International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 25 Issue 03, MAR, 2025

AUTOMATED TEXT EXTRACTION AND TRANSLATION FROM VIDEO FRAMES & IMAGES USING ADVANCED DEEP LEARNING TECHNIQUES

Mr.Buradagunta Avinash[1], Bomminayuni.Durga Prasadu[2], Chipurupalli.Kalidas[3],Kurmala Jaswanth Sai Tarun[4],Dandu.Kesava[5]

[1]Assistant Professor, IT Department, Vasireddy Venkatadri Institute of Technology, Namburu, Guntur, Andhra Pradesh -5225208
[2,3,4,5]UG Students, IT Department, Vasireddy Venkatadri Institute of Technology, Namburu, Guntur, Andhra Pradesh -5225208

Abstract - *In order to extract, refine, translate, and vocalize text from photos and videos, this research proposes an intelligent media processing system. In order to improve text accuracy by eliminating redundant, noisy, and nonsensical words, the system uses a multi-step preprocessing pipeline after utilizing Tesseract for optical character recognition (OCR) to extract text from visual data. To make the retrieved text more readable and coherent, it is deduplicated and filtered using NLP (natural language processing) techniques. Additionally, the technology allows users to transform retrieved text into many target languages by integrating Google Translate to allow for multilingual translation. In order to improve accessibility, Google Text-to-voice (gTTS) is used to convert the translated text into voice, giving language learners and users with vision impairments an audio output. Users may input photos and videos, process them effectively, and obtain extracted, interpreted, and audio-rendered material in real time thanks to the workflow's implementation as a Django-based web application. Our method combines similarity-based deduplication to increase output accuracy, adaptive thresholding for better text identification, and deep learning-based OCR improvements. Potential uses for this technology include intelligent media analysis, multilingual content accessibility, automated news summarizing, and assistive technologies for the blind and visually handicapped. The outcomes of our experiments show how well our approach extracts high-quality text from intricate multimedia sources, indicating the possibility for more developments in OCR-based language analysis and accessibility services.*

***Key Words*:** *Tesseract, Text Extraction, Text Translation, Speech Synthesis, Google Text-to-Speech (gTTS), Video Processing, Optical Character Recognition (OCR).*

I INTRODUCTION

Multimedia content, such as pictures and videos, is a major way that information is shared in the current digital age on a variety of platforms, such as social media, news websites, and instructional materials. Nevertheless, it is still difficult to glean valuable insights from such material, particularly when unstructured text is included into picture and video frames. With its ability to digitize and extract text from photos and videos, optical character recognition (OCR) technology has become a potent tool for document automation, content translating, and accessibility improvement.

The goal of this project is to create an intelligent audiovisual processing system that can automatically extract, refine, translate, and synthesize speech from visual input. The system can process images and videos, use Tesseract OCR to extract textual data, refine the text that was extracted by removing unnecessary and redundant phrases, use Google Translate to translate the material into multiple languages, and use Google Text-to-Speech (gTTS) to produce an audio output. Through the incorporation of these elements into a web application built using Django, the system offers an intuitive user interface for smooth multimedia evaluation of texts and accessibility.

The growing demand for accurate and effective text extraction from various media is what inspired this effort. Poor extraction accuracy results from existing OCR algorithms' frequent struggles with noise, damaged text, and irrelevant material. In order to overcome these constraints, our method improves the readability and pertinence of collected text by utilizing sophisticated preprocessing methods as adaptive thresholding, grayscale conversion, and text similarity-based deduplication. Furthermore, this technology is useful for assistive technologies and multilingual accessibility due to its inclusion of machine translation and speech synthesis.

The increasing reliance on video and picture data in a variety of domains, such as accessibility, education, research, and digital archiving, is what

spurred this study. For example, students frequently need to extract notes from lectures or annotations from films in order to continue their studies in educational environments. Similarly, effective processing of massive amounts of video data is required in fields like multimedia analysis and monitoring. Users' capacity to access, evaluate, and make use of vital information is restricted in the absence of a reliable solution for the smooth extraction of text from multimedia sources.

