Journal of Information Systems Engineering and Management
2025, 10(23s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Interactive Image Exploration for Visually Impaired Readers using Speech Captioning Model.

Pritam Langde[1*], Shrinivas Patil[2]
Research Scholar, Department of Electronics and Telecommunication Engineering, DKTE Research Center, Ichalkaranji, India
Professor, Department of Electronics and Telecommunication Engineering, DKTE Research Center, Ichalkaranji, India

ARTICLE INFO

Received: 16 Dec 2024
Revised: 02 Feb 2025
Accepted: 20 Feb 2025

ABSTRACT

This paper presents a system for assisting blind individuals for reading printed document. This system utilizes a collaborative approach by combining Optical Character Recognition (OCR) and the Scale-Invariant Feature Transform (SIFT) algorithm to recognize text and extract images. The proposed system utilizes SIFT features to extract and recognize the content of captured image documents. Additionally, it employs OCR technology to read the text content. Subsequently, the system transforms the identified text into speech using Text-to-Speech (TTS) technology, and delivers auditory responses to the user. The system underwent testing using a dataset consisting of printed text documents from Higher Secondary School History Books (HSSHB) and achieved a commendable level of accuracy. In order to facilitate computer usage for individuals with visual impairments, we employed the NVDA (Non-Visual Desktop Access) open-source software. The designed system exhibits the characteristics of being cost-effective, small in size, highly effective, and user-friendly. The results indicate that the system will enhance user-friendliness when reading documents by combining text and images. The system exhibited a commendable accuracy rate of approximately 92% in discerning printed text within the documents. The system demonstrated an impressive accuracy rate of around 91% in the field of image detection. This level of precision encompasses the collective results from the "Good" and "Moderate" categorizations, highlighting its proficiency in identifying and counting images within documents. The system achieved a commendable level of accuracy in word detection, with 82% of the documents classified as "Good."

Keywords: SIFT, OCR, Image processing, NVDA, Visually Impaired people

1. INTRODUCTION

Recent reports from the World Health Organization (WHO) have revealed that a considerable proportion of the worldwide populace [1-2][14-16] suffers from visual impairments. As of 2023, this figure is estimated to amount to 286 million individuals. Visual impairments frequently arise from congenital anomalies or uncorrected refractive errors, underscoring the urgent requirement for alternatives to conventional medical treatments. India is home to an astonishing 22 percent of the globally affected visually impaired population. However, the country performs only 46,000 transplant surgeries annually, which falls far short of the substantial demand caused by the high prevalence of visual impairments and the exorbitant cost of eye transplant procedures, which exceeds 1,50,000 Indian Rupees [1]. Eye transplant surgery is therefore prohibitively expensive for some [2]. Individuals who have visual impairments encounter a distinct array of challenges on a daily basis, which transcend the mere absence of sight. These obstacles affect all aspects of everyday life, including the ability to participate in a visually oriented society and obtain vital information. In sightlessness is a significant obstacle for individuals who are blind [3]. They face significant obstacles in reading, information access, and day-to-day activities. Long have the blind and visually impaired traversed a distinct linguistic terrain in a society influenced by the written word. Their script language, Braille, is a physical manifestation of the ingenuity and perseverance of the human race. However, it cannot be denied that Braille, despite

* Corresponding Author:
Email: pritamlangde@gmail.com

being of great value, poses a significant obstacle to learning—one that has caused many individuals to find it exceedingly difficult to master this script [4]. Amidst this obstacle, it is critical to acknowledge that members of the visually impaired community possess a remarkable ability to perceive and comprehend their surroundings via auditory stimuli. As evidence of the adaptability of the human species, this auditory prowess enables them to interact with their surroundings and extract vital information.

These blind individuals are unable to establish a certain degree of connection with any objects due to their inability to read content that is not provided in braille format. Consequently, these individuals encounter difficulties in realizing their complete intellectual capabilities due to the limited availability of readily obtainable books, magazines, and textual materials [5]. Engaging in reading, which is essential for personal development and intellectual stimulation, presents a formidable obstacle for individuals with visual impairments, impeding their ability to delve into the extensive corpus of written information. Due to the importance of reading to the visually impaired, the majority of them depend on Braille or audio books to read. However, these approaches are not without their drawbacks. The progression of computer vision and machine learning presents a promising prospect for the creation of a system capable of aiding visually impaired individuals in the comprehension of printed text.

The task of content-based multimedia database indexing and retrieval is an essential component of modern information retrieval systems. It involves the automatic extraction of descriptive features that are relevant to the subject materials, which can include images, video, and other multimedia content [6-8]. Text detection and recognition in images has been emerged as pivotal components in the development of advanced image and video annotation and retrieval systems. This integration brings together cutting-edge Optical Character Recognition (OCR) technology with text-based searching methodologies, addressing the critical need for efficient and effective ways to work with textual content embedded within visual media. Recent advancements in research have shifted the focus towards extracting more descriptive features and higher-level entities, such as text and human faces, from multimedia content. This shift has been driven by the recognition that text, especially in the form of captions, carries valuable content information. The integration of SIFT (Scale-Invariant Feature Transform) enhances the feature extraction process, enabling more accurate recognition of text in diverse and complex backgrounds.

This section presents previous relevant studies concerning blind people's spatial auditory perception and use of interactive VR environments in cognitive studies. A detailed literature review, as well as background studies related to the current work, can also be found in recent published article by Naz et al. [9]. A selection of the literature is reviewed in this article, with a focus on a particular application for document readers, their successful outcomes, and their innovative approaches. Text recognition is utilized on the majority of paper. Perera et al. [10] demonstrated a method for visually impaired individuals to describe actions using video data. The demonstrated method extracts a feature set from the projection histograms of the foreground mask for each frame. Musale and Ghiye [11] demonstrated an efficient system for the visually impaired known as a smart reader. The OCR (Optical Character Recognition) capabilities of MATLAB were employed to transform the image into text. Ashmafee and Sabab [12] investigