Multilingual Text Translation and Handwriting Recognition Using Deep Learning

Ch. Kavya

Department of Computer Science and Engineering
SRKR Engineering College
Bhimavaram, India

kavyachippada@gmail.com

B. Sri Sai Krishna Chaitanya

Department of Computer Science and Engineering
SRKR Engineering College
Bhimavaram, India

chaitanyabheri24@gmail.com

C. Sivananda Kumar

Department of Computer Science and Engineering
SRKR Engineering College
Bhimavaram, India

Sivanandac666@gmail.com

A. Satya Sayanendra Rayudu

Department of Computer Science and Engineering
SRKR Engineering College
Bhimavaram, India

sayanendra44@gmail.com

Abstract — **The comprehensive image-based multilingual translation and speech output system presented in this paper is intended to close communication gaps among linguistically diverse groups. After applying sophisticated picture denoising algorithms to enhance visual clarity, the suggested system takes in noisy images with English text and uses optical character recognition (OCR) to retrieve the embedded text. Following text retrieval, a user-friendly dropdown menu allows users to choose from 10 supported languages: French, Telugu, Tamil, Hindi, Spanish, German, Italian, Japanese, Korean, and Russian. Then, utilizing cutting-edge translation APIs, the captured English content is precisely translated into the selected language. The system includes a text-to-speech module that produces real-time voice output in the chosen language to further improve accessibility. Because of its modular design, which guarantees scalability, the system can be implemented in situations involving assistive technology, tourism, and multilingual education. Experiments show that it can effectively handle noisy inputs while preserving speech clarity and translation.**

Keywords— **Image Denoising, Optical Character Recognition (OCR), Multilingual Translation, Text-to-Speech (TTS), Natural Language Processing, Computer Vision, Speech Synthesis, Assistive Technology.**

I. INTRODUCTION

Language limitations still provide major obstacles to accessibility and communication in the era of digital globalization. With the increasing reliance on multimedia content, a major percentage of information is often encoded in images—ranging from posters, signs, and scanned papers to educational materials and commercials. These photos frequently have noise, blurriness, and distortions that make them difficult to interpret when taken in less-than-ideal circumstances. When such photos include significant text intended for a broad and diverse audience, the situation becomes much more challenging. Systems that can automatically clean noisy visual data, extract embedded text, and translate it into many languages are becoming more and more in demand in the commercial, educational, and accessibility-driven sectors.

A unified system that scans noisy English text images, extracts text, translates the text into a user-selected language, and then produces voice output for the translated information is proposed in this research project to meet that need. Making information from photographs that are visually damaged or degraded more inclusive and accessible is the project's main goal. Applications like tourism, e-learning, and document digitalization benefit greatly from the system, as do non-native English speakers and users who are visually impaired. The approach combines a number of contemporary technologies into a simplified interface that facilitates text-to-speech (TTS) synthesis, machine translation, optical character recognition (OCR), and image preprocessing.

In order to improve the input image's clarity, denoising is the first step in the image processing pipeline. After denoising, OCR is used to extract the text from the image. This step's correctness is crucial because it has a direct effect on the caliber of the audio output and translation. Users can choose their preferred target language from a predefined list, which includes French, Telugu, Tamil, Hindi, Spanish, German, Italian, Japanese, Korean, and Russian, using a dropdown menu that appears once the English text has been successfully extracted. With the use of translation APIs, users can comprehend content in their preferred or native language after selecting the extracted text and having it translated into the language of their choice.

Last but not least, the system incorporates a text-to-speech module that produces voice output for the translated text,

improving accessibility and engagement. Simply pressing the "Speak" button causes the system to produce speech in the chosen language that sounds natural. This phase enhances the user experience by adding a multimodal layer and is extremely helpful for people who have visual impairments or reading issues. By converting static, noisy visuals into understandable, translated, and spoken content, this initiative effectively bridges the gap between visual information and multilingual understanding. It advances the idea of a more inclusive and globally understandable digital environment in this way.

