

Extraction of Text from Image with Audio Conversion Using Deep Learning

Ms. Adari.Devi Charishma^{*1}, Mrs.P.Sri Jyothi^{*2}

¹MCA Student, Department of Master of Computer Applications,
Vignan's Institute of Information Technology(A), Beside VSEZ,Duvvada,Vadlapudi Post,
Gajuwaka, Visakhapatnam-530049.

²Assistant Professor, Department of Information Technology,
Vignan's Institute of Information Technology(A), Beside VSEZ,Duvvada,Vadlapudi Post,
Gajuwaka, Visakhapatnam-530049.
vignaniit.edu.in

Abstract:

In diverse scenarios, the ability to extract text from images proves to be immensely valuable, facilitating the transformation of both printed and handwritten text into a digital format. This capability not only streamlines data entry processes but also opens avenues for enhanced accessibility and information retrieval. However, there exist situations where visual access to text is inherently constrained, particularly for individuals with visual impairments. In such instances, an additional imperative arises—to seamlessly convert the extracted text into audio, thereby catering to the diverse needs of users with varying modes of information consumption. The focal challenge at hand revolves around the development of a robust system that seamlessly extracts textual information from images and subsequently converts it into an audio format. The overarching objective is to craft an automated solution adept at processing images containing textual elements, employing advanced Optical Character Recognition (OCR) techniques for accurate text recognition, and finally transmuting the extracted textual content into a comprehensible audio representation. This interdisciplinary pursuit not only aligns with the broader technological advancements in image processing and machine learning but also addresses the crucial aspect of inclusivity by catering to the unique needs of individuals with visual impairments. Through the proposed system, a seamless bridge is established between visual content and auditory comprehension, unlocking new dimensions of accessibility and usability in the digital landscape..

Keywords: Optical Character Recognition (OCR), Handwritten Text, Visual Content ,Auditory Comprehension.

1. INTRODUCTION

A groundbreaking program that ingeniously amalgamates speech synthesis and computer vision technologies is introduced herein, showcasing its prowess in extracting text from photos and seamlessly converting it into an audio format. This innovative approach holds significant promise, particularly in rendering textual content within photographs accessible to individuals with visual impairments or those who prefer audio-centric content consumption. The underpinning technology leverages sophisticated Optical Character Recognition (OCR) techniques, which meticulously analyze the visual composition of images to discern and extract textual elements. Remarkably, the evolution of OCR algorithms has reached a stage where they can accurately decipher text even in intricate photos, spanning diverse fonts, sizes, and backgrounds. Upon successful extraction of text, the subsequent phase involves the transformation of this textual data

into an audio file.

Text-to-speech (TTS) synthesis methods take center stage in this process, employing voice databases and linguistic principles to produce an audio output that closely mimics natural human speech patterns. This synthesis not only enhances the immersive quality of the auditory experience but also broadens the accessibility spectrum.

The practical applications of this text extraction from photos with audio conversion are manifold. Beyond aiding individuals with visual impairments by allowing them to access written information through a simple picture, this technology caters to diverse preferences, enabling those who opt for listening over reading. In contexts where reading is impractical, such as while driving or multitasking, this innovation proves particularly invaluable. Moreover, the versatility of this technique extends across various industries. Its utility in automatically transcribing handwritten notes or scanning documents simplifies the conversion of analog information into digital text. By making printed information accessible to visually impaired students, it fosters inclusivity in educational environments. Beyond these applications, it finds relevance in document analysis, content indexing, and archiving, showcasing its potential to streamline various professional processes. In conclusion, the extraction of text from photos with audio conversion not only holds immense promise for enhancing accessibility and information retrieval but also embodies a technological synergy between computer vision and voice synthesis. As these fields progress, this transformative technology opens new avenues for inclusive communication and information dissemination, marking a significant step forward in the pursuit of comprehensive accessibility and convenience across diverse disciplines

2. LITERATURE SURVEY

The most important step in the software development process is the literature review. This will describe some preliminary research that was carried out by several authors on this appropriate work and we are going to take some important articles into consideration and further extend our work.

