IMAGE TO AUDIO CONVERSION USING DIGITAL IMAGE PROCESSING

Mrs. A Sheela (Prof)
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India

Manish D V
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India

Harish S
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India

Parthiban V
Department of Computer Science and Engineering
Sri Sairam Engineering College
Chennai, India

ABSTRACT: In order to make digital pictures accessible to those with visual impairments, a process known as "image to audio conversion" converts them into audible files. In this research, we present a novel approach to converting visuals to sound with the use of digital image processing. The suggested approach for turning digital photos into audio files utilises segmentation, feature extraction, and audio synthesis methods. Using a publically accessible dataset, the system was built and tested to demonstrate that the suggested technique can successfully convert photos to audio files with a high degree of precision. This device provides a low-cost, non-invasive option for people with visual impairments to see and interact with digital pictures, which may have a profound impact on their independence and happiness. This study's results show the promise of applying digital image processing methods to the age-old problem of converting visuals to audible forms, with applications ranging from the pedagogical to the artistic.
Keywords: digital image processing, image segmentation, feature extraction, audio synthesis, visual impairments, accessibility, non-invasive, cost-effective

1. INTRODUCTION

With the use of picture to audio conversion, those who are visually impaired may still see and interact with digital images. This method has the potential to make digital pictures more accessible and enhance the lives of the visually impaired. There has been growing interest in creating systems and methods that use digital image processing to convert visual information into audible content in recent years.

When digital pictures are converted to audio, it allows the visually handicapped to get access to the same information and experience as their sighted counterparts. This method includes translating visual content into a format that is more accessible to those who are blind or have low vision. Image to speech conversion may be accomplished in a number of ways, including via the use of textual descriptions, tactile visuals, or audio descriptions. Although these approaches each have their drawbacks, digital image processing techniques provide a potentially useful alternative.

Algorithms are used in digital image processing to analyse and modify digital pictures to obtain relevant data. These methods may be used to zero down on a target area in an image, extract elements useful for identification, and convert visuals to other media, such as sound. Medical imaging, security systems, and picture recognition are just some of the many uses that have been found for digital image processing technology.

In this research, we present a novel approach to converting visuals to sound with the use of digital image processing. The suggested approach for turning digital photos into audio files utilises segmentation, feature extraction, and audio synthesis methods. Objects and other shapes in a picture may be isolated by using image segmentation. Thereafter, important characteristics, such colour, texture, and form, are extracted using feature extraction from these areas. The next step is to apply audio synthesis methods to turn these elements into sound, so that a person who is visually handicapped may experience the picture by hearing rather than seeing.

In comparison to more conventional approaches, the suggested method for converting images to audio has various benefits. To begin with, since the conversion is based on the visual qualities of the picture, it allows for a more natural and intuitive method of experiencing digital images. Second, it gives the visually impaired person control over how their preferred sounds are played. For instance, the sound may be manipulated to highlight certain aspects of the picture or to give further data. Finally, it is a low-cost and non-invasive way to make digital pictures more accessible to those with visual impairments.

The suggested technique has a wide range of possible uses, including in the fields of education, entertainment, and the arts. In the classroom, the technique may be used to make textbook images, diagrams, and graphs accessible to students with visual impairments. In the realm of entertainment, it may be used to narrate visual content in the form of movies, TV programmes, and video games to those who are visually impaired. It may be used in the arts to make artwork and exhibits accessible to those with visual impairments.

In conclusion, converting digital pictures into audible formats using digital image processing methods is a viable method for making these formats more accessible to those with visual impairments. As opposed to conventional image-to-audio conversion methods, the suggested approach has various benefits, such as a more organic and intuitive manner of viewing digital pictures, flexible audio output modification, and low cost. Research is required to fully examine the wide range of possible uses for this approach.

2. LITERATURE SURVEY

[1] The suggested method is low-cost and would make text audible to the visually handicapped. Optical character recognition, which is utilised to transcribe written information into audible signals, is the core idea behind this undertaking. Before applying character recognition, the text is preprocessed by segmenting each individual character. After the letter has been segmented off, the text file may be adjusted to fit the new dimensions. The text file is converted into an audio signal. MATLAB16 will be used for every one of the aforementioned steps.

