I. J. Engineering and Manufacturing, **2023, 6, 1-8**
Published Online on December 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2023.06.01

An Integrated Pipeline with Internal Image Processing for Efficient Image to Text to Speech Conversion

Shreyas Reddy*
International Institute of Information Technology, Bhubaneswar, Odisha, India
Email: b419056@iiit-bh.ac.in
ORCID iD: https://orcid.org/0000-0002-8166-2385
*Corresponding Author

Rashmi Ranjan Das
International Institute of Information Technology, Bhubaneswar, Odisha, India
Email: b419045@iiit-bh.ac.in
ORCID iD: https://orcid.org/0009-0005-3083-8415

Anjali Mohapatra
International Institute of Information Technology, Bhubaneswar, Odisha, India
Email: anjali@iiit-bh.ac.in
ORCID iD: https://orcid.org/0009-0005-7184-9217

Received: 27 May, 2023; Revised: 01 August, 2023; Accepted: 23 August, 2023; Published: 08 December, 2023

Abstract: Optical Character Recognition Systems (OCR) is a tool that helps computers read text from pictures of papers. It makes it easier for machines to understand what the words say without needing a person to read it out loud. It allows for easy digitizing of historical documents, archival material, and medical records thereby saving on their retrieval times. However, the accuracy of OCR systems heavily relies on the quality of the input images. To negate the contribution of the quality of input images to the accuracy of OCR systems, in this paper, we propose an integrated image pre-processing pipeline integrated with the OCR systems that enhances the quality of input images for efficient image to text conversion. This method results in an easily understandable text output with a lower Character Error Rate (CER) in comparison to the current methods. In addition, we explore a technique for converting text from a document or image into machine-readable form and then converting it to audio output using gTTS, a Python library that interfaces with Google Translate's text-to-speech API. We assess the effectiveness of this approach and illustrate that it substantially enhances OCR precision when compared to other existing methods. This paper presents a clear overview of the growth phases and significant obstacles, accompanied by compelling comparisons of results achieved through various methods.

Index Terms: Optical Character Recognition(OCR), Text-to-speech(TTS), Image processing, Character Error Rate(CER).

1. Introduction

There exists a population of approximately 285 million individuals with visual impairments among which approximately 30% of them are blind. One potential strategy to improve reading abilities could be through the use of character recognition techniques. Recent developments in the domain of digital image processing and the increase in computational power in recent times have made it possible to implement OCR systems effectively.

Optical character recognition involves the use of specialized software that can recognize individual characters from printed or handwritten text and convert them into a machine-readable format. They have revolutionized the way we handle, store, and analyze text data. With this technology, it is possible to digitize and preserve valuable records, as well as improve accessibility for people with visual impairments. The integration of this technology has been extensively embraced in diverse sectors, particularly in the banking industry and in fields such as healthcare, publishing, and education where an efficient and precise recognition of text is of utmost importance. Text detection, which entails

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

finding the location of text in an image, is one of two crucial steps in OCR. The extraction of text from the image, or text recognition, comes next. Only a small number of the OCR engines used in current research studies are free and open source for usage. In accordance with the amount of noise in the document images, their accuracy ranges from 70% to 98%.

While text-to-speech (TTS) and optical character recognition (OCR) technologies exist to help address the challenges faced by the visually impaired, they are often not integrated into a single, user-friendly pipeline. This lack of integration can make the process of converting textual images to speech output cumbersome and time-consuming. Therefore, there is a need to design an integrated package that combines OCR and TTS technologies to allow for seamless and efficient conversion of textual images to speech output. Such a package would provide individuals with visual impairments or reading difficulties with greater accessibility to written information, thereby improving their independence, quality of life, and participation in society. The motivation behind this problem statement is to develop a pipeline that enables individuals to access and comprehend textual information easily, regardless of their visual capabilities or reading abilities.

Literature studies on OCR suggest that OCR engine accuracy heavily depends on the quality of input images. Poor image quality can result in misinterpretations, omissions, and errors, leading to inaccurate and unreliable OCR outputs. Therefore, enhancing the quality of input images through pre-processing techniques has become a crucial step in the OCR pipeline. In this paper, we thus propose an integrated image pre-processing pipeline for the efficient image to text conversion by OCR systems. The pipeline consists of several stages, including image binarization, noise reduction, and morphological transformations.