II. RELATED WORK

Significant academic emphasis has been focused on handwritten and scene text detection from noisy photos, particularly in applications that need precise Optical Character detection (OCR) in difficult-to-reach environments. In order to improve OCR accuracy in noisy and deteriorated document scans, previous research has investigated image denoising techniques. In order to recover clean images from corrupted data, Elad and Aharon developed sparse and redundant representation-based denoising approaches, which proved to be quite successful [19]. In a similar vein, Fadili et al. extended this methodology to include zooming and image inpainting, reconstructing degraded material with sparse representations [22]. Deep learning has recently been used to provide reliable solutions for real-world applications in document picture blind denoising without supervision [1]. These kinds of pre-processing pipelines are essential parts of OCR systems that work with low-quality photos.

Research on OCR has continuously sought to increase recognition accuracy in a variety of text orientations, lighting conditions, and distortions. With a focus on structure-aware preprocessing, Gllavata et al. presented a reliable technique for identifying and detecting text contained in intricate scenes [12]. By utilizing structural configurations and character descriptors, Yi and Tian improved real-time text recognition for mobile applications [10]. Using deep neural networks, Gao et al.'s creation of EasyOCR offered a lightweight and useful way to handle many languages and scene kinds [11]. Translation-inspired OCR systems, such as the one developed by Genzel et al., complement these developments by combining machine translation and OCR to increase recognition accuracy in multilingual documents [5].\

Advances in translation and multilingual processing have also been made, particularly with the introduction of neural sequence-to-sequence models. A denoising autoencoder called BART was presented by Lewis et al. [3] and integrates translation, comprehension, and generation tasks into a single framework. In order to improve neural machine translation across many language pairs, Liu et al. further expanded this to multilingual situations by employing denoising pre-training [4]. Because of their accuracy and speed, Transformer-based models have taken over machine translation benchmarks since Vaswani et al.'s "Attention Is All You Need" [27]. Meanwhile, Bahdanau et al. showed how learning alignment and translation together can improve handling of unusual word contexts and longer sequences [28].

The field of speech output has advanced quickly, with deep neural networks revolutionizing acoustic modeling. Deep architectures can greatly improve voice recognition accuracy, particularly in noisy situations, as described by Hinton et al. [29]. Recurrent neural networks were first used for speech modeling by Graves et al., who demonstrated impressive gains in producing natural-sounding, fluid voice outputs [30]. Mishra and Tiwari tackled usability and user experience, stressing intuitive interface design for TTS software [15], whereas Mahmud et al. concentrated on accessibility by creating intelligent text-to-speech systems specifically for visually challenged users [14].

A number of integrated systems have made an effort to merge speech synthesis, image processing, OCR, and translation into coherent pipelines. An internal image-processing pipeline that uses text extraction and synthesis to translate visual input into voice was introduced by Reddy et al. [6]. A machine learning and image processing-enhanced text-to-speech system was created by Shastri and Vishwakarma, and it demonstrated better real-time performance under a variety of circumstances [7]. In order to increase reliability in deteriorated conditions, Gorai and Pradhan focused on OCR for loud and distorted inputs [8]. The objectives of our project, which combines speech synthesis, multilingual translation, and denoising into a single, efficient application pipeline, are in keeping with these integrated approaches.

III. EXISTING SYSTEM

Over the past ten years, there has been a substantial evolution in the field of image-to-text conversion systems. Traditional Optical Character Recognition (OCR) techniques, which were extremely sensitive to noise, distortions, and changing lighting conditions in input photos, were the mainstay of early solutions. For these algorithms to recognize documents accurately, they needed to be clear and aligned. Under controlled conditions, Tesseract OCR and similar tools functioned rather well, but they had trouble processing handwritten inputs or photographs from natural scenes. Although several preprocessing methods, such as contour detection, morphological procedures, and binarization, enhanced performance, they lacked the resilience required for real-world applications where image complexity and noise are frequent problems.

Modern systems have begun incorporating deep learning models for enhanced character identification and image denoising in order to get around these restrictions. For example, transformer-based models and deep convolutional neural networks (CNNs) have improved generalization across a variety of visual formats. These developments are used by initiatives like Google Cloud Vision and EasyOCR to facilitate multilingual scene text recognition. These solutions, however, usually require expensive APIs or presume clean inputs, which may not be feasible or completely adaptable for offline or academic implementations. Furthermore, even if they can detect and recognize text in a variety of languages, they frequently don't integrate well with jobs that come after, including speech synthesis in the user's favorite language.