In this study, a novel deep learning-based text extraction technique tailored for low-resolution photos is presented. The initial step involves pre-processing the low-resolution images to enhance text edges and contrast. Subsequently, a Refine Net deep learning network is trained on these pre-processed photos to extract text. To bolster the precision of text extraction, Refine Net incorporates a multi-level feature fusion technique. The authors evaluated their approach on the ICDAR dataset and compared it with state-of-the-art methods. The results showcased the superior performance of their strategy in terms of F1-score, precision, and recall. This innovative approach proves effective in extracting text from low-resolution photos, holding potential applications in industries such as remote sensing and video surveillance.

In a paper titled "Upload an Image Using Image Processing" by Deepak Chaudhary, Prateek Agrawal, and Vishnu Madam from ICAICR2019, the authors propose a method for validation using image processing techniques. The paper outlines a comprehensive process involving image acquisition, reprocessing, segmentation, feature extraction, validation, and verification. The method aims to enhance the quality of a bank cheque image, extract crucial information, and validate it against predefined rules and standards. The proposed technique holds promise for improving the accuracy and efficiency of validation processes, particularly in the context of financial transactions.

Another paper, "Image Security Enhancement in Online" by K. Dhanva, M. Harikrishnan, and P. U. Babu from ICICCT2018, addresses the security challenges associated with transmitting and storing images electronically. The authors propose a

three-tier security framework incorporating strong encryption techniques to safeguard image transmission over networks. The framework aims to provide robust protection throughout the lifecycle of images in online systems, from transmission to storage. Through encryption, digital watermarking, and secure storage mechanisms, the authors strive to mitigate potential security threats and enhance overall image security in the online domain.

The paper titled "Text Extraction based on Pytesseract" by Wu, Y. Cheng, J., & Li, X. (2019) introduces a method leveraging Pytesseract for text extraction. The proposed technique involves two stages: text detection using the EAST model and text recognition using a convolutional neural network (CNN). The authors evaluated their method on a dataset of traffic signs, demonstrating high accuracy in both text detection and recognition. The results suggest that their approach effectively extracts text from traffic signs, contributing to improved safety and efficiency in transportation systems.

In the paper "GTTS is Based on Converting the Text to Speech" (ArXiv:1904.03670, 2019), the authors explore the conversion of text to speech using the Google Text to Speech API (gTTS). The gTTS API, a widely used utility, allows simple translation of entered text into audio, saved as an mp3 file. The API supports multiple languages and provides options for adjusting the audio velocity. However, as of the latest update, changing the voice of the generated audio is not feasible. This paper sheds light on the practical use of gTTS for text-to-speech conversion, showcasing its simplicity and versatility in generating spoken content in various languages and speeds.

3. EXISTING SYSTEM

Within the realm of translation, the inclusion of images stands out as a pivotal component, influencing the effectiveness and accuracy of the translation process. Traditionally, the transformation of speech involves a meticulous and time-consuming approach. The conventional method necessitates a thorough examination of the speech, demanding manual entry of details. This arduous task not only consumes a significant amount of time but also entails a substantial workforce. As we navigate the nuances of this traditional process, the importance of streamlining and modernizing translation practices becomes evident. The integration of innovative technologies and methodologies could potentially alleviate the burdens associated with manual entry, offering a more efficient and resource-effective alternative to enhance the overall translation experience.

Limitations of the Existing System:

- 1) Time-Consuming Process:**The current system suffers from a fundamental limitation due to its inherently time-consuming nature, impeding efficiency.
- 2)Complex Execution:**The complexity embedded in the process poses a challenge, making execution intricate and prone to errors.
- 3)High Manpower Dependency:**The system's heavy reliance on substantial manpower raises concerns about scalability and operational costs.

4)Questionable Adaptability:The system's adaptability to evolving technological landscapes is uncertain, potentially rendering it obsolete in the face of advancements.

5)Logistical and Cost Challenges:The significant manpower requirement leads to logistical challenges and increased costs, impacting the feasibility of the existing methodology.

4. PROPOSED SYSTEM

The proposed system aims to enhance the limitations of the existing system with the following things.They are as follows:

1) Efficient Text Extraction:The proposed system integrates a deep learning algorithm for text extraction from images, ensuring a more efficient and accurate process compared to the existing system.

2)Enhanced Automation with OCR:By incorporating Tesseract OCR (Optical Character Recognition), the system automates the extraction of details from text, reducing the need for manual input and minimizing errors.

3)Text-to-Speech Integration:The inclusion of a Text-to-Speech (TTS) component enriches the system, enabling the conversion of extracted text into spoken words, enhancing accessibility and user experience.

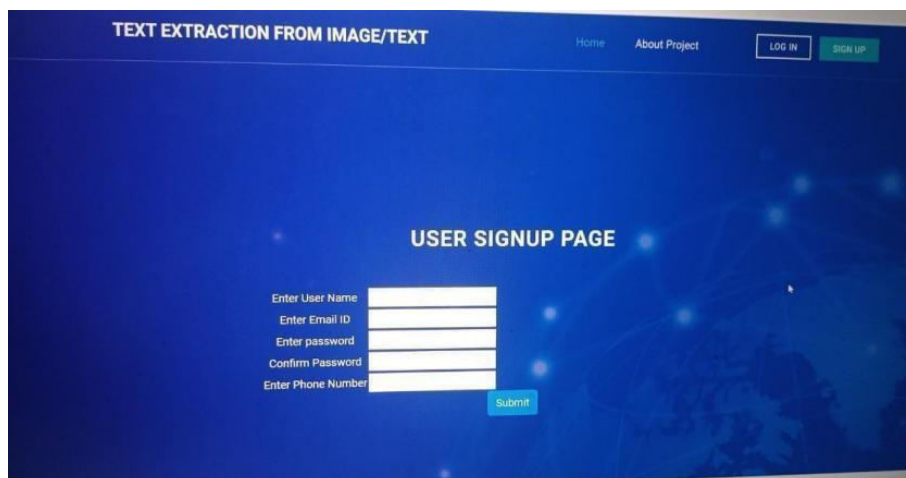
4)Multilingual Capability:Leveraging Google translation, the proposed system offers a seamless conversion of extracted data into multiple languages, broadening its applicability and catering to diverse user preferences.

5)Streamlined and Faster Process:The proposed system streamlines the overall process, leveraging advanced technologies to significantly reduce time consumption and enhance the efficiency of text extraction and translation.

5. EXPERIMENTAL RESULTS

From the below two figures it can be seen that proposed model is more accurate in order to prove our proposed system.

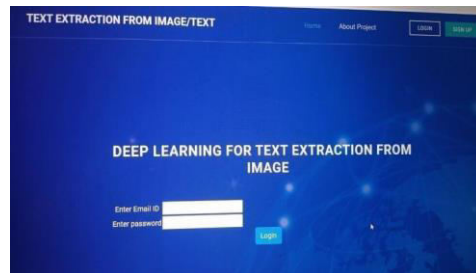
User Signup Page



The screenshot displays the 'USER SIGNUP PAGE' of a web application. The page has a dark blue background with a subtle pattern of light blue dots. At the top, the text 'TEXT EXTRACTION FROM IMAGE/TEXT' is visible on the left, and navigation links for 'Home', 'About Project', 'LOG IN', and 'SIGN UP' are on the right. The main heading 'USER SIGNUP PAGE' is centered. Below it, there are five input fields with labels: 'Enter User Name', 'Enter Email ID', 'Enter password', 'Confirm Password', and 'Enter Phone Number'. A blue 'Submit' button is located at the bottom right of the form area.

Explanation:Here the user try to signup with his basic details.

User Login Page



Explanation:Here the user try to login with valid details what he registered.

User Choose Input:



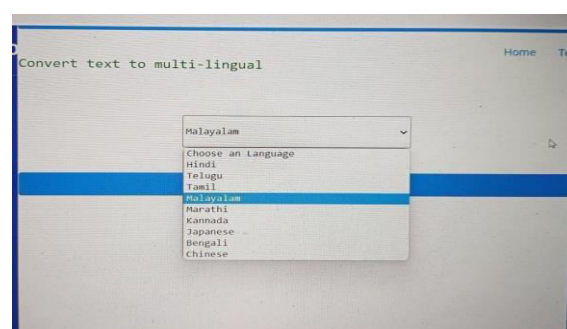
Explanation:Here the user try to choose input as image file and then try to convert.