2. Literature Review

A literature survey of research studies and articles on OCR and text-to-speech technologies (TTS) reveals several key findings. Firstly, a study [1] discusses an algorithm for accurate text recognition subject to various environmental conditions. The study concluded that accuracy gains could be secured by correcting for the orientation of the captured image and the levels of illumination in the image. This finding helps conclude that the quality of the input image largely determines the performance of the OCR system. Therefore, it is important to have an integrated image preprocessing pipeline to improve the quality of the image before extracting the text from the image.

Secondly, another study published [2] evaluated the effectiveness of Image-to-text-to-speech (ITTTS) software for reading printed materials and found that it significantly improved reading speed and comprehension among individuals with visual impairments. Additionally, the study found that users preferred the ITTTS software over traditional screen readers, as it provided a more natural reading experience. Therefore, in this paper, we propose to integrate the TTS engine alongside the EasyOCR engine to provide a seamless ITTTS conversion.

Thirdly, another study [3] examined the usability and accessibility of ITTTS software for individuals with different types of visual impairments. The study found that the software was easy to use and highly accessible, even for users with severe visual impairments. This shows the immense impact ITTTS systems have on society and highlights the importance of improving the quality of such systems which can be done through our proposed pipeline.

Some other research works deal with assisting visually impaired people by leveraging the latest advances in OCR technology [4,6]. These works highlight and discuss the various methods like binarization and morphological transformations and their effects on the input image quality to enhance the performance of OCR. To address the problem of bad quality of the input image, in our paper, we propose an integrated image preprocessing pipeline that receives the input image from the user and subjects it to processing to improve its quality. Overall, the literature survey suggests that ITTTS software has great potential for improving the accessibility of digital content for individuals with visual impairments or reading difficulties. However, further research is needed to address the current limitations of the technology and to continue advancing its capabilities.

3. Proposed Methodology

To address the research gap defined in the introduction section, we propose a methodology mainly involving designing an internal image preprocessing, choosing an appropriate OCR engine to convert textual images to text, and then using a TTS engine to convert extracted text to speech. It is of prime importance to ensure that the input image to the ITTTS system is of high quality to ensure higher accuracy of the image to speech conversion. While the EasyOCR engine has a built-in image pre-processing stage, not all of the components in the pipeline help to enhance the quality of our input image [7] . This problem of input image quality could be tackled and effective output could be ensured by providing an internal pre-processing pipeline as part of the ITTTS system. The proposed pipeline for image to speech conversion is shown in Fig.1. The pre-processing pipeline mainly involves:

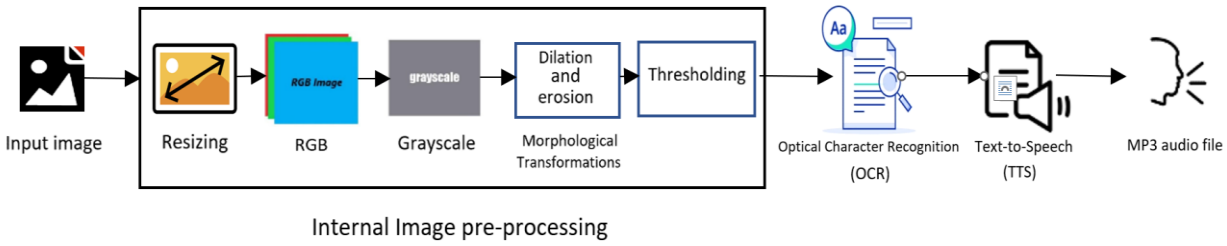

Fig.1. An integrated pipeline proposed for image to speech conversion (ITTTS)

A. Loading the image(input)

The user may load any text-containing image from the computer. A good image quality must be guaranteed in order for the output to be fairly close to the ground truth label. The input image must be checked to make sure it is not fuzzy, noisy, or distorted [8]. Any of these characteristics in the input source will almost certainly result in errors in the output. So, for the OCR engine to be highly accurate, image quality must be very high.