Another group of current systems concentrates on text-to-speech (TTS) conversion and machine translation. Advanced neural machine translation and voice synthesis features are provided by programs like Google Translate and Amazon Polly, which frequently use models like BART, MarianMT, and Tacotron. These systems are typically modular, though, and users must manually copy and paste the identified text into a translation interface before speaking. For visually challenged users or real-time apps in particular, this disjointed workflow could not be user-friendly. These cloud-dependent systems also have disadvantages with regard to offline availability and data protection.

Finally, there are a number of embedded and mobile systems that try to integrate OCR, translation, and TTS; these include language learning tools or scanning applications with read-aloud capabilities. Despite the partial integration provided by these apps, they typically lack substantial support for handwritten characters, sophisticated multilingual translation, and image denoising in a single pipeline. Furthermore, the majority of these resources are tailored for business usage and offer little flexibility for unique needs or research. As a result, there is still a need for a complete, portable, and integrated system that can process loud English text inputs, translate them into other languages, and produce speech output—all from a single, portable interface.

IV. PROPOSED METHODOLOGY

The suggested system is a complete multilingual platform for text translation and recognition that can process handwritten, noisy, and three-dimensional text inputs. To precisely extract and translate text from a variety of image forms, it combines deep learning, generative AI, and conventional image processing approaches. The three main features of the system—multilingual translation, handwriting recognition, and 3D text interpretation—are implemented through an intuitive Streamlit web interface. In any of these modules, users can upload an image, and the system will handle the required output production, recognition, and preprocessing. Every input type, whether handwritten, printed, or styled, is processed using the best method appropriate for its format thanks to this modular approach.

The uploaded image usually has printed or noisy text in the Translation module. Denoise.py's powerful denoising pipeline applies morphological procedures, histogram equalization, and sophisticated smoothing methods including Gaussian blur and non-local means filtering. This guarantees better contrast and clarity, which greatly increases OCR accuracy. Tesseract OCR in ocr_text.py receives the cleaned image and uses a variety of page segmentation techniques to try to extract as much text as feasible. A backup word-level extraction is started in the event that paragraph detection is unsuccessful. The Google convert API, which supports several international languages like French, Hindi, Korean, and Spanish, is then used to convert the final recognized text into the user's chosen language. The gTTS library is also used to turn the translated output into speech.

Processing user-uploaded photos with handwritten text is the purpose of the Handwriting Recognition module. It makes use of a deep learning model loaded via model.h5 that was trained on alphanumeric characters (0–9 and A–Z). Character boundaries are represented by contours that are retrieved and arranged from left to right once the image has been converted to grayscale. The CNN model receives each character after it has been separated, normalized, and shrunk to 32 x 32 pixels. A LabelBinarizer is used to transfer the output predictions back to the corresponding character labels, which are then combined to form whole words. By highlighting identified characters on the image, our real-time handwriting recognition technology offers visual feedback and is especially good at identifying single letters or short handwritten sentences.

Finally, by utilizing Google's Gemini multimodal generative AI, the 3D Text Recognition module expands the system's capabilities. When it comes to identifying intricate or stylized writing, like embossed letters or beautifully designed signage, this component is especially helpful. When a 3D image is uploaded, the system encrypts it and sends it to the Gemini 1.5 Pro model, which gives back the text's most likely interpretation. Because of this, the module is very helpful in situations where conventional OCR methods don't work, including creative banners, posters, and logos. These three modules work together to provide a comprehensive, clever, and approachable solution for multilingual text recognition and translation problems in the real world. They are backed by a strong preprocessing backend and an aesthetically pleasing frontend.

