IMAGE TO AUDIO CONVERSION TO AID VISUALLY IMPAIRED PEOPLE BY CNN

Dr.Sivaganesan D
Department of Computer Science and Engineering
PSG Institute of Technology and Applied Research
Coimbatore, India
sivaganesan@psgitech.ac.in

Venkateshwaran M
Department of Computer Science and Engineering
PSG Institute of Technology and Applied Research
Coimbatore, India
m.venkateshwaran2001@gmail.com

Dhinesh S P
Department of Computer Science
Sri Krishna Arts and Science College
Coimbatore, India
dhineshponnarasan@gmail.com

***Abstract*— This study suggests an innovative method for helping people who are blind or visually handicapped by turning visuals into sounds. In the proposed system, audio descriptions are produced in real-time together with significant features that are extracted from photos using deep learning algorithms. The proposed work is developed to be user-friendly, which includes a simple interface that aids blind individuals to easily capture and process images using a mobile device. A user research was undertaken to assess the efficiency of the suggested method, and the results were encouraging in terms of precision and usability. This initiative offers a promising technique to give people who are blind or visually impaired an alternate means of perceiving and interacting with their environment, therefore improving their quality of life. The suggested picture to audio converter system aims to overcome the drawbacks of current assistive devices that rely on braille or textual descriptions. Blind people can more easily interpret visual information that is necessary for daily life, such as recognising items, interpreting signs, or navigate unfamiliar situations, through offering audio descriptions of images. The system makes use of recent deep learning developments that have significantly improved picture identification as natural language processing. As a result, the suggested technique has the ability to offer audio descriptions that are more precise and comprehensive than current methods. This technology has the potential to be implemented into a variety of products, from cellphones to intelligent glasses, and could significantly improve the lives of people who are blind or visually impaired.**

***Key Words*--- Deep learning, Visually impaired, Artificial Intelligence, Convolutional Neural Networks, Smartphones**

I. INTRODUCTION

In order to grasp what's going on about us and make wise judgements, visual information is vital to our daily life. However, for those with visual impairments, a major barrier to daily tasks can be an absence of ability to access visual information. There are many assistive devices available, such as reading[8] screens and braille displays, but their capacity to provide a thorough knowledge of visual content is frequently constrained. Systems for converting images to audio offer a potential remedy for this issue by enabling visually [10] challenged people to understand visual data through audio descriptions. In this article, we suggest a deep learning-based method for converting photos to audio descriptions that offer an evaluation of the system based on user feedback. In order to improve the standard of life for those with visual impairments, the system is made to be user-friendly and accessible.

Convolutional neural networks, the most widely used DL algorithms, have become the reference or go to point for image processing activities that focus on object identification and classification. Tools like these are considered to be essential for detection and analyzing purposes of the rapidly changing human activities in an image or video sequence. Computational requirements for training and hosting complex CNN networks, which includes actions like regional convolutional neural network, pose as a challenge to notice in many applications. These challenges have been relieved by advances in graphics processing units (GPU).The computer vision[3] community now makes extensive use of GPUs, especially those who are interested in deep learning for the purpose of increasing the speed in training the model and referring to it.

II. LITERATURE SURVEY

Authors: Sneha.C. Madre, S.B. Gundre et al [1] suggested solution is affordable and enables persons who are blind to understand the text. The fundamental idea behind this project is the conversion of written characters into audio signals using optical character recognition. Prior to character recognition, the text has been preprocessed by dividing each character. Following the process of segmentation the letter is extracted, and the text file is resized. Utilising MATLAB16, the aforementioned procedures will all be completed.

P Rohit, M S Vinay Prasad, S J Ranganatha Gowda, D R Krishna Raju,Imran Quadri, et al [2] proposed to give visually impaired persons with speech output-based scene perception, our study outlines the development of a system that operates in real time based on item detection, classification, and location estimate in an outdoor setting. The

technology is affordable, portable, straightforward, and simple to wear. The module is integrated onto the stick, while the pi-camera is utilised to take the picture while being moved in the appropriate direction by a controller. The system is then modified to better meet the demands of the user using the insightful information gleaned from the feedback.

Abhishek Mathur, Akshada Pathare, et al [5] proposed OCR is a process that turns printed, typed, or handwritten text into machine-encoded text. The picture will be examined, and the programme will interpret any English text and convert the results to speech. The purpose of presenting the output as voice or speech is to provide the data on the document to those who are blind.