B. Rescaling

Resizing the image helps to reduce noise and blur, which can make it easier for OCR algorithms to recognize individual characters. OCR algorithms may struggle with images that are too large or too small, as they may not be able to properly recognize individual characters. OCR accuracy can be improved if the image is resized to an optimal size.

C. Converting RGB image to Grayscale

In this method, a multicolor image is mapped to a black-and-white image. By increasing the contrast between the text and the background, this can make it simpler for OCR algorithms to distinguish between individual characters. EasyOCR performs this conversion internally. However, it results in poorer output if the input image is unevenly dark.

D. Morphological transformations

We apply morphological transformations to the image to remove noise. Noise can be random changes in the density of the image, which can be seen as graininess in films and the discrepancy between elements in digital images. Dilation operation is applied to increase the size of white areas in the image. Then an erosion operation is applied to erode the boundaries of the foreground object.

E. Thresholding

This technique turns all pixel values into black below a predetermined threshold and white above it. Thresholding can help to remove noise and other artifacts from the image, which can improve OCR accuracy. More importantly, it also reduces the number of shades of gray in an image to just two (black and white), which can simplify the image and make it easier for OCR algorithms to process.

4. Optical Character Recognition Engine

After the input image is ensured to be of high quality, we use an efficient EasyOCR engine for the conversion of textual image to human readable text. The OCR part of the pipeline is a primary component in designing the ITTTS pipeline that we propose in this paper and it acts as an intermediary between the internal image preprocessing pipeline and the TTS engine.

EasyOCR is a popular open-source Python package used for optical character recognition (OCR) tasks, offering an intuitive interface and support for over 70 languages. Its popularity and effectiveness are reflected in the numerous scholarly publications that have cited its use in recent years. For example, in a study by Zheng et al. (2021) [9], EasyOCR was used to extract text from images of documents in order to perform text analysis for historical research. Similarly, in a study by Gao et al. (2020) [10], EasyOCR was utilized to extract text from various sources, including newspapers, journals, and government documents, to create a dataset for training and evaluating natural language processing (NLP) models. We preferred EasyOCR over other tools like Tesseract because EasyOCR provides us with pre-trained models for various languages. They also perform well on noisy or low-quality images. It is also designed to be fast and can process multiple images in parallel making it suitable for use.

5. Text-to-Speech Converter (TTS)

The TTS engine part of the pipeline is the final component in designing the ITTTS pipeline that we propose in this paper and it converts the text extracted by the OCR engine into a human speech audio file.

Text-to-speech (TTS) technology has been in development for decades, and its applications are increasingly important in the digital age. According to a study by Zheng et al. (2021) [9], TTS can play a crucial role in improving accessibility for individuals with visual impairments, as well as for those who prefer audio formats for reading. In our integrated pipeline, after the text is extracted by the OCR engine from the input image, the textual output is then passed to the gTTS engine for conversion from text to speech. We chose the gTTS engine because gTTS allows for customization of voice, pitch, speaking rate, and volume to create a more personalized listening experience.

6. Experiments and Results

This section includes the evaluation of the model using various image examples. We aim to prove that the proposed methodology including the internal image preprocessing pipeline performs significantly better in terms of accuracy and Character Error Rates(CER) than a pipeline with no integration of the preprocessing module, OCR engine, and TTS engine. To set up this experiment, we use a public dataset containing textual images and compare the performances of the 2 methods mentioned above on these images. The results show that the accuracy of the model strongly depends on the quality of the input image.

A. Data Acquisition

The dataset we use for evaluating the performance of our internal image preprocessing pipeline consists of 30 test images in either .jpg or .png format obtained from the internet. Each image consists of text with different backgrounds, contrasts, different font sizes, and colors to introduce noise variety into the dataset. The dataset also consists of manually annotated text in a .csv file representing the same text as present in each of the test images. The dataset has been open-sourced on Kaggle [11] for further usage. This dataset allows for a comparison of the efficiencies of the ITTTS pipeline with internal image preprocessing against the same pipeline without the internal preprocessing module. By making sure of the same test images to compare both pipelines, we remove the chances of the quality of the input image affecting the results obtained.