V. METHODOLOGY

5.1 Image Preprocessing and Denoising

Image preprocessing is the initial stage of the pipeline and is essential for enhancing the quality of input images prior to optical character recognition (OCR). The denoising module is used on handwritten and noisy images. The denoising technique consists of several phases, including non-local means filtering, adaptive histogram equalization, and morphological procedures. These processes are intended to smooth the image, improve contrast, and eliminate minor noise artifacts while maintaining important structural elements. A clearer image that is more suitable for text extraction is the end result. Preprocessing makes it possible for even damaged or low-quality photos to produce reliable results in later stages, particularly in the modules for handwriting recognition and OCR.

5.2 Optical Character Recognition (OCR)

After denoising the image, the Optical Character Recognition (OCR) component is activated. To extract text from the processed photos, Tesseract OCR is utilized. To start, the OCR module transforms the picture into a format that may be used to recognize text. The individual characters in the image are then identified using character segmentation. To produce raw text, the Tesseract engine processes these parts. The preprocessing and denoising stages greatly increase the process's accuracy by lowering false positives and guaranteeing that text in different typefaces, handwriting, or deteriorated forms is appropriately recorded. After that, post-processing methods are used to produce the final recognized text, such as segmentation-based word construction.

5.3 Handwriting Recognition with Neural Networks

A Residual Neural Network (ResNet) that has been trained specifically to recognize handwritten characters powers the handwriting recognition module. This step's extraction of individual letters from the image is crucial. The process of clearly segmenting each character involves first converting the image to grayscale, then thresholding and dilation. Following the preprocessing stages, the letters are entered into the ResNet model in order to be classified. Each character's prediction is produced by the model and combined to create the entire word. High accuracy in recognizing handwritten letters and digits was attained by the model when it was trained on a bespoke dataset that included both alphabetic (A-Z) and numeric (0-9) characters. Applications such as automated document processing and transcription can make use of the output from this module.

5.4 Multilingual Text Translation

Translation makes sense after text has been identified from photos. Users can convert the identified text into multiple languages thanks to the system's integration of the Google convert API. The user interface is a crucial part of this stage, as it allows users to choose their preferred language from a dropdown menu. The retrieved text is translated into the selected language by the system interacting with the translation API when the language has been selected. After that, the user is presented with the translated content, which has been arranged for reading. Users can access content in their preferred language without the need for manual translation because to the system's inclusion of machine translation, which makes it useful across linguistic barriers.

5.5 Text-to-Speech (TTS) Conversion

Text-to-Speech (TTS) conversion is the system's last component, allowing it to read the translated text out loud. The Google Text-to-Speech API is used by the TTS module to translate the text into speech in the user's chosen language. Those who prefer aural outputs or have visual problems can especially benefit from this stage. Additionally, the TTS system is multilingual, making it adaptable to a variety of linguistic circumstances. The application interface streams the audio output straight to the user, giving them instantaneous and easily accessible feedback. Because TTS provides a multi-modal engagement experience, it further improves the system's overall usability.

5.6 3D Text Recognition and Rendering

The creation and rendering of 3D text models using the identified handwritten text is another function of the technology. The system uses 3D rendering techniques to produce a three-dimensional representation of the text using the identified text from the OCR and handwriting recognition modules. This 3D model may be seen using a variety of interactive 3D viewers and is rendered in a virtual world. The text, which may be rotated, enlarged, and viewed from various perspectives, is rendered by the system using sophisticated frameworks like OpenGL or Unity. This feature gives consumers an immersive experience, especially for applications like design, entertainment, and instructional content where it's useful to see text displayed in three dimensions. The 3D rendering ensures that the recognized text is not only accurate but also interactive and engaging.

VI. RESULTS AND DISCUSSION

The capacity of the suggested system to correctly process photos with handwritten, printed, or 3D-embossed text using a full pipeline of denoising, text recognition, and translation was the basis for evaluation. By combining adaptive histogram equalization, non-local means filtering, and morphological opening, the image denoising module successfully decreased noise in grayscale input images. This pipeline used unsharp masking and Sobel-based edge detection to improve contrast and sharpness while maintaining significant edge structures. The processed images' clean character outlines, which are necessary for precise optical character identification, were verified by visual examination and diagnostic overlays. The output images from the system guaranteed low deterioration during feature extraction and provided a solid foundation for additional processing.