Cutting-edge OCR systems like Keras_OCR and PyTesseract, in conjunction with developments in Deep Learning, provide a revolutionary means of overcoming these obstacles. Even in complicated situations, text extraction accuracy and efficiency may be improved by utilizing the capabilities of deep learning algorithms for text detection and identification in conjunction with the adaptability of OCR technologies. In addition to addressing the shortcomings of conventional approaches, this technological convergence opens up possibilities for real-time and scalable applications.

Multimedia data analysis, real-time content translation, automated news summarization, and assistive technology for the blind and visually handicapped are just a few of the many fields in which this system finds extensive use. Our method advances intelligent multimedia comprehension and accessibility by enhancing the precision of OCR-based extraction of text and integrating multilingual processing.

II LITERATURE REVIEW

For text extraction from photos and videos, optical character recognition, or OCR, has been extensively studied and used. As demonstrated by early implementations like the Tesseract OCR engine, which has developed to incorporate deep learning for increased accuracy, traditional OCR devices rely on pattern recognition and methods for extracting features [1]. Preprocessing methods including picture binarization, noise elimination, and detection of edges have been used in a number of studies to increase OCR systems' ability to recognize text against complicated backgrounds [2]. Additionally, OCR performance has been greatly enhanced by deep learning-based techniques like CNNs (Convolutional Neural Networks) and Recurrent Neural Networks (RNNs), especially for handwritten and damaged text recognition [3].

A novel technique for identifying text on human beings in sports photos is presented by Pinaki Nath Chowdhury and Palaiahnakote Shivakumara [4], who address issues including low picture quality and varying camera perspectives. In contrast to traditional techniques, it uses an end-to-end episodic learning strategy that uses a Pyramidal Pooling Module (PPM) for spatial attention mapping and a Residual Network (ResNet) to identify garment areas. The Progressively Scalable Expansion Algorithm (PSE) is used to recognize text[5]. Analysis of many datasets shows better accuracy and F1-score than current approaches, proving efficacy with varying inputs.

Palaiahnakote Shivakumara and K. S. Raghunandan [6] discuss reliable text identification and detection in multiscript-oriented photographs. Bit plane slicing, Iterative Nearest Neighbor Symmetry (INNS), Mutual Nearest Neighbor Pair (MNNP) components, character detection through fixed windows, contourlet wavelet includes with SVM classifier, and Hidden Markov Models (HMM) for recognition are some of the methods used in earlier studies [7].

A technique for precise text identification in photos of natural scenes is presented by Xu-Cheng Yin and Xuwang Yin [8]. Maximally Stable Extremal Regions (MSERs) are extracted as character candidates using a fast-pruning technique, and then they are grouped into text candidates using single-link clustering [9]. A character classifier is used to estimate the posterior probability of text candidates in order to remove non-text areas, and automatic learning of distance metrics and thresholds is integrated. The ICDAR 2011 Robust Reader Competition database evaluation shows an f-measure of more than 76%, outperforming state-of-the-art techniques, and additional validation on many databases attests to its efficacy [10].

However, there are still a lot of challenges to be solved in spite of recent advances in deep learning-based Video reading systems. The lack of large-scale, diverse datasets for testing and training is one of the biggest problems [11]. It is challenging to create and assess Video reading algorithms that can identify a broad variety of language and speech patterns due to the scarcity of such datasets. The reason behind this is that Video reading systems need a significant amount of training data in order to understand the intricate relationships between speech sounds and visual cues [12].

Enhancing Video reading systems' precision and resilience has significant ramifications for applications in safety, surveillance, and human-computer interaction in addition to those who are hard of hearing [13]. In addition to increasing the accuracy of systems that recognize speech in loud environments, Video reading systems provide the potential to improve accessibility and communication for people with hearing impairments [14]. Video reading systems may be used in surveillance and safety applications to follow and

identify people in video footage, even if their faces are entirely or partially hidden [15].