Javavrinda Vrindavanam, Raghunandan Srinath, et al [7] proposed the study which is necessary because, in a world that is getting more and more digital, there are less and fewer ways for individuals who are blind or visually impaired to interact with the world, and using an image describer to access digital media can help those people. Processing of images that the visually challenged cannot see, creation of descriptions that are appropriate, and audio processing. Results are converted.

III. SYSTEM ANALYSIS

A. Existing System:

One such system is the picture Captioning system, which creates a textual description of a picture using a mixture of computer vision and NLP approaches. Text-to-speech [4] synthesis methods can then be used to turn this written description into an audio file. In order to recognise items, scenes, and other visual elements, the Image Captioning system first analyses intelligent [6] of an image's content using computer vision[19] techniques. Following that, a written account of the image is created using this information along with environmental knowledge and semantic guidelines. The textual description is analysed using Natural Language Processing (NLP) techniques, and then it is transformed into an audio format that people can easily understand.

B. Problem Definition:

Convolutional neural network (CNN)-based image to audio conversion[15] is the process of transforming an input picture into an appropriate sound signal via a deep learning model created expressly for this purpose. The primary goal of this work is to provide a sound file that accurately captures the key elements and substance of the picture that was input. Using CNNs to convert images to audio presents a number of difficulties, including the need to model complex audio waveforms, train to extract significant information from images, and create high-quality audio outputs from image data. The answer to this issue may be used for a variety of purposes, such as the creation of multimedia content, audio-based picture browsing, and image captioned for the blind[9].

C. System Architecture:

Every entity presently incorporated into the system are succinctly and clearly described in this graphic. The figure demonstrates the relationships between the various activities and decisions. The entire procedure as well as how it was handled[20] might be described as a picture. The functional relationships between various entities are depicted in the image below.

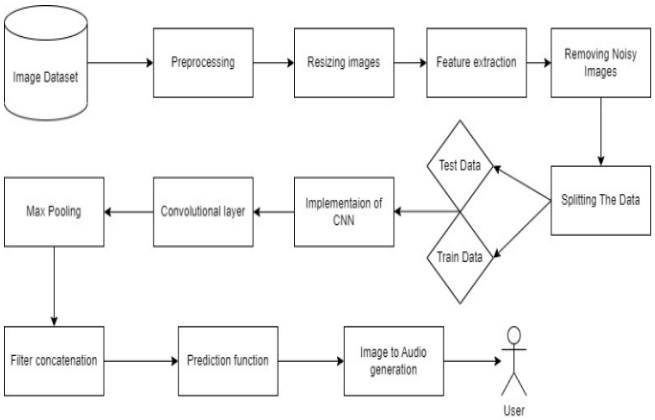

FIGURE III-1: Architecture Diagram

D. Data Flow Diagram:

A data-flow diagram (DFD) can be used to show how information moves through a process or system. A diagram of data flow does not show any loops or decision-making processes because information only flows in one direction. Data-flow graphs can be represented in a variety of ways. Procedures, flow, storage, and terminators are all parts of structured data modelling (DFM). Data flow diagram symbols are Process, Data Store, Data Flow, External Entity. In a level DFD, the entire system is depicted as a single process. Here, the entire assembly procedure for the system, comprising all intermediate steps, is documented. This and two-level data flow diagrams make up the "basic system model".