Image to Text:



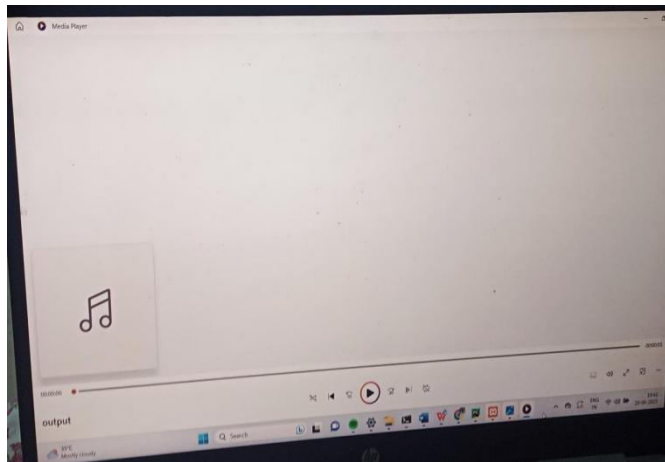
Explanation:Here the user try to choose input as image file and then try to extract the text from that input image.

Change Text to Multilingual



Explanation: Here the user try to convert the text into multilingual.

Output is Displayed as Audio



Explanation: From the above window we can see the output is formed as audio.

6. CONCLUSION

The technology operates through the utilization of Optical Character Recognition (OCR) software, a tool designed to convert textual content within an image into machine-encoded text. This converted text can then be audibly presented using Text-to-Speech (TTS) software, facilitating accessibility for users who may face challenges in reading the text independently. This inclusive approach benefits individuals with visual impairments or those unfamiliar with the language depicted in the image. Despite its merits, the technology does come with certain limitations. OCR, while powerful, may encounter difficulties in accurately recognizing all text within an image, particularly if the text is handwritten or presented in an unconventional font. Furthermore, TTS software might not consistently reproduce the original tone and inflection of the text, introducing nuances that could impact the overall user experience. As the technology continues to evolve, addressing these limitations becomes pivotal for ensuring its widespread effectiveness and usability across diverse contexts..

Declaration

1. All authors do not have any conflict of interest.
2. This article does not contain any studies with human participants or animals performed by any of the authors.

References

- 1) Shi, B., Bai, X., & Yao, C. (2015). "Scene Text Recognition with LSTM Recurrent Neural Networks." In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2015. This paper introduces a method for recognizing text in images using a deep LSTM recurrent neural network.
- 2) Hannun, A., et al. (2014). "DeepSpeech: An Open Source Speech Recognition Engine." arXiv:1412.5567. This paper presents a deep learning-based approach to

speech recognition, utilizing a combination of convolutional and recurrent neural networks.

- 3) Xu, H., Yu, K., & Dong, J. (2018). "A Comprehensive Survey on Deep Learning for Speech Recognition." arXiv:1812.06448. This survey paper provides an overview of recent advancements in deep learning for speech recognition, covering various neural network architectures and training techniques.
- 4) Breuel, T. (2009). "OCROPUS: A Free Document Analysis and Optical Character Recognition System." In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2009. This paper introduces a free and open-source OCR system with pre-processing and post-processing steps to enhance accuracy.
- 5) Amodei, D., et al. (2015). "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin." In International Conference on Machine Learning (ICML), 2015. This paper presents an end-to-end deep learning model for speech recognition achieving state-of-the-art results on benchmark datasets.
- 6) Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks." In Proceedings of the 23rd International Conference on Machine Learning (ICML), 2006. This influential paper introduces Connectionist Temporal Classification (CTC) as a technique for sequence labeling with recurrent neural networks.
- 7) Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation*, 18(7), 1527-1554. This paper by Geoffrey Hinton and colleagues introduces a fast learning algorithm for training deep belief networks.
- 8) Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems (NIPS)*, 2012. This landmark paper presents the architecture known as AlexNet, which significantly advances image classification using deep convolutional neural networks.
- 9) Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory." *Neural Computation*, 9(8), 1735-1780. This foundational paper introduces Long Short-Term Memory (LSTM) networks, a crucial advancement in recurrent neural network architecture.
- 10) LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE*, 86(11), 2278-2324. This classic paper by Yann LeCun and colleagues discusses the application of gradient-

based learning to document recognition, contributing to the development of modern OCR systems.