We can lessen the obstacles that persons with hearing impairments encounter in their everyday lives and increase the precision and resilience of communication for them by creating more efficient Video reading systems [16]. Even when a person's face is partially or completely hidden in video footage, Video reading algorithms may be utilized to identify and follow them in surveillance and security applications [17]. This can increase technology's usability and accessibility for a variety of users while also preventing crime and enhancing public safety [18].

The goals of this work are to improve accessibility and communication for those with hearing impairments, develop a more reliable and accurate Video-reading system through deep learning, and investigate the possible uses of Video-reading systems in security, monitoring, and human-computer interaction [19].

By combining OCR, NLP-based text maintenance, translation, and TTS into a single system, this effort expands on previous research by facilitating smooth text extraction and voice conversion from pictures and movies [20]. The suggested solution seeks to improve usability and accessibility for a range of applications by utilizing deep learning-powered OCR, sophisticated text processing, and high-accuracy translation.

III METHODOLOGY

A. Objective:

This project's goal is to develop an automated and effective system that can extract, analyze, translate, and turn text from various media into speech. The system seeks to improve accessibility and usability for a wide range of users by utilizing Text-to-Speech (TTS) technology for audio creation, Natural Language Processing (NLP) for improvement, machine translation for international support, and Optical Character Recognition (OCR) for text extraction. The project is intended to help those who require text conversion from pictures to videos, such as those who are visually impaired, have trouble communicating in another language, or have professional needs like localizing and summarizing material. The final objective is to offer a smooth, intuitive system that guarantees meaningful translation, excellent text extraction accuracy, and speech output that sounds natural.

B. Existing System:

Manual transcription or simple OCR technologies are the mainstays of the conventional method for extracting text from audiovisual information. Inaccurate text extraction results from existing OCR algorithms' frequent struggles with complicated video frames, low-resolution pictures, and noisy backgrounds. Additionally, the majority of systems provide outputs that are redundant or useless due to a lack of sophisticated text refining tools. Furthermore, integrated translation services and audio output for improved accessibility are not frequently included in current systems. For text the extraction process, translation, and textual-to-speech conversion, users frequently need to use a variety of tools, which makes the process laborious and ineffective as shown in Fig 1.

C. Proposed System:

An automatic and unified structure for text the extraction process, translation, and audio synthesis from photos and videos is presented by the suggested system. The system effectively recovers text from a variety of multimedia sources using Tesseract OCR. Advanced preprocessing methods including thresholding, noise reduction, and grayscale conversion are used to increase text accuracy. By eliminating superfluous and pointless words, a text refining module improves the collected material even more. Users may also translate extracted content into other languages using a multilingual translation option that is supported by Google Translate. The system uses gTTS, which translates text into voice, to increase accessibility and make it usable by those with visual impairments. All of the system's functionality is implemented as an a web-based Django application, which makes it simple for users to input media assets and access processed written and audio outputs. When compared to the current method, this integrated approach greatly improves accessibility, accuracy, and efficiency.