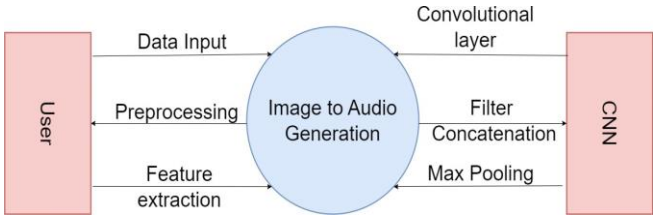

FIGURE III-2: Data Flow Diagram level 0

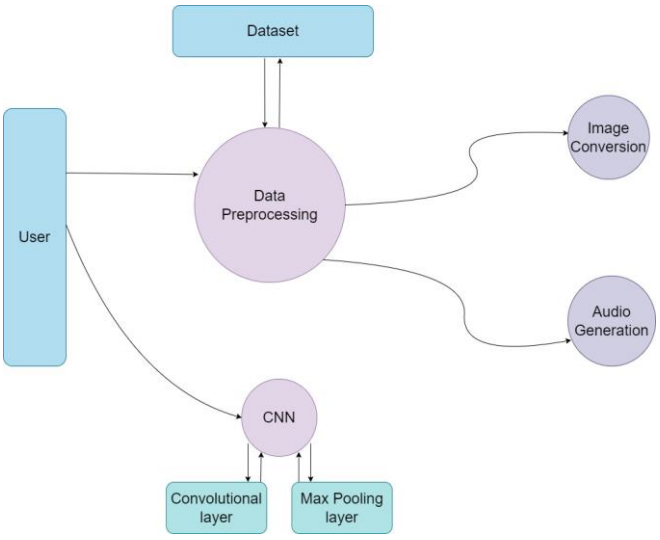

FIGURE III-3: Data Flow Diagram level 1

E. ER Diagram:

An entity-relationship diagram (ERD) is a visual representation of the connections between database entities. Through the use of data object descriptions, additional information can be added to the relationships and entities shown in an ERD. Entity-relationship models (ERMs), which are used in software engineering, are used to express concept and abstract data descriptions. Symbols used are external entity, Attribute and Relationship.

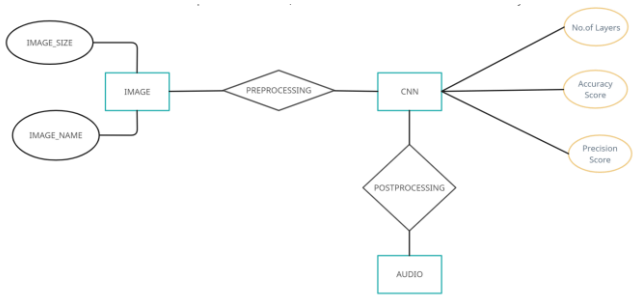

FIGURE III-4: ER Diagram

F. Activity Diagram:

In its most basic form, an activity diagram is a diagram that shows the order in which tasks are carried out. It shows the order of steps which make up the entire process. Although they aren't quite flowcharts, they have a similar function.

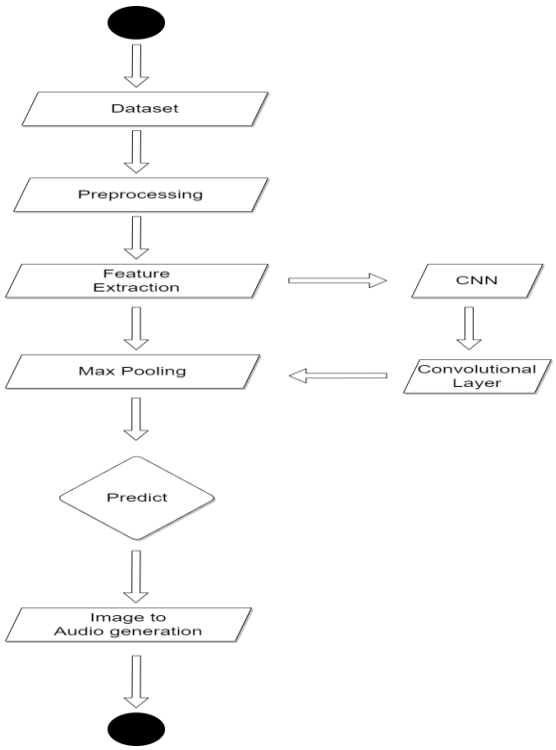

FIGURE III-5: Activity Diagram

G. Use case Diagram:

A use case diagram frequently shows the potential relationships among the user, the data set, and the algorithm. It is made at the beginning of the process.

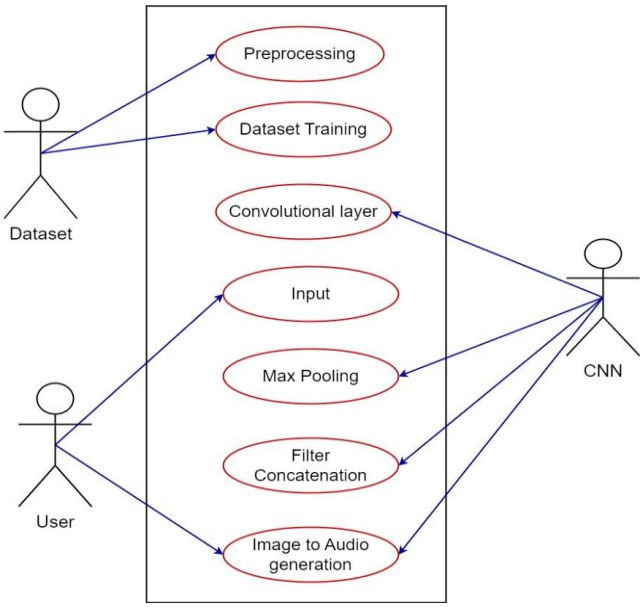

FIGURE III-6: Use Case Diagram

H. Sequence Diagram:

These are an additional sort of interaction-based diagram intended to show how the system functions. They keep track of the circumstances in which things and processes work together.