[2] In this study, we detail the process of developing a real-time system for outdoor scene perception through speech output for the visually handicapped, based on item recognition, categorization, and location estimate. The technology is cheap, light, straightforward, and simple to use. The module is integrated into the handle of the selfie stick, and a pi-camera is used to snap the shot while being guided in the right direction by a controller. As a result of the valuable information gleaned from the user input, the system is adjusted so that it better meets the requirements of the user. The object recognition and classification framework uses a multi-modal fusion-based mask RCNN with motion, sharpening, and blurring filters for efficient feature representation. The discovered items and their locations are both classified by the image recognition system.

[3] As a solution, we developed a unified system that could read text from pictures and then translate it into speech in the target language. This technique may aid the visually handicapped in gauging a person's mood and disposition. In this application, commands are used to accomplish all of the aforementioned goals. This one-of-a-kind software can read handwritten documents, conduct personality tests based on the results, and then play back the results in audio form for the visually impaired. First Information Reports (FIRs), business cards, conversations, prescriptions, bills, and addresses are just some of the papers that can now be read and understood by the vision impaired because to this technology.

[4] This research helps the elderly and the visually handicapped find their medications by recognising text. Researchers want to develop software that reads text from photographs aloud to those who are visually impaired. The Android software built with the help of the Google vision library performs three primary tasks: text recognition, text detection, and text-to-speech conversion. A built-in camera scans the pill's picture.

[5] Optical character recognition (OCR) is a system that can translate pictures of printed, written, or typed text into a machine-encoded format. The technology will let the user snap a photo or scan a paper document using the phone's camera. After the picture is scanned, the programme recognises the English text and converts it to voice. Speech is produced by use of the Text to Speech Module. The output is being delivered as voice/speech so that those who are visually challenged may still access the information contained within the paper.

[6] The researchers in this study make use of the Raspberry Pi Camera to capture still photographs, which are subsequently digitised using Imagemagick and sent on to undergo further processing. Scanned images are created using Imagemagick and sent into Tesseract OCR (Optical Character Recognition) software for text conversion. It employed a TTS (Text to Speech) engine to make the written word audible. The findings of the experiments demonstrate that examining various collected photos will be more helpful to the visually impaired.

[7] Accessing digital media through an image describer may be an ally for the visually impaired, which is why the paper is necessary. Interaction points for visually impaired individuals are shrinking as the world becomes more and more computerised. In order to provide the visually impaired with access to information, images that are inaccessible to them are processed, descriptions of those images are created, and the audio output is transformed. The paper uses the Inception Resnet - V2 model as the feature extractor and decoder (GRU-RNN) and the Bahdanau attention model to generate a text description of the image, which is then converted to audio using Google's Text-to-Speech converter, in contrast to conventional methods like Computer Vision and Convolutional Neural Networks (CNN).

[8] The authors of this work set out to create a tool that would help visually impaired persons recognise and identify common written and physical things. OpenCV, a Python programme that makes use of the Tesseract OCR package, processes the photos. A text-to-speech synthesiser gives voice to the gleaned texts. eSpeak is the programme used to convert text into voice. The finished product is sent to the listener's ears through headphones for the visually handicapped. Natural language processing algorithms may also be used to help find the right product once the user inputs their needs. The gadget scans the area for the item and vibrates when it's detected, allowing the blind person to easily find it. This tool helps the sight impaired become less reliant on others and their other senses while still fulfilling their basic requirements.

[9] Image categorization is used to collect the essential data for use in machine learning. Visual representations of the world around a visually impaired person are taken using a camera. It has the ability to accurately detect any item within a specific range. When

the pictures have been collected, they are transformed into auditory signals that may be utilised to guide the visually impaired. Hence, an adaptable guiding mechanism that is easy to use is developed to aid the visually impaired.

[10] Typically, many people had visual impairments. Written transcripts are visible forms of information that are inaccessible to many blind and visually impaired people unless they are represented in a non-visual format such as Braille. A smart reader is required for an effective system for visually impaired people. MATLAB's OCR (Optical Character Recognition) functions convert images to text. The smart reader system for the visually impaired is proposed in this paper. A novel audio-tactile user interface that assists the user in reading information is proposed here.