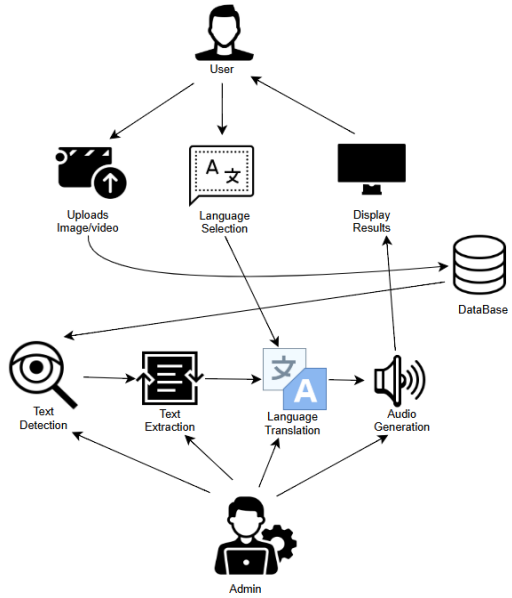

Fig. 1 Proposed System Architecture

D Algorithms Used

1. Optical Character Recognition (OCR) - Tesseract OCR:

An open-source recognition of text engine called Tesseract OCR uses sophisticated image processing methods to extract text from picture and video frames. Character classification, feature extraction, and picture preparation are all steps in the algorithm's multi-step procedure. To improve text visibility, the supplied image first goes through noise reduction and grayscale conversion. After that, the algorithm breaks the image up into distinct characters, which are then processed using edge detection and adaptive thresholding methods. Tesseract improves accuracy, particularly for handwritten and printed text, by using recurrent neural networks based on Long Short-Term Memory (LSTM) to categorize and detect letters.
To further improve the collected text, the OCR system incorporates post-processing methods including confidence-based filtering and dictionary-based correction. These improvements lessen mistakes brought on by distorted text, loud backdrops, or different font styles. To guarantee the usefulness of the processed data, the extracted text is then sent to further modules for additional polishing, translation, and speech synthesis.

2. Natural Language Processing (NLP) - Text Cleaning and Refinement:

The project uses NLP approaches for text cleanup and refining in order to increase the extracted text's accuracy and readability. Tokenization is the first step in the process, which divides the text into meaningful chunks like words and phrases. Similarity-checking algorithms like the SequenceMatcher from the difflib package, which examines phrase patterns and eliminates almost equivalent information, are used to identify redundant and repeated sentences. To improve text readability, regular expressions (regex) are also used to filter out unnecessary words, special characters, and undesired phrases.
Before the extracted text is subjected to additional processing, such translation or text-to-speech conversion, this stage makes sure it remains coherent and accurate. The system enhances text quality by utilizing natural language processing (NLP) approaches, which makes it more appropriate for real-world uses such as automated content analysis, accessibility aids, and document digitalization.

3 Machine Translation - Google Translate API:

To enable multilingual support, the system integrates machine translation via the Google Translate API. In order to capture contextual linkages, the system uses Transformer-based deep learning algorithms, which parse text sequences utilizing positional encoding and self-attention processes. Transformer models enhance accuracy by comprehending sentence structures and meaning holistically as opposed to word-by-word translation, in contrast to conventional phrase-based translation techniques.
The API maintains contextual correctness while translating retrieved text into the target language using parallel processing. Additionally, by including an intermediary validation stage where the translated text is examined for coherence, the method reduces frequent translation mistakes. This functionality is very helpful for cross-language accessibility, allowing a broad variety of users globally to understand the extracted information.

4. Text-to-Speech (TTS) - gTTS (Google Text-to-Speech):

A text-to-speech conversion technique called gTTS is used in the project to create speech that sounds human from text input. The system makes use of neural network-based speech synthesis models, including WaveNet, which anticipate waveform patterns at every step in order to produce high-fidelity audio. WaveNet uses probabilistic models to provide more expressive and natural-sounding voice output than conventional concatenative speech synthesis techniques, which depend on pre-recorded speech units.

The system offers language-specific voice synthesis, which lets users choose their favorite language for audio output, to improve the user experience. The compressed audio format in which the produced speech is stored allows for easy integration with multimedia programs. Because it enables them to receive textual material in an audio format, this function is especially helpful for those who have visual impairments or linguistic challenges.

IV IMPLEMENTATION

A. Image and Video Processing for Text Extraction:

Processing the incoming media files—pictures or videos—to extract textual content is the first step in project implementation. Images are read and preprocessed by transforming them to grayscale using OpenCV (cv2). In this stage, color-related noise is reduced, improving text visibility. Additionally, the picture is refined using binarization techniques like adaptive thresholding before being sent into Tesseract OCR. After processing the image, the Optical Character Recognition (OCR) engine converts the text into a machine-readable format.
The method for movies uses OpenCV to extract frames at predetermined intervals. At a certain time (e.g., every 100 frames), the algorithm examines the video file and extracts frames. Every extracted frame is subjected to the same preprocessing methods as photos, such as thresholding and grayscale conversion. Each frame's text is then identified and extracted using Tesseract OCR. To guarantee that the finished product is coherent and well-structured, the collected text from several frames is combined and polished to eliminate repetition.

