**REVIEWS ON TOPIC OPTICAL CHARACTER RECOGNITION USING TESSERACT OCR ENGINE**

NAME: ABHISHEK CHIGADOLLI

USN: 2GI20EC004

1. TITLE: “Optical Character Recognition Using Tesseract Engine”

REVIEW :

After reading the document "Optical Character Recognition Using Tesseract Engine," it is evident that the Tesseract OCR Engine is a trustworthy and efficient optical character recognition tool. The document describes the Tesseract OCR Engine's architecture, preprocessing stages, and recognition process in depth. It also provides a comparison of the Tesseract OCR Engine to other OCR technologies. The experimental findings reported in the document demonstrate the accuracy and efficiency of the Tesseract OCR Engine in recognizing text from diverse image formats. Overall, the document gives useful information about the Tesseract OCR Engine's capabilities and performance for OCR projects.

Link

<https://www.ijert.org/research/optical-character-recognition-using-tesseract-engine-IJERTV10IS090157.pdf>.

1. TITLE: "Reading Digits in Natural Images with Unsupervised Feature Learning"

REVIEW:

"Reading Digits in Natural Images with Unsupervised Feature Learning" is a research study that provides a novel approach to digit recognition in natural photos using unsupervised feature learning. The research delves into the suggested method, which entails training a deep autoencoder to learn features from a huge unlabeled dataset of natural photos and then using the learnt features to recognize digits in fresh images. The experimental results reported in the research show that the proposed strategy is effective in delivering state-of-the-art performance on numerous benchmark datasets. Overall, the research gives useful insights into the application of unsupervised feature learning for digit recognition in natural photos and indicates its potential for enhancing digit recognition system accuracy and efficiency.

Link:

<https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33418.pdf>

1. TITLE:"Combined Orientation and Script Detection Using the Tesseract OCR Engine"

REVIEW :

The article "Combined Orientation and Script Detection Using the Tesseract OCR Engine" examines the capabilities of the Tesseract OCR engine for detecting text orientation and script identification at the same time. The article describes the difficulties encountered in text detection, particularly in multilingual publications, and how the Tesseract engine can reliably recognize the orientation and script of text. The paper also describes how Convolutional Neural Networks (CNN) are used to train the Tesseract engine to recognize text orientation and script. The research concludes that the combined orientation and script detection of the Tesseract OCR engine can assist enhance the accuracy and speed of text recognition in a variety of applications, including document scanning and translation.

Link:

<https://tesseract-ocr.github.io/docs>

1. TITLE: "Image Processing and OCR Analysis for the Detection of Textual Information from Digital Images"

Review

The article "Image Processing and OCR Analysis for the Detection of Textual Information from Digital Images" explains how to detect and recognize textual information from digital images. Image preprocessing, text identification, text segmentation, and optical character recognition (OCR) are all stages of the proposed method. The authors also offer experimental data that show how good their system is at detecting and recognizing text from various sorts of photographs. The proposed method could be useful in a range of applications, including document digitization and automatic image annotation. However, the article would benefit from more extensive explanations and discussions of the suggested method's many stages.

Link: <https://www.matec-conferences.org/articles/matecconf/pdf/2020/05/matecconf_cscns2020_03034.pdf>

1. TITLE: "Unknown-box Approximation to Improve Optical Character Recognition Performance"

Review:

The paper "Unknown-box Approximation to Improve Optical Character Recognition Performance" presents a "unknown-box" approximation to improve the accuracy of optical character recognition (OCR) systems. This method entails breaking the input image into smaller boxes and treating some of these boxes as "unknown" throughout the OCR system's training phase. This forces the system to learn to recognize characters that are partially obscured or otherwise difficult to identify. The results suggest that the proposed strategy can increase OCR accuracy, especially in difficult photos with poor contrast or substantial noise.

Link: <https://arxiv.org/pdf/2105.07983.pdf>

1. TITLE: "Content-based Image Retrieval Using Tesseract OCR Engine"

Review

The paper "Content-based Image Retrieval Using Tesseract OCR Engine" provides a novel method for retrieving images from a huge image library based on their textual content by utilizing the Tesseract OCR engine. The suggested method extracts text from photographs and uses it to search the database for comparable images. The authors test the proposed method on a 1000-image dataset and report encouraging results. According to the study's findings, the proposed method can be used to effectively retrieve images based on their textual content, which can be useful in a variety of applications such as digital libraries, online archives, and image-based search engines.

Link: [https://thesai.org/Downloads/Volume12No7/Paper\_76- Content\_based\_Image\_Retrieval\_using\_Tesseract\_OCR\_Engine.pdf](https://thesai.org/Downloads/Volume12No7/Paper_76-%20%20%20%20%20%20Content_based_Image_Retrieval_using_Tesseract_OCR_Engine.pdf)

1. TITLE: "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning"

REVIEW:

This is a research paper titled "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning" by Adam Coates et al. published on arXiv in 2010. The research describes an unsupervised feature learning method for recognising text in natural scene photos. The system is divided into two parts: text detection and character recognition. A sliding window technique is utilised in the text detection stage to scan the image and recognise sections of text using a convolutional neural network (CNN) trained with unsupervised feature learning. A separate CNN is trained to recognise characters within the identified text regions during the character recognition stage. On various benchmark datasets, the system achieves state-of-the-art performance, demonstrating the efficiency of unsupervised feature learning for text identification and recognition in natural scene photos.

Link

<https://arxiv.org/ftp/arxiv/papers/1003/1003.5891.pdf>

1. TITLE: "Printed Gujarati Character Recognition Using CNN and LSTM"

Review:

"Printed Gujarati Character Recognition Using CNN and LSTM" For recognising written Gujarati characters, the authors suggested a deep learning-based technique that blends convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. On the dataset utilised in the study, the proposed method attained an accuracy of 93.43%, exceeding some of the existing approaches for recognising Gujarati characters. According to the study, the proposed method can be used for a variety of applications such as OCR, licence plate identification, and handwriting recognition.

Link <https://www.irjet.net/archives/V9/i4/IRJET-V9I489.pdf>

1. TITLE: “Adapting the Tesseract Open Source OCR Engine for Multilingual OCR”

Review

The goal of this study is to adapt the Tesseract open-source OCR engine for multilingual OCR. The paper highlights the problems of developing a multilingual OCR system and the changes made to the Tesseract engine to allow multilingual OCR. The authors also explain the system's evaluation on several datasets, demonstrating that the customized Tesseract engine performs well in recognizing text in multiple languages. The report finishes with the multilingual OCR system's possible applications in numerous fields.

Link:<https://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/35248.pdf>

1. TITLE: "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study"

Review

The paper "Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study" gives a case study of the Tesseract OCR engine's performance evaluation on scanned images of printed documents. The study examines the accuracy of the Tesseract OCR engine for various font types and sizes, as well as its comparison to other OCR engines. According to the study's findings, the Tesseract OCR engine performs well in terms of accuracy and speed, and it may be utilised as a dependable OCR tool for a variety of applications. The study also emphasises the necessity of pre-processing processes for boosting OCR accuracy, such as picture binarization and deskewing.

Link:<https://www.researchgate.net/profile/Chirag-Patel-12/publication/235956427_Optical_Character_Recognition_by_Open_source_OCR_Tool_Tesseract_A_Case_Study/links/00463516fa43a64739000000/Optical-Character-Recognition-by-Open-source-OCR-Tool-Tesseract-A-Case-Study.pdf>

**References:**

1. Nikita Kotwal , Gauri Unnithan , Ashlesh Sheth , Nehal Kadaganchi, 2021, Optical Character Recognition using Tesseract Engine, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 09 (September 2021)
2. Smith, Ray. "An overview of the Tesseract OCR engine." *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE, 2007.
3. Unnikrishnan, Ranjith, and Ray Smith. "Combined script and page orientation estimation using the tesseract ocr engine." *Proceedings of the international workshop on multilingual OCR*. 2009.
4. Sun, Yueyue, and Xuechen Zhao. "Research and implementation of license plate recognition based on android platform." *MATEC Web of Conferences*. Vol. 309. EDP Sciences, 2020.
5. Randika, Ayantha, et al. "Unknown-box approximation to improve optical character recognition performance." *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*. Springer International Publishing, 2021.
6. Adjetey, Charles, and Kofi Sarpong Adu-Manu. "Content-based image retrieval using Tesseract OCR engine and levenshtein algorithm." *International Journal of Advanced Computer Science and Applications* 12.7 (2021).
7. Rakshit, Sandip, and Subhadip Basu. "Recognition of handwritten Roman script using Tesseract open source OCR engine." *arXiv preprint arXiv:1003.5891* (2010).
8. Desai, Nidhi, Mosin Hasan, and Prashant Swadas. "Printed Gujarati Character Recognition: A Review." (2022).
9. Smith, Ray, Daria Antonova, and Dar-Shyang Lee. "Adapting the Tesseract open source OCR engine for multilingual OCR." *Proceedings of the International Workshop on Multilingual OCR*. 2009.
10. Patel, Chirag, Atul Patel, and Dharmendra Patel. "Optical character recognition by open source OCR tool tesseract: A case study." *International Journal of Computer Applications* 55.10 (2012): 50-56.