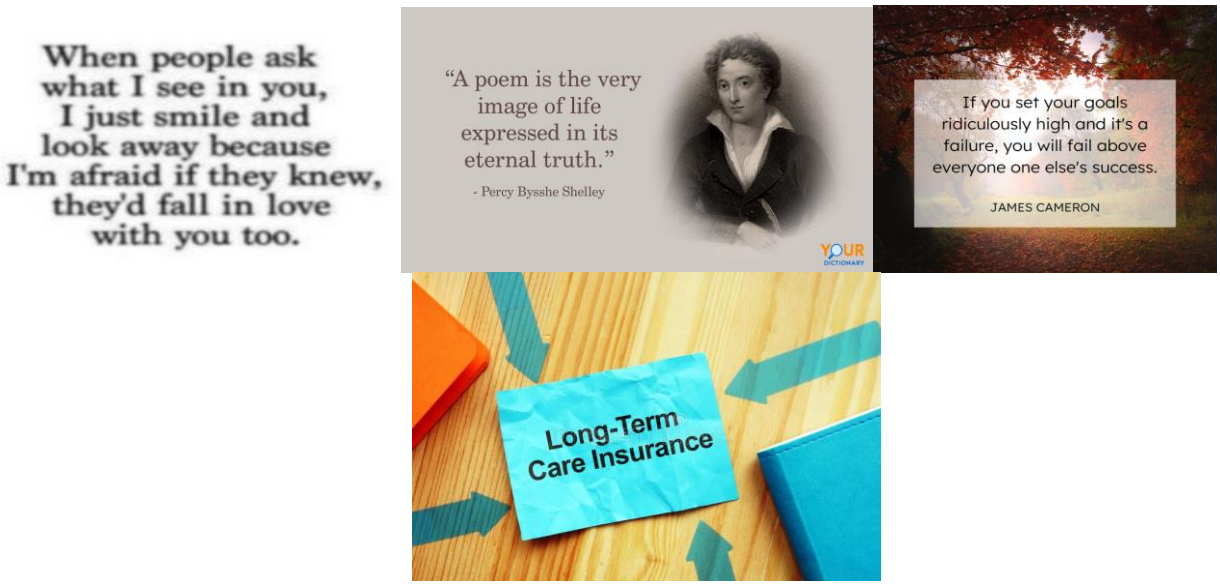

Fig.2. Sample test images present in the dataset

B. Hardware Specifications

The computer used for the experiments was an HP Pavilion CS1000TX with the following specifications: Intel Core I5-8265U with up to 3.9 GHz base frequency, 8 GB of DDR4-2400 SDRAM, 1 TB 5400 rpm SATA. The EasyOCR engine used for image to text conversion leverages the computation power of a 2 GB NVIDIA Geforce MX130 graphics card. No additional hardware was used. The specifications were chosen based on the computational demands of the experiments and the availability of the hardware. The use of this computer enabled us to perform internal image preprocessing and text extraction from the image by the OCR engine in a timely and efficient manner.

C. Results

A set of experiments are conducted to validate our proposed pipeline architecture. We compare the performance of the integrated pipeline with the image preprocessing module and the performance without the preprocessing module before the OCR engine. We compare them using various metrics used in the OCR literature such as character error rates (CER), word error rates (WER), similarity indices, and Levenshtein distances.

The above 4 metrics are computed for both pipelines and the mean of statistics is described in Table 1. and the variance of statistics is described in Table 2:

Table 1. Mean results of evaluated metric

Mean results of evaluated metrics				
Statistics	CER	WER	Similarity Index	Levenshtein Distance
With preprocessing	0.057	0.130	0.960	4.533
Without preprocessing	0.083	0.174	0.949	6.266

Table 2. Variance in results of evaluated metrics

Variance in results of evaluated metrics				
Statistics	CER	WER	Similarity Index	Levenshtein Distance
With preprocessing	0.001	0.013	0.0008	19.636
Without preprocessing	0.006	0.028	0.01	46.960

Table 1. shows that there is an average increase in the similarity index from 0.949 to 0.960 between the OCR output and the ground truth when our proposed pipeline is used in comparison to the pipeline without preprocessing. There was a significant average decrease of 5% in the word error rate after using the proposed image processing pipeline. Similar trends could be observed for the Character Error Rate(CER) metric that decreased by 3% by using our pipeline. The Levenshtein Distance decreased considerably on using our pipeline showing that the output from our pipeline was closer to the ground truth.