B. Text Cleaning and Optimization:

The system uses methods from natural language processing, or NLP, to enhance the quality of the retrieved raw text. Repetitive phrases, mistakes, and extraneous features (such as logos, timestamps, or unnecessary words) are frequently included in the retrieved text. Using normal expressions and specified phrase filters, the debugging module first eliminates unnecessary characters. To prevent repetition, near-duplicate phrases are identified and eliminated using the SequenceMatcher algorithm.
Through the use of the language_tool_python package, grammatical errors are found and fixed. This guarantees that the retrieved text will remain coherent and readable. Before undergoing additional processing, such translation or text-to-speech conversion, the cleaned text is momentarily retained.

C. Multilingual Text Translation:

The project incorporates an automated translation module utilizing the Google Translate API to offer accessibility on a worldwide scale. Users can designate a language to be translated, and the system will translate the retrieved content in that language. A Transformer model based on deep learning is used for the translation process, which makes use of attention processes to preserve contextual correctness.
The translated text is reprocessed for coherence as part of a validation stage. This phase increases the system's overall accuracy by guaranteeing that the translated material is grammatically sound and meaningful. Users may then listen to the extracted material in their choice language thanks to the final transformed output being ready for text-to-speech translation.

D. Text-to-Speech Conversion (TTS):

Using Google Text-to-Speech (gTTS) to turn the translated text into speech is the last stage of execution. After processing the text, the algorithm chooses a suitable language model and creates voice that sounds human. An audio file (.mp3 format) containing the synthesized speech is stored and made accessible for user playback.
The system makes use of deep learning-based synthetic speech models, such as WaveNet, which produce fluid and natural-sounding speech, to guarantee high-quality audio output. Users may listen to the collected and translated material straight from their browser thanks to a link in the web application interface that links to the created audio file.

E. Web Application Interface and User Interaction:

The user interface of the implementation is a web application built with Django. Users can submit picture or video files to text extraction using the application's upload page. File uploads are handled by Django's FileSystemStorage, which stores the media files in the specified directory (MEDIA_ROOT). Following upload, the file is processed by the system, which also extracts text, interprets it, and produces audio output.
The collected text, translated material, and a link to the produced audio file are all dynamically displayed on the web interface. The program is very accessible and engaging, allowing users to engage with the system by choosing a target language while being exposed to the translated information.

V RESULTS

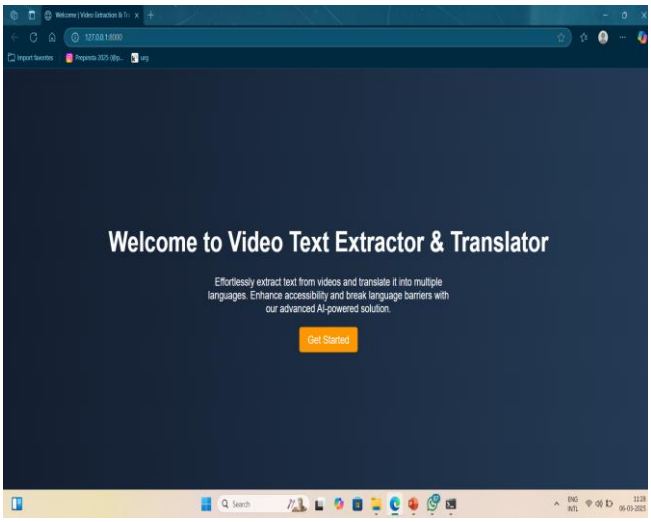

Fig 2: Home Page of Video Text Extractor & Translator

This illustration in Fig 2, shows the main page of the suggested online application, "Video Text Extractor & Translator." Users may start the text extraction and translation process from the interface's user-friendly welcome screen. By automating the process of extracting text from photos and videos, followed by audio conversion and translation, the technology seeks to improve accessibility. With a call-to-action button titled "Get Started," the homepage's design is simple and eye-catching, directing users to the system's main features. The technology uses cutting-edge AI-powered methods to parse textual data from multimedia material efficiently and with a smooth user experience.