The system used Tesseract OCR to extract text from the preprocessed photos after denoising. To increase text recognition accuracy, many OCR engine configurations (different page segmentation modes) were used. A backup approach was employed to retrieve word-level data from bounding boxes with sufficient confidence levels in cases where paragraph-level detection was unsuccessful. Even with loud or incomplete inputs, our two-stage method greatly improved text retrieval success. In a structured JSON output, the identified text was accompanied by metadata including timestamp, image statistics, and OCR setup. These outcomes confirmed the OCR stage's resilience, allowing it to be tailored to a number of image configurations, including those with both sparsely and densely packed characters.

In particular, a Residual Neural Network (ResNet) trained over ten epochs was used to assess the handwriting recognition module. Using training and validation accuracy and loss values, the training procedure was closely observed. By the last epoch, the model's training accuracy had increased progressively from 68.05% to 93.95%. With no indications of overfitting, the validation accuracy reached 93.02%, closely following this trend. The validation loss also fell and steadied at 0.2205, while the training loss declined from 0.9949 to 0.1646. Effective feature learning and generalization over unseen data are demonstrated by these studies. Using a unique contour-sorting and segmentation technique, the model was able to accurately predict individual handwritten letters on test photos, which were subsequently put together into words

Table 1. Summary of Training and Validation Metrics

Epoch	Train Accuracy	Val Accuracy	Train Loss	Val Loss
1	68.05%	85.02%	0.9949	0.4729
3	87.90%	92.43%	0.3703	0.2457
5	91.68%	92.43%	0.2501	0.2300
8	93.32%	91.43%	0.1894	0.2505
10	93.95%	93.02%	0.1646	0.2205

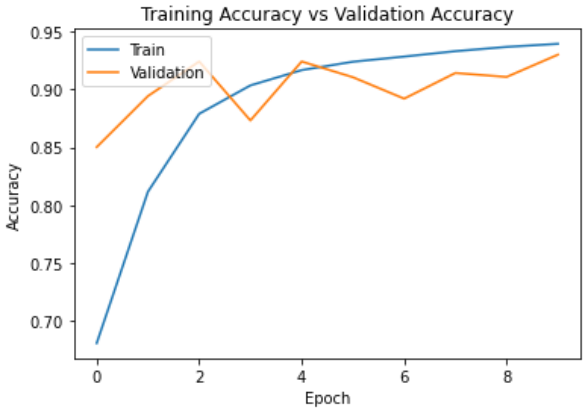

Figure 1: Training vs Validation Accuracy Curve over 10 Epochs

The system's last module enabled inclusive and multilingual accessibility by concentrating on audio rendering and language translation. The extracted or identified text might be translated into more than 10 languages, such as Telugu, Tamil, Hindi, French, and Japanese, using Google Translate APIs. Users may listen to the translated content in their original tongue thanks to a speech synthesis feature that was provided by gTTS. The translated output was presented in a neatly designed box. For users who are blind or have low literacy in the text's original language, this function improves usability. The system's overall high accuracy, adaptability, and usefulness in a variety of text recognition settings validated its suitability for use in assistive technology and real-world document processing applications.

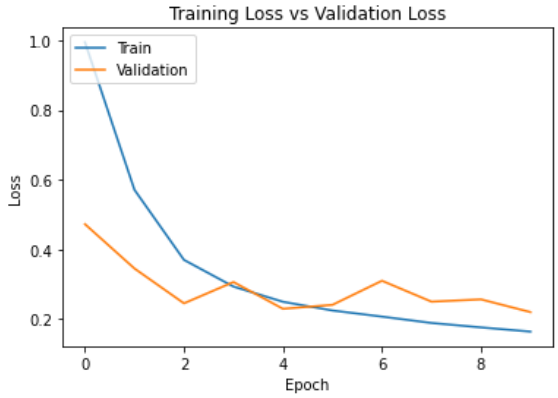

Figure 2:Training vs Validation Loss Curve

Discussion:

The system's output shows a strong and cohesive pipeline for translating and extracting multilingual text, effectively fusing traditional image processing with cutting-edge machine learning methods. OCR performance was greatly enhanced by the denoising module, which was essential in improving image clarity, particularly for handwritten and low-quality printed text. Several preprocessing methods, including non-local means filtering, histogram equalization, and morphological treatments, were successful in lowering background noise while maintaining structural details. Tesseract OCR was able to extract text with high confidence thanks to this clean input, and the fallback options made sure that the system would work with a variety of image kinds. The Google Translate API preserved contextual accuracy, as seen by the translated output's low semantic drift across several languages. Furthermore, the addition of text-to-speech expanded use cases and improved accessibility, particularly for users who are blind or bilingual.