While table 2. shows the variance of the computed metrics. It can be observed that the Levenshtein distance has a very high variance of 46.96 without preprocessing and it reduced drastically to 19.64 when our pipeline was used. It can also be observed that the variance is lesser for the Character Error Rates, Word Error Rates, and Similarity Index showing that the proposed pipeline performance does not degrade irrespective of the quality of the input image by the user. This decrease in variance across all metrics could be attributed to the internal image preprocessing module integrated into the proposed pipeline.

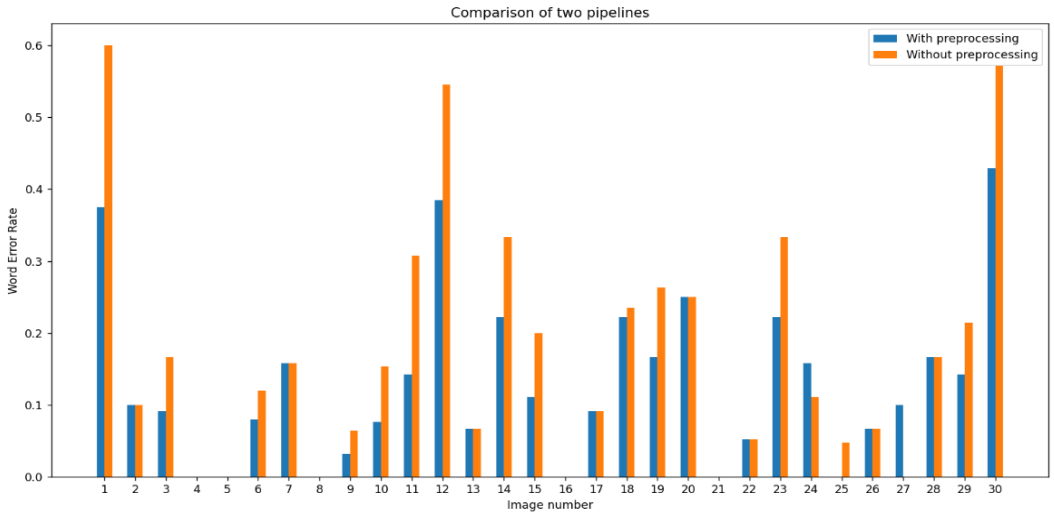

Fig.3. Comparison of Word Error Rates with the two methods

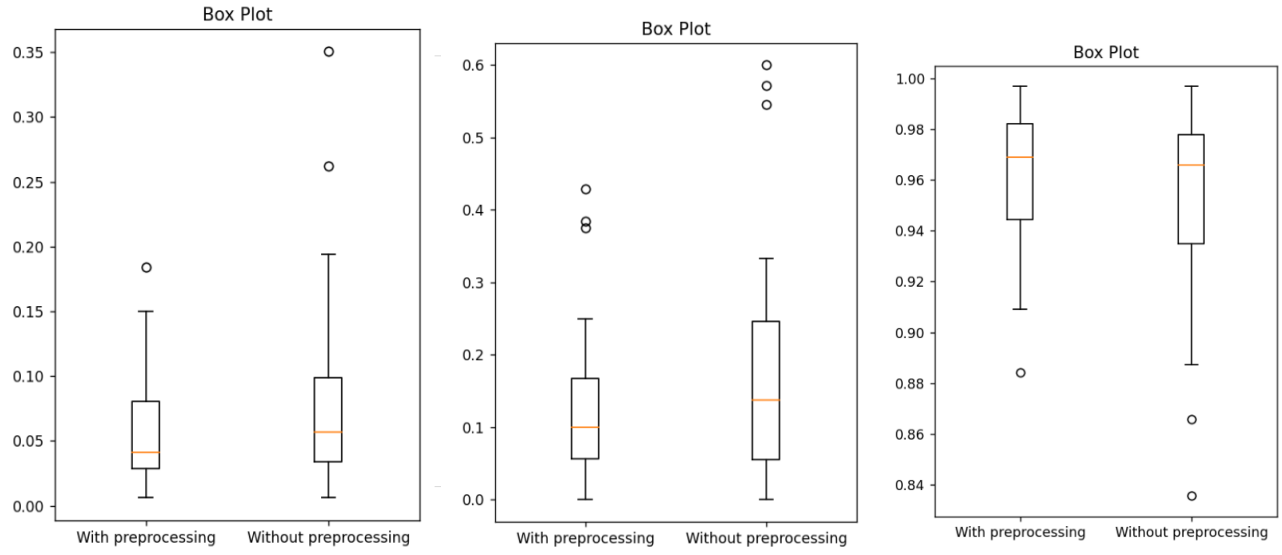

Fig.4. The figures show the boxplots for computed metrics, (a) Word Error Rate, (b) Character Error Rate, (c) Similarity Index respectively.

The boxplots in Fig.4. shows the distribution of Word Error Rates, Character Error Rates, and Similarity indices for the two pipelines under observation, namely the pipeline without a preprocessing module and the proposed pipeline with an integrated preprocessing module. It can be observed in all the 3 box plots that there exist outliers that are obtained from poor quality input images. In boxplot (a) for the Word Error Rates, it can be observed that the outliers have high WER reaching 0.35 without preprocessing in comparison with the outlier having WER of 0.18 with our proposed pipeline. Therefore, we can conclude that our pipeline exceeds the performance of a traditional pipeline without preprocessing across all the 4 metrics.

7. Conclusion and Future Work

A. Conclusion

In conclusion, we have implemented an integrated pipeline presented in this research paper that provides an efficient solution for converting images to speech by utilizing Optical Character Recognition (OCR) and Text-to-Speech (TTS) systems. By addressing the critical dependence of OCR accuracy on input image quality, the proposed internal image preprocessing pipeline effectively enhances the accuracy of image-to-text conversion.We have also open sourced a package in Python [12] that offers the ITTTS service to application developers. The package could be installed easily on the local machine including its dependencies. The work's significance lies in its holistic approach, which bridges the gap between image preprocessing, OCR, and TTS systems, offering a comprehensive solution for enhanced accessibility and usability. This integrated pipeline provides a tangible advancement over existing methods by producing more accurate and intelligible audio output from textual images. By effectively tackling challenges associated with image quality and ensuring smoother integration between text and speech conversion, this research contributes to the seamless conversion of textual information into accessible audio content.

The experiments conducted in this study demonstrate the effectiveness of the proposed pipeline in terms of accuracy and error rates in comparison to methods without the image preprocessing module proposed [13,14] . The suggested technique is a much better way to extract text from the input image and convert it into an audio output. The results also indicate that the proposed pipeline can handle different types of input images and produce high-quality speech output.

For future work, there are several directions to consider. First, the proposed integrated pipeline could be fine-tuned and optimized further to handle a wider range of image qualities and backgrounds, thus enhancing its robustness. Additionally, exploring ways to adapt the pipeline for different languages and accents could significantly extend its usability. The pipeline's applicability in domains such as education, healthcare, and information accessibility for visually impaired individuals could be further investigated and expanded upon. Moreover, future work could delve into optimizing the OCR and TTS engines for specific use cases or languages, potentially leading to even better performance. Continuously updating and expanding the pipeline based on new advancements in OCR and TTS technologies would ensure its relevance in an ever-evolving technological landscape.

References

[1] Sonia Bhaskar, Nicholas Lavassar and Scott Green, Implementing Optical Character Recognition on the Android Operating System for Business Cards, EE 368 Digital Image Processing.