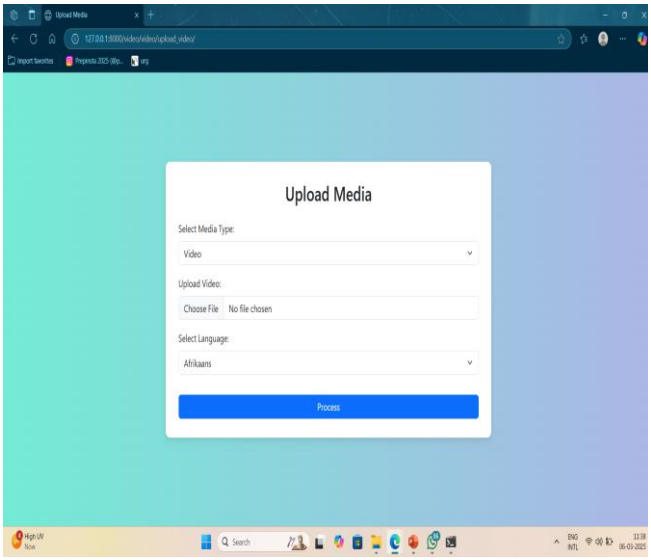

Fig 3: Media Upload Interface

The media upload interface, seen in Fig 3, allows users to choose and submit an image or video file for processing. Users may pick the target language for translation, upload the required file, and choose the media format using the interface. The material is processed by the system to extract text, improve it, translate it, and provide an audio output when the user uploads the file. By making it simple to choose video files and language settings, the interactive form guarantees user ease. To effectively handle a variety of media types and guarantee high text extraction accuracy, the system integrates strong backend processing algorithms. This interface streamlines the user workflow for efficient multimedia text translation and acts as the entry point to the main features of the suggested system.

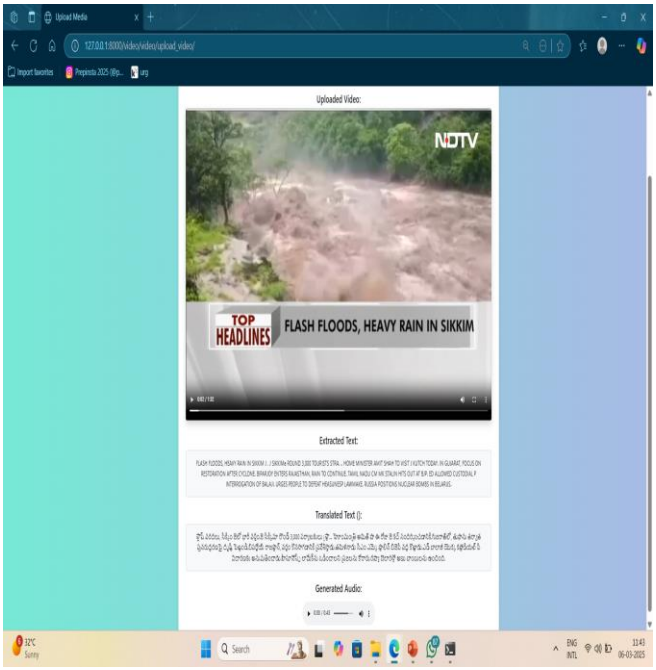

Fig 4: Video Processing and Extracted Text Output

The handling of a video file with text retrieved from video frames is shown in Fig 4. The extracted material is shown beneath the video once the technology correctly identifies and extracts language from the news segment. An audio output of the translated text is produced for playback, and the translated text is shown in the chosen target language. For users who are visually handicapped or non-native speakers, this feature enhances accessibility. The technology also makes sure that the extracted material is correctly structured and processed prior to translation, which improves the translated content's readability and usefulness. The system produces high-quality outputs with few mistakes by combining natural language processing (NLP) for translation and deep learning models for text recognition. Users may more easily absorb information in the language they want thanks to the produced audio, which further improves accessibility.