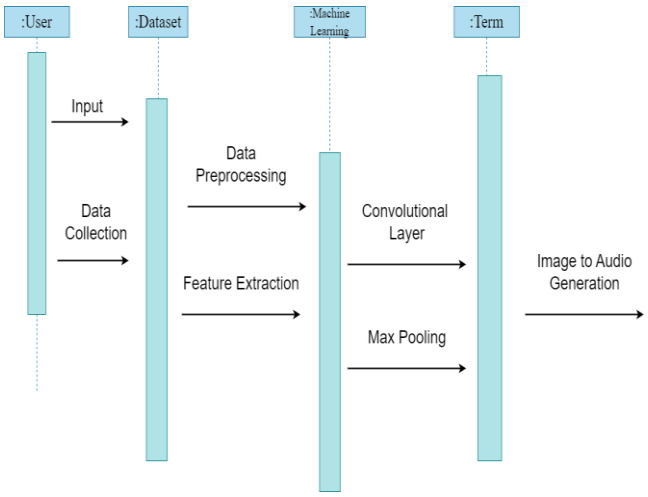

FIGURE III-7: Sequence Diagram

I. Class Diagram:

This is basically a "context diagram," which is another term for a context diagram. It merely refers to the procedure's 0 Level, which is its highest point. The system is

depicted as a whole as one procedure, with an abstract representation of the relationship with externalities.

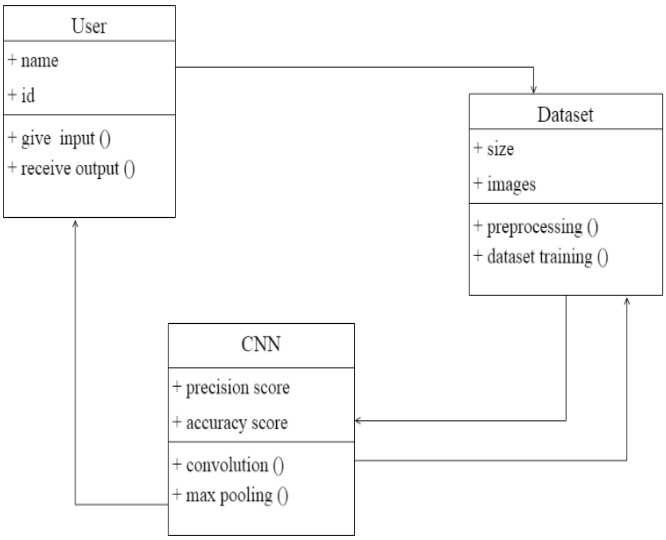

FIGURE III-8: Class Diagram

IV. PROPOSED MODEL

The steps for image to audio conversion[16] using CNN involves the following:

Preprocessing: The system first preposses the input to extract data which is pertinent. In order to do this, visual features from the image would need to be obtained via a convolutional neural network (CNN).

Encoding: The LSTM (Long Short-Term Memory) networks would then be used to encode the visual attributes into a series of vectors. The links among the visual characteristics and the related audio descriptions would be taught to the LSTM network.

Decoding: Encoded audio vector sequence from decoder would entail transforming the encoded vector sequence into an audio waveform via a text-to-speech[14] synthesis method.

Postprocessing: The output audio stream would next undergo post-processing to enhance its clarity and quality. Techniques like noise reduction, equalisation, and amplification may be used in this.

J. Module 1:Image pre-processing

This module is vital in preparing the provided picture data for the CNN-LSTM training process. The pre-processing[11] module uses a variety of techniques and algorithms to enhance the quality of the incoming photographs, identify useful traits, and scale the values of pixels to a predetermined range.