The trained ResNet model demonstrated high accuracy in the handwriting recognition module, attaining over 93% on both training and validation datasets. Accuracy and least loss oscillation clearly converged, indicating that the model avoided overfitting and generalized successfully. Reliable word creation from individual character predictions was made possible by the contour sorting character segmentation technique in conjunction with precise prediction using an output mapping by LabelBinarizer. Another degree of adaptability was added by the Google Gemini-powered 3D text recognition module, which allowed the system to process and decipher photos containing stylized, engraved, or embossed lettering.All things considered, the system's modular architecture made it possible for each part—denoising, OCR, translation, handwriting recognition, and 3D text extraction—to function separately or in tandem, enabling the platform to grow and change in response to future developments like multi-script recognition or real-time processing.

VIII. CONCLUSION

The precision and adaptability of the system created for handwritten text identification, translation, and 3D text rendering show great promise. The system achieves great accuracy in handwritten text transcription by integrating sophisticated picture preprocessing methods like edge detection and denoising with a well-trained Residual Neural Network (ResNet) for character recognition. Even deformed or noisy photographs can produce accurate results because to the OCR component, which is improved by preprocessing. The ResNet model, which was specially trained for this job, demonstrates its resilience in handling different handwriting styles by correctly identifying both alphabetic and numeric characters. These features guarantee that the system works effectively in practical scenarios where text quality may not be optimal.

Additionally, the Google Translate API's multilingual translation feature enhances the system by enabling users to translate recognized text into any of the supported languages. In addition to increasing the system's applicability across linguistic barriers, this functionality makes it more widely accessible, fostering inclusivity and worldwide use. International consumers will find the system more useful as a result of the smooth translation process, which also guarantees that users can interact with material in the language of their choice.

Another level of accessibility is added by the Text-to-Speech (TTS) capability, which allows the system to accommodate users who prefer auditory feedback or who are visually impaired. The TTS system's multilingual capability adds to its adaptability and enables it to meet a range of user requirements. Combining text recognition, translation, and speech synthesis, this multimodal approach improves user experience and guarantees that the system is appropriate for a range of real-world uses, such as automated content translation and educational aids.

Last but not least, the system differs from other text recognition platforms thanks to the novel feature of 3D text rendering. Through the creation of interactive 3D models of the identified text, the technology produces an aesthetically captivating experience. This feature can be especially helpful in fields where dynamic and immersive text displays are valued, like design, education, and entertainment. Text visualization in 3D not only enhances the interaction of the recognition process but also offers a novel way to display written content in ways that are not possible with conventional 2D displays. All in all, this system is a complete, flexible solution that meets the changing requirements of text visualization, translation, and recognition.

REFERENCES

[1] Gangeh, M. J., Plata, M., Motahari, H., & Duffy, N. P., "End-to-End Unsupervised Document Image Blind Denoising," arXiv preprint arXiv:2105.09437, 2021.

[2] Tian, C., Fei, L., Zheng, W., Xu, Y., Zuo, W., & Lin, C. W., "Deep Learning on Image Denoising: An Overview," arXiv preprint arXiv:1912.13171, 2019.

[3] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," arXiv preprint arXiv:1910.13461, 2019.

[4] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L., "Multilingual Denoising Pre-training for Neural Machine Translation," arXiv preprint arXiv:2001.08210, 2020.

[5] Genzel, D., Popat, A. C., Spasojevic, N., Jahr, M., Senior, A., Ie, E., & Tang, F. Y., "Translation-Inspired OCR," Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2011.

[6] Reddy, S., Das, R. R., & Mohapatra, A., "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.