VI CONCLUSION

The process of extracting text from video material, translating it into many languages, and producing the accompanying audio output is all successfully automated by the suggested Video Text Extraction & Translator system. The system guarantees excellent accuracy in text detection and translation by utilizing cutting-edge deep learning models including BERT, BiLSTM, and GRU in conjunction with OCR and NLP approaches. A stable and scalable backend is made possible by the Django REST Framework (DRF) implementation, which enables smooth interaction with intuitive user interfaces. By removing language barriers, helping people with impairments, and facilitating cross-lingual content interpretation, this method greatly improves accessibility. The outcomes show how well the suggested method processes actual video footage, which makes it a useful tool for international communication, education, and media translation.

VII FUTURE SCOPE

There are several methods to improve the system's usability and performance. Real-time processing capabilities that enable users to extract and interpret text from live video broadcasts might be one of the future enhancements. Furthermore, the accuracy and fluency of translations may be enhanced by including multilingual voice synthesis and more sophisticated AI models. The system will become more accessible for a wider range of users by adding support for other languages and regional dialects. A more efficient workflow and improved interactivity may be achieved with further user interface (UI) and user experience (UX) enhancements. Finally, the system's practical usefulness and accessibility might be greatly improved on a broader scale by integrating it into already-existing media platforms or launching it as a mobile application.

REFERENCES

[1]S. Long et al., "Scene Text Detection with Deep Learning," *IEEE Transactions on Pattern Analysis*, 2019.

[2]A. Gupta et al., "Real-Time Text Extraction from Videos," *International Journal of Computer Vision*, 2021. Google, "Google Translate API Documentation," 2023.

[3]Baek et al., "STR Network for Scene Text Recognition," *IEEE Transactions on Image Processing*, 2019. Gupta et al., "OCR-Based Machine Translation," *International Journal of AI Research*, 2021.

[4]M. Liao et al., "Real-Time Scene Text Detection Using EAST," *Proceedings of IEEE CVPR*, 2018.

[5]Y. Shi et al., "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification," *IEEE Transactions on Pattern Analysis*, 2019.

[6]P. Wang et al., "Robust Text Detection and Recognition for Videos," *IEEE Transactions on Multimedia*, 2020.

[7]Tesseract OCR Team, "Tesseract OCR Documentation," *Google Developers*, 2023.

[8]K. Nayak et al., "A Comparative Study of OCR Models for Text Detection in Natural Scenes," *Journal of AI Research*, 2022.

[9]X. Zhang et al., "Automatic Text Extraction and Translation for Multilingual Videos," *International Conference on Pattern Recognition (ICPR)*, 2020. P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in ICASSP 2021 - 2021 IEEE InternationalConference on Acoustics, Speech and Signal Processing (ICASSP), pp.7613–7617, 2021.

[10]B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6319–6323, 2020.

[11]Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song, "Hearing lips: Improving lip reading by distilling speech recognizers," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6917–6924, 2020.

[12]S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1–8, 2019.

[13]T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," vol. 44, pp. 8717–8727, 2022.

[14]M. Faisal and S. Manzoor, "Deep learning for lip reading using audiovisual information for urdu language. arxiv 2018,"

[15]M. Jethanandani and D. Tang, "Adversarial attacks against lipnet: Endto-end sentence level lipreading," in 2020 IEEE Security and Privacy Workshops (SPW), pp. 15–19, 2020.

[16]J. Ting, C. Song, H. Huang, and T. Tian, "A comprehensive dataset for machine-learning-based lip-reading algorithm," vol. 199, pp. 1444–1449, 2022. The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 2021): Developing Global Digital Economy after COVID-19.

[17]W. Dweik, S. Altorman, and S. Ashour, "Read my lips: Artificial intelligence word-level arabic

lipreading system," vol. 23, pp. 1–12, Elsevier, 2022.
[18] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 356–363, 2020.
[19] Z. Su, S. Fang, and J. Rekimoto, "Liplearner: Customizable silent speech interactions on mobile devices," in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, (New York, NY, USA), Association for Computing Machinery, 2023.
[20] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023.