The photographs must first be resized to a specific resolution by the preprocessing programme. This phase ensures every one of the images have the same size in order to effectively train the CNN-LSTM. The CNN-LSTM will be better able to identify patterns and characteristics from the images[12] by standardising the picture[13] size. Images of different sizes could make network training challenging.

Enhancing contrast is a key function of the pre-processing module. Contrast enhancement methods are used to increase the visibility of visual characteristics and help the CNN-LSTM find significant elements that can be used to produce audio. Several popular techniques for improving contrast include gamma correction, adaptive histogram equalisation, and histogram equalisation. These techniques work by boosting contrast and changing the intensity levels of the image's pixel values.

Feature extraction is yet another vital task carried out by the pre-processing module. This method involves finding and extracting important components in order to transform the image to audio. Multiple characteristics may be extracted, based on the assignment at hand as well as the content of the image data. Among the regularly used feature extraction techniques are detection[17] of edges, texture analysis, and colour analysis. The CNN-LSTM must be able to identify important patterns and features in the visual input in order to produce trustworthy audio.

Normalisation is yet another crucial phase of the pre-processing module. A consistent scale is created by transforming the image's value of pixels during the normalisation process. This step is necessary to guarantee the CNN-LSTM may be taught effectively.

The pre-processing programme raises the quality[18] of the input photographs, extracts important data, and adjusts the pixel values to a consistent scale. This module is essential for the CNN-LSTM to recognise and learn significant trends and features in the visual data that might be utilised for precise audio creation.

K. Module 2: CNN-LTSM Architecture:

The second element entails edges, shapes, or textures are just a few examples of the specific themes or qualities that the filters are designed to recognise in the visual input.

The amount of filter per layer and the dimension of the filter are two hyper parameter settings that can be changed to improve CNN performance. Pooling layers are used to decrease the result if the convolutional layers in order to reduce the complexity of the data and prevent overfitting. The maximal pooling layer, the most common type of pooling layer, selects the highest possible amount within a specific area of the input data.

After the CNN-LSTM structure has been defined, the CNN-LSTM is taught using a collection of labelled images and the corresponding audio labels. The CNN-LSTM trains to map the provided imagery to the audio labels by adjusting the biases and weight using an optimisation strategy like gradient descent.

This module describes the CNN-LSTM parameters and the way the CNN-LSTM trains to generate sounds from the previously processed picture data. The CNN-LSTM design module allows the CNN-LSTM net to understand the complex mappings among its input visuals and associated audio signals, allowing for efficient training.

L. Module 3: Post Processing:

The output of CNN is a set of feature vectors that depict the probability to generate each audio sample. LSTM's The post-processing module analyses this set of feature vectors before producing the final audio stream. In order to achieve this, the post-processing module may perform procedures including denoising, spectrogram generation, and inverse Fourier transform. Denoising helps to eliminate any possible noise from a feature vector sequence. The series of vectors of features is represented as a spectrogram in the time domain to efficiently process and handle the audio input. A series of feature vectors are transformed from the domain of frequencies to the duration plane utilising the inverse Fourier transformation to produce the final audio output.

Waveform normalisation is used to ensure that the audio signal's amplitude falls within a specific range. Spectral smoothing is used to reduce any jarring changes or anomalies that may be found in the audio source. It analyses the LSTMs produced by CNN and transforms them into an effective representation of the sound signal. The post-processing tool takes certain the audio signal generated accurately and superbly represents the input visual data.

V. SYSTEM IMPLEMENTATION AND RESULTS

Step1: Click on browse files and choose an image in the local storage of the device.