[2] Abdullah-Al Mahmud, Ahmed Sabbir Arif, Md. Mahbubur Rahman, and Muhammad Abul Hasan, ”Development of an intelligent text-to-speech (ITTTS) system for visually impaired people,” Journal of Assistive Technologies, vol. 11, no. 2, pp. 91-99, 2017

[3] Mishra, A., Tiwari, V. (2019). Usability and Accessibility Evaluation of Intelligent Text to Speech (ITTTS) Software for Visually Impaired Users. Journal of Accessibility and Design for All, 9(1), 106-129.

[4] Aditya Bakshi, Sunanda Gupta et al., “3T-FASDM: Linear Discriminant Analysis based 3-Tier Face Anti-Spoofing Detection Model using Support Vector”, International Journal of Communication Systems, Wiley, 2020, vol 33, issue 12.

[5] Aditya Bakshi, Sunanda Gupta “An Efficient Face Anti-Spoofing and Detection Model Using Image Quality Assessment Parameters” in Multimedia Tools and Applications, 2020.

[6] Shakti, Aditya Bakshi “An Optimal Energy Efficient Spatial-Temporal Correlation Method for Data Aggregation in Wireless Sensor Networks” published in International Journal of Control Theory and Applications, ISSN : 0974-5572,Number 45(2016).

[7] Aditya Bakshi, Sunanda Gupta “A Taxonomy on Biometric Security and its Applications” International Conference on Innovations in Information and Communication Technologies.

[8] Aditya Bakshi and Sunanda Gupta” A Comparative Analysis of Different Intrusion Detection Techniques in Cloud Computing” published in 2nd International Conference on Advanced Informatics for Computing Research ,2018, CCIS 956, pp. 358–378.

[9] Zheng, C., Wang, B., Liu, Y., Yang, M., Han, J. (2021). EasyOCR: End-to-End Scene Text Recognition. Pattern Recognition, 114, 107778. doi: 10.1016/j.patcog.2021.107778.

[10] Gao, Z., Yang, Y., Chen, Y., Deng, L., Wang, Y. (2020). EasyOCR: A Practical Scene Text Recognition System. In 2020 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE. doi: 10.1109/ICME46284.2020.9102593.

[11] https://www.kaggle.com/datasets/shreyaspj/tiocr

[12] https://pypi.org/project/img2speech/

[13] Chucai Yi & Yingli Tian, 2014 Scene Text Recognition in Mobile Applications by Character Descriptor and Structure Configuration, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 7, JULY 2014

[14] Julinda Gllavata', Ralph Ewerth' and Bemd Freisleben’ 2003 , A Robust Algorithm for Text Detection in Images, Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis (2003).

Authors’ Profiles

Shreyas Reddy is currently a student pursuing a bachelor’s degree in Information Technology from the International Institute of Information Technology, Bhubaneswar, India. His research interests include working on designing efficient Machine Learning and Deep learning models for various business cases. He also has experience working on projects relating to upper arm prosthetics and gait analysis. He will be graduating with a bachelor's degree from IIIT in 2023.

Rashmi Ranjan Das is currently a student pursuing a bachelor’s degree in Information Technology from the International Institute of Information Technology, Bhubaneswar, India. His interests include designing Web applications for deploying useful ML models and formulating ML problem statements from real world problems. He will be graduating with a bachelor's degree from IIIT in 2023.

Dr. Anjali Mohapatra received a Msc. degree in Computer Science Engineering from Utkal University, Bhubaneswar, India in 2001. She has also received a phD in Computer Science from Utkal University, Bhubaneswar, India in 2008. She is currently a HOD of the Computer Science Department at the International Institute of Information Technology, Bhubaneswar, India. She is also an assistant professor in the Computer Science department at the International Institute of Information Technology, Bhubaneswar, India. Her research activities have been focused on computational biology,Bioinformatics and pattern recognition.

How to cite this paper: Shreyas Reddy, Rashmi Ranjan Das, Anjali Mohapatra, "An Integrated Pipeline with Internal Image Processing for Efficient Image to Text to Speech Conversion", International Journal of Engineering and Manufacturing (IJEM), Vol.13, No.6, pp. 1-8, 2023. DOI:10.5815/ijem.2023.06.01