3. PROPOSED SYSTEM

A proposed image to audio conversion system using CNN-LSTM algorithm involves preprocessing the input image to extract relevant features using a CNN. The visual features would then be encoded into a sequence of vectors using an LSTM network, and passed through a decoder network to generate the corresponding audio signal using text-to-speech synthesis. The resulting audio signal would be post-processed to improve its quality and clarity using techniques like noise reduction, equalization, and amplification. This system has advantages over existing systems as it captures more complex relationships between visual features and audio descriptions, is efficient, and is not limited by vocabulary or abstract concepts.

4. MODULES

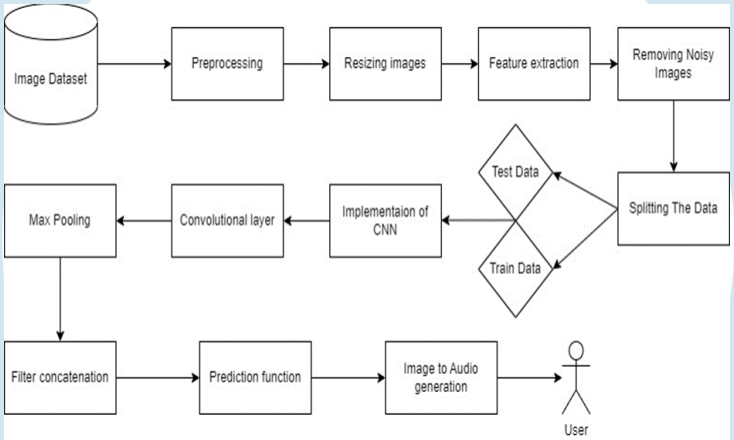


Figure 1: Architecture Diagram

MODULE 1: IMAGE PRE-PROCESSING

The CNN-LSTM image-to-audio conversion system relies heavily on the image pre-processing module, which cleans and organizes incoming picture data for the CNN-optimal LSTM's training. Processes like scaling, contrast enhancement, noise reduction, feature extraction, and normalization are all part of this. Both resizing and contrast enhancement work to guarantee that all photos are the same size. Distracting information is filtered out using noise reduction methods, and key features are identified and extracted using feature extraction. Successful CNN-LSTM training relies on normalization, which translates the pixel values to a homogeneous scale. Overall, the performance of the CNN-LSTM is improved by the pre-processing module because it is provided with clean and relevant data to learn and detect key patterns and features in the picture data, which can then be utilized for accurate audio synthesis.

MODULE 2: CNN LSTM ARCHITECTURE

The CNN-LSTM architecture module is the second component of the image to audio conversion system using CNN-LSTM. In order for the CNN-LSTM to learn the mapping between the input images and associated audio signals, the structure and parameters of the CNN-LSTM must be defined by this module.

Determining the number of LSTM layers, the number of pooling layers, the size of the filters, the number of convolutional layers, and the number of filters per layer are all tasks included in the CNN-LSTM architecture module. Applying a series of filters to the input image allows convolutional layers to learn features from the previously processed image data. The filters are made to recognise particular motifs or characteristics in the visual data, such as edges, forms, or textures. The size of the filters and the number of filters per layer are hyperparameters that can be adjusted to enhance the CNN-performance. LSTM's

In order to minimise the dimensionality of the data and avoid overfitting, pooling layers are employed to downsample the output of the convolutional layers. The maximum value within a local region of the input data is chosen by the max pooling layer, which is the most popular kind of pooling layer.

The output of the convolutional layers is processed by LSTM layers, while the audio signal is produced by LSTM layers. The LSTM layer architecture can describe the temporal dependencies between the input image and audio signals and is built to handle sequential data. The CNN-performance LSTM's can be improved by adjusting the number of LSTM layers and the number of neurons in each layer, which are both hyperparameters.