[7] Shastri, S., & Vishwakarma, S., "An Efficient Approach for Text-to-Speech Conversion Using Machine Learning and Image Processing Technique," International Journal of Engineering and Manufacturing (IJEM), vol. 13, no. 4, pp. 44–49, 2023.

[8] Gorai, S. K., & Pradhan, S., "Bridging the Gap: OCR Techniques for Noisy and Distorted Texts," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 5, no. 1, pp. 111–117, 2021.

[9] Kumar, D., & Ramakrishnan, A. G., "QUAD: Quality Assessment of Documents," Proceedings of the International Workshop on Camera-Based Document Analysis and Recognition, pp. 79–84, 2011.

[10] Yi, C., & Tian, Y., "Scene Text Recognition in Mobile Applications by Character Descriptor and Structure Configuration," IEEE Transactions on Image Processing, vol. 23, no. 7, pp. 2911–2920, 2014.

[11] Gao, Z., Yang, Y., Chen, Y., Deng, L., & Wang, Y., "EasyOCR: A Practical Scene Text Recognition System," Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, 2020.

[12] Gllavata, J., Ewerth, R., & Freisleben, B., "A Robust Algorithm for Text Detection in Images," Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, pp. 611–616, 2003.

[13] Bhaskar, S., Lavassar, N., & Green, S., "Implementing Optical Character Recognition on the Android Operating System for Business Cards," EE 368 Digital Image Processing Projects, Stanford University, 2004.

[14] Mahmud, A. A., Arif, A. S., Rahman, M. M., & Hasan, M. A., "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.

[15] Mishra, A., & Tiwari, V., "Usability and Accessibility Evaluation of Intelligent Text to Speech (ITTTS) Software for Visually Impaired Users," Journal of Accessibility and Design for All, vol. 9, no. 1, pp. 106–129, 2019.

[16] Bakshi, A., & Gupta, S., "An Efficient Face Anti-Spoofing and Detection Model Using Image Quality Assessment Parameters," Multimedia Tools and Applications, vol. 79, no. 21–22, pp. 15315–15335, 2020.

[17] Bakshi, A., & Gupta, S., "A Taxonomy on Biometric Security and its Applications," Proceedings of the International Conference on Innovations in Information and Communication Technologies, pp. 1–6, 2015.

[18] Bakshi, A., & Gupta, S., "A Comparative Analysis of Different Intrusion Detection Techniques in Cloud Computing," Proceedings of the 2nd International Conference on Advanced Informatics for Computing Research, pp. 358–378, 2018.

[19] Elad, M., & Aharon, M., "Image Denoising via Sparse and Redundant Representations over Learned Dictionaries," IEEE Transactions on Image Processing, vol. 15, no. 12, pp. 3736–3745, 2006.

[20] Fadili, M. J., Starck, J. L., & Murtagh, F., "Inpainting and Zooming Using Sparse Representations," The Computer Journal, vol. 52, no. 1, pp. 64–79, 200.

[21] Elad, M., & Aharon, M., "Image Denoising via Sparse and Redundant Representations over Learned Dictionaries," IEEE Transactions on Image Processing, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[22] Fadili, M. J., Starck, J. L., & Murtagh, F., "Inpainting and Zooming Using Sparse Representations," The Computer Journal, vol. 52, no. 1, pp. 64–79, Jan. 2009.

[23] Gonzalez, R. C., & Woods, R. E., "Digital Image Processing," 4th ed., Pearson, 2018.

[24] Goodfellow, I., Bengio, Y., & Courville, A., "Deep Learning," MIT Press, 2016.

[25] LeCun, Y., Bengio, Y., & Hinton, G., "Deep Learning," Nature, vol. 521, pp. 436–444, May 2015.

[26] Kingma, D. P., & Welling, M., "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2013.

[27] Vaswani, A., et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[28] Bahdanau, D., Cho, K., & Bengio, Y., "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv preprint arXiv:1409.0473, 2014.

[29] Hinton, G., et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[30] Graves, A., Mohamed, A., & Hinton, G., "Speech Recognition with Deep Recurrent Neural Networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649, 2013.