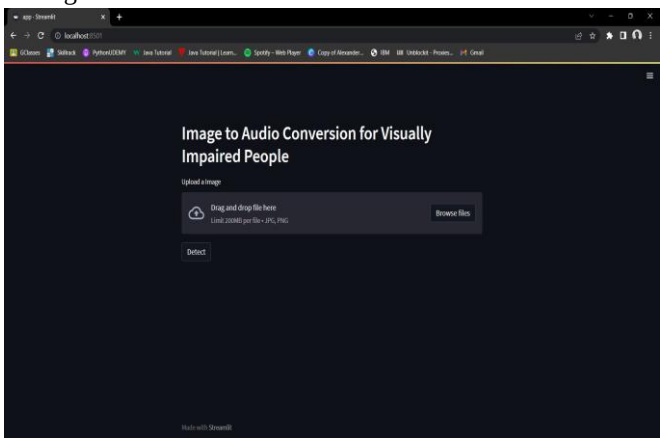

FIGURE V-1: Upload Image Page

Step 2: After uploading select detect for the image choosen

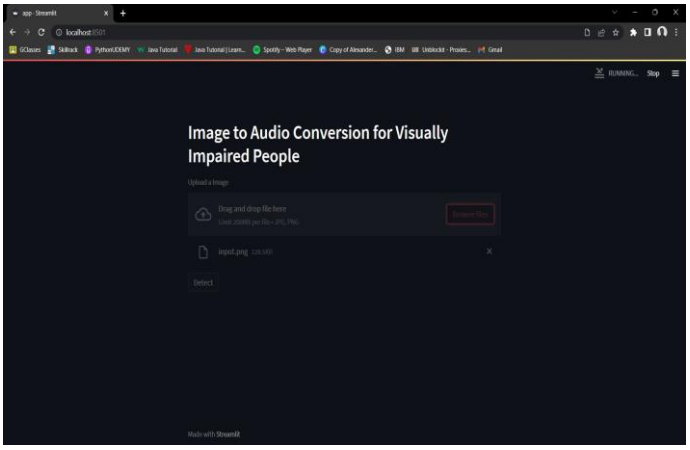

FIGURE V-2: Detect Image

Step3: The uploaded image pops up with the caption describing the image.

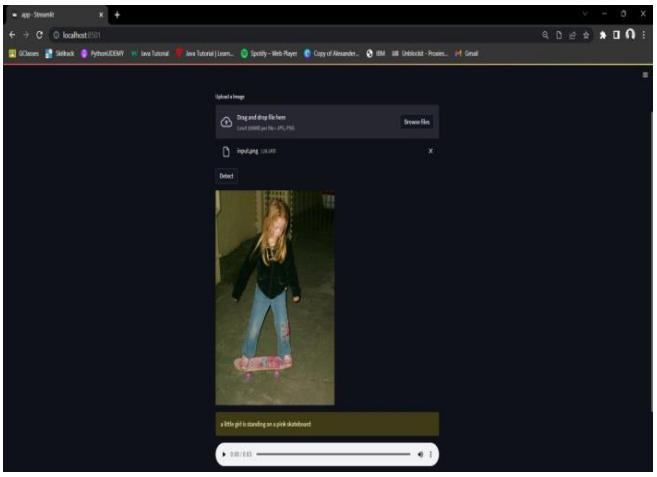

FIGURE V-3: Audio with image Page

Step4: Click on the audio button to hear the audio which would describe the image.