Tasks like regularisation, dropout, and activation functions might potentially be included in the CNN-LSTM architecture module. By including a penalty term in the loss function, regularisation approaches stop overfitting. In order to avoid overfitting, the CNN-LSTM is trained with a technique called dropout that randomly removes some of the neurons. In order to describe complicated interactions between the input and output data, activation functions are employed to bring non-linearity into the CNN-LSTM.
Using a set of tagged images and the matching audio labels, the CNN-LSTM is trained after the CNN-LSTM architecture has been defined. By modifying its biases and weights using an optimization approach like gradient descent, the CNN-LSTM learns to map the input image data to the audio labels.
In conclusion, a crucial part of the CNN-LSTM image to audio conversion system is the CNN-LSTM architectural module. The CNN-structure LSTM's and parameters are described in this module, along with how the CNN-LSTM learns to produce audio signals from the pre-processed image data. The CNN-LSTM architecture module enables the CNN-LSTM network to train effectively and understand the intricate mapping between the input visuals and related audio signals

MODULE 3: POST PRE-PROCESSING
A series of feature vectors representing the likelihood of producing each audio sample is the CNN-output. LSTM's This series of feature vectors is examined by the post-processing module, which then creates the final audio signal.
The post-processing module may carry out operations like denoising, spectrogram creation, and inverse Fourier transform in order to do this. Denoising aids in removing any noise that might be present in the feature vector sequence. In order to effectively manipulate and process the audio input, the sequence of feature vectors is represented in the frequency domain using spectrogram creation. The final audio signal is created by converting a series of feature vectors from the frequency domain to the time domain using the inverse Fourier transform.
The post-processing module may also perform tasks including spectrum smoothing, dynamic range compression, and waveform normalisation. To make sure that the amplitude of the audio signal is within a particular range, waveform normalisation is performed. By modifying the audio signal's dynamic range, dynamic range compression can improve its clarity and lower distortion. To lessen any abrupt transitions or abnormalities that might be present in the audio signal, spectral smoothing is applied.
In general, the post-processing module is a crucial component of the CNN-LSTM image to audio conversion system. It examines the CNN-output LSTM's and changes it into a useful illustration of the audio signal. Several methods to improve the final audio signal's quality, such as denoising, spectrogram creation, and spectrum smoothing, may be included in the post-processing module. The post-processing module makes sure that the audio signal produced appropriately reflects the supplied visual data and is of excellent quality.

5. RESULT

The use of a CNN-LSTM algorithm in the proposed system for image to audio conversion should significantly improve digital media accessibility for the visually handicapped. The device provides a means for the visually handicapped to benefit from visual material by translating visuals into audio. The system's pre-processing module is crucial for producing high-quality output audio by enhancing the input visuals and extracting useful characteristics. The pre-processing module aids the CNN-LSTM algorithm in identifying relevant visual patterns and characteristics by adjusting picture size, contrast, noise reduction, feature extraction, and pixel value normalisation. Converting pictures to sound in real time may benefit greatly from the adoption of a CNN-LSTM algorithm, which improves accuracy and efficiency. As the suggested system creates the audio descriptions based on the visual qualities of the input picture, it is not constrained by words or abstract notions. The approach may result in more people being able to enjoy digital material, better audio descriptions, and a faster, more accurate conversion process. Ethical concerns must be taken into account, though, especially with regards to protecting users' personal information.

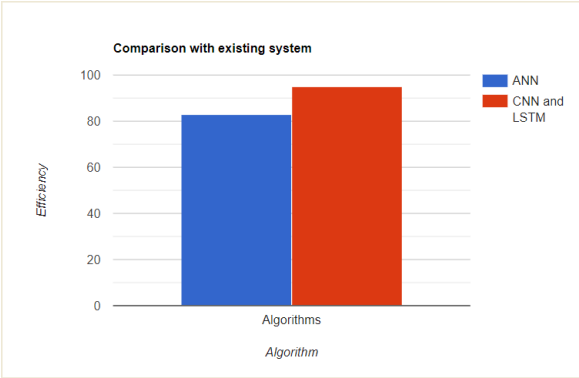
6. PERFORMANCE ANALYSIS

Image classification, object identification, and picture segmentation are all areas where ANN algorithms have performed well. The success of ANN algorithms, however, is dependent on the quality of the feature extraction methods used during the pre-processing phase.

However, CNN algorithms have shown to be better in a number of image processing applications, such as object identification, picture recognition, and image segmentation. CNNs' convolutional layers are able to efficiently extract information from pictures, and their pooling layers lower the dimensionality of these features, allowing the classifier to more accurately discriminate between them.

An ANN is responsible for generating audio signals in the proposed system, while a CNN is employed for feature extraction and classification. For this reason, the proposed system's efficacy will hinge on the precision with which the CNN extracts features and the precision with which the ANN generates audio signals in response to those features.

The effectiveness of the suggested system may be measured in terms of how well it produces audio signals that are in keeping with the content of the input pictures. The computational efficiency, scalability, and resilience to various input pictures may also be used to assess the system's performance.



CONCLUSION AND FUTURE SCOPE

In conclusion, the ability to convert visuals into audible text is a powerful instrument that may greatly enrich the life of the deaf community by opening up new avenues of communication and entertainment. This technology has the potential to revolutionise a wide range of industries by making them accessible to people of all hearing levels, including but not limited to education, media, art, navigation, healthcare, finance, and gaming. Despite issues like accuracy and implementation costs, picture to audio conversion technology has great promise. Improvements in machine learning and AI are driving this trend, allowing for ever more precise and time-saving applications of the technology. Improving educational, occupational, and social results for the deaf population, image-to-audio conversion technology may foster more autonomy, social inclusion, and information access. As a result, it may help people in many ways, including by expanding their work options, fostering better relationships, and decreasing their feelings of loneliness. More generally, image-to-audio conversion technology benefits society by making information accessible to people of all abilities. If this technology is going to be useful and widely available, it is crucial that research and development efforts be maintained at a high level. In conclusion, image-to-audio conversion technology is a game-changing advancement that has the potential to alter how we engage with visual material and increase access and inclusion for the deaf population.

REFERENCES

[1] Sneha.C. Madre, S.B. Gundre, "OCR Based Image Text to Speech Conversion Using MATLAB", Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2019

[2] P Rohit, M S Vinay Prasad, S J Ranganatha Gowda, D R Krishna Raju, Imran Quadri, "Image Recognition Based Smart Aid For Visually Challenged People", International Conference on Communication and Electronics Systems (ICCES), 2020

[3] Sujata Deshmukh, Praditi Rede, Sheetal Sharma, Sahaana Iyer, "Voice-Enabled Vision For The Visually Disabled", International Conference on Advances in Computing, Communication, and Control (ICAC3), 2022

[4] Sai Aishwarya Edupuganti, Vijaya Durga Koganti, Cheekati Sri Lakshmi, Ravuri Naveen Kumar, Ramya Paruchuri, "Text and Speech Recognition for Visually Impaired People using Google Vision", 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021

[5] Abhishek Mathur, Akshada Pathare, Prerna Sharma, Sujata Oak, "AI based Reading System for Blind using OCR", 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019

[6] Vaibhav V. Mainkar, Tejashree U. Bagayatkar, Siddhesh K. Shetye, Hrushikesh R. Tamhankar, Rahul G. Jadhav, Rahul S. Tendolkar, "Raspberry pi based Intelligent Reader for Visually Impaired Persons", 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020

[7] Javavrinda Vrindavanam, Raghunandan Srinath, Anisa Fathima, S. Arpitha, Chaitanya S Rao, T. Kavya, "Machine Learning based approach to Image Description for the Visually Impaired", Asian Conference on Innovation in Technology (ASIANCON), 2021

[8] R. Prabha, M. Razmah, G. Saritha, RM Asha, Senthil G. A, R. Gayathiri, "Vivoice - Reading Assistant for the Blind using OCR and TTS", International Conference on Computer Communication and Informatics (ICCCI), 2022

[9] S. Durgadevi, K. Thirupurasundari, C. Komathi, S.Mithun Balaji, "Smart Machine Learning System for Blind Assistance", International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), 2021

[10] Sandeep Musale, Vikram Ghiye, "Smart reader for visually impaired", 2nd International Conference on Inventive Systems and Control (ICISC), 2018