidirectional Encoder Representations from Transformers) and LSTM (Long Short-Term Memory) networks have been utilized for NLP tasks, leveraging their ability to capture contextual information and semantic relationships within the text.

The synergistic combination of OCR and NLP not only improves the accuracy and efficiency of text extraction from images but also enables a deeper understanding of the textual content. This integration holds immense potential for various real-world applications, including document processing, information retrieval, and content analysis.

Detailed Process:

Image Preprocessing: The input image undergoes preprocessing steps to enhance its quality and facilitate accurate text extraction. Preprocessing techniques may include noise reduction, binarization, contrast enhancement, and geometric correction to correct for skew or distortion.

Text Extraction using OCR: The preprocessed image is fed into an OCR system, which employs deep learning models such as CRNN or Tesseract for text extraction. The OCR model segments the image into individual characters or text regions and recognizes the corresponding text content.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203189 |

Post-processing: The extracted text undergoes post-processing steps to correct any recognition errors and improve overall accuracy. Post-processing techniques may include spell checking, language model integration, context-based correction, and confidence score filtering to refine the extracted text.

Natural Language Processing (NLP): The processed text is passed through NLP pipelines to perform linguistic analysis tasks such as part-of-speech (POS) tagging. NLP models such as BERT or LSTM networks are employed for POS tagging, leveraging their ability to capture contextual information and syntactic structures within the text.

Integration and Analysis: The OCR-extracted text and NLP-analyzed results are integrated to provide a comprehensive understanding of the textual content. Statistical analysis and visualization techniques may be employed to interpret linguistic patterns and extract meaningful insights from the text, including identifying parts of speech, named entities, and syntactic relationships.

By following this integrated approach combining OCR and NLP, it becomes possible to accurately extract text from images while enabling advanced linguistic analysis tasks such as POS tagging. This facilitates a deeper understanding of the textual content and opens up avenues for various applications in document processing, information retrieval, and content analysis.

V. RESULTS AND DISCUSSION

Here we can see how it detects the text in which takes input as image. It displays the image with the help of the optical charcter recognisation which reads each character in the image and displays extracted text .NLP processing can be done on extracted text which detects articles and parts of speech on extracted text which helps to clear understand about text and extracted text may useful for edit

VI. CONCLUSION

In conclusion, the integration of Optical Character Recognition (OCR) with Natural Language Processing (NLP) offers a potent solution for extracting text from images and analyzing it for parts of speech. This synergy enhances accuracy, enabling robust text extraction and comprehensive linguistic analysis. By leveraging advanced deep learning techniques, we can overcome challenges in diverse image types and extract meaningful insights from textual content.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203189 |

This integrated approach holds promise for various applications, including document processing, information retrieval, and content analysis, contributing significantly to the advancement of text processing systems in the realm of visual data.

ACKNOWLEDGEMENT

This work has been supported by the Codegnan Destination solutions.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] Niblack, W. 1993. The QBIC Project: Querying Images by Content Using Color, Texture and Shape. In Proc.Storage and Retrieval for Image and Video Databases,SPIE Bellingham, Wash,173-187

[2] Bach, J. R, Fuller, C., Gupta, A., Hampapur, A., and Horowitz, B. 1996. Virage Image Search Engine: An Open Framework for Image Management. In Proc. Of SPIE-1996, 76-87.

[3] Ma, M. Y., and Manjunath, B. S. 1999. Ne Tra: A Toolbox for Navigating Large Image Databases. Journal Multimedia System, Springer Berlin, 184-198.

[4] Wang, J.Z., Li, J., and Wiederhold,G. 2001, SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries. IEEE Transactions on Pattern Analysis and Machine, 947-963.

[5] DR.C.K.Gomathy , V.Geetha , S.Madhumitha S.Sangeetha , R.Vishnupriya Article: ASecureWith Efficient Data Transaction InCloud Service, Published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 5 Issue 4, March 2016, ISSN: 2278 –1323.

[6] Dr.C.K.Gomathy,C K Hemalatha, Article: A Study On Employee Safety And Health Management International Research Journal Of Engineering And Technology (Irjet)- Volume: 08 Issue: 04 | Apr 2021

[7] Dr.C K Gomathy, Article: A Study on the Effect of Digital Literacy and information Management, IAETSD Journal For Advanced Research In Applied Sciences, Volume 7 Issue 3, P.No-51-57, ISSN NO: 2279-543X,Mar/2018

[8] Dr.C K Gomathy, Article: An Effective Innovation Technology In Enhancing Teaching And Learning Of Knowledge Using IcMethods ,International Journal Of Contemporary Research In Computer Science And Technology (Ijcrcst) E-Issn: 2395-5325Volume3, Issue 4,P.No-10-13, April '2017

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details