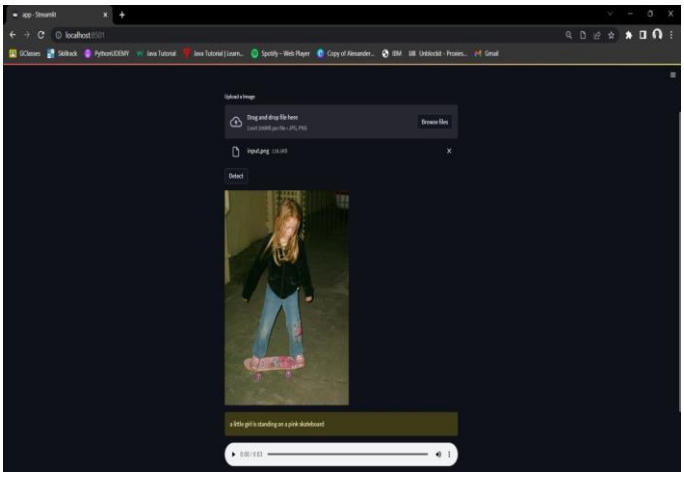

FIGURE V-4: Play Audio

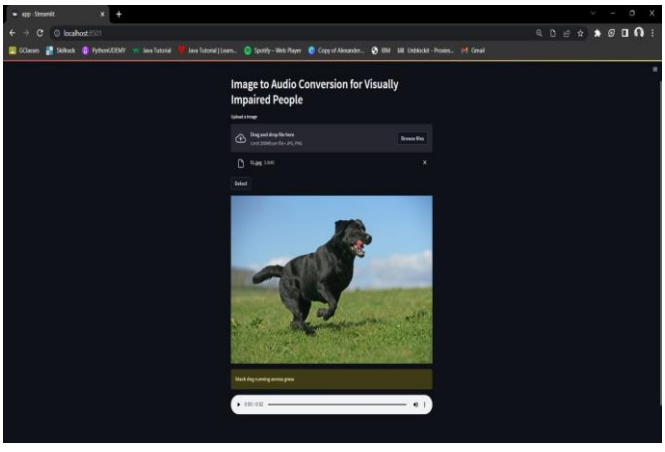

FIGURE V-5: Sample image of dog

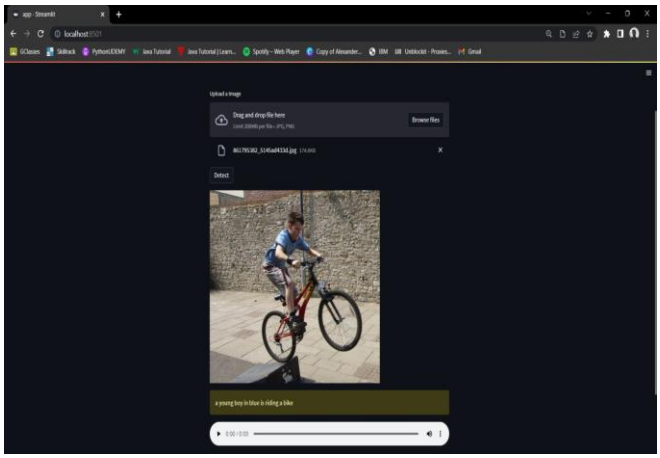

FIGURE V-6: Sample image of boy with cycle

VI. CONCLUSION AND FUTURE WORK

In conclusion, this work provides a deep learning-based method for providing visually impaired people with audio descriptions of photographs. Convolutional neural networks (CNNs) are used in the proposed system to extract significant features from input photos and produce related audio descriptions in real-time. The method also includes attention algorithms that help the resulting audio descriptions be more accurate and pertinent by focusing on the images' key elements. The overall goal of this paper is to advance the use of CNNs in the development of efficient and dependable picture to audio conversion systems, that have the potential to improve our society's accessibility and inclusivity by allowing people who are blind or visually impaired to interact with and perceive pictures through audio descriptions. Future research could examine the applicability of this strategy to more complicated image types, such videos, and look into the system's potential for commercial application.

VII. 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), 2019

[11] Archana A. Shinde, D.G.Chougule "Text Pre-processing and Text Segmentation for OCR" IJCSET , January 2012.

[12] Benjamin Z. Yao, Xiong Yang, Liang Lin, MunWai Lee and Song-Chun Zhu, "I2T: Image Parsing to Text Description".

[13] Bernard Gosselin Faculté Polytechnique de Mons, Laboratoire de Th´eorie des Circuits et Traitement du Signal, "From Picture to Speech: An Innovative Application for Embedded Environment".

[14] Huizhong Chen1, Sam S. Tsai1, Georg Schroth, David M. Chen, Radek Grzeszczuk and Bernd Girod, “Robust text detection in natural images with edge-enhanced maximally stable extremal regions”, International Conference on Image Processing • September 2011

[15] Jisha Gopinath, Aravind S, Pooja Chandran, Saranya S S, “Text to Speech Conversion System using OCR”, International Journal of Emerging Technology and Advanced Engineering, January 2015.

[16] Itunuoluwa sewon, Jelili Oyelade, Olufunke Oladipupo, “Design and Implementation of Text To Speech Conversion for Visually Impaired People”, International Journal of Applied Information Systems (IJAIS, 2014).

[17] Yao Li and Huchuan Lu, “Scene Text Detection via Stroke Width”, 21st International Conference on Pattern Recognition (ICPR 2012) November 11-15, 2012. Tsukuba, Japan.

[18] Alías F. Sevillano X. Socoró J. C Gonzalvo X. (2008), „Towards high-quality next-generation text-to-speech synthesis‟, IEEE Trans. Audio, Speech, Language Process, Vol. 16, No. 7. pp. 1340-1354.

[19] Balakrishnan G. Sainarayanan G. Nagarajan R. and Yaacob S. (2007) „Wearable real-time stereo vision for the visually impaired‟, Vol. 14, No. 2, pp. 6–14.

[20] Chucai Yi. YingLiTian.AriesArditi. (2014), „Portable Camera-based Assistive Text and Product Label Reading from Hand-held Objects for Blind Persons‟, IEEE/ASME Transactions on Mechatronics, Vol. 3, No. 2, pp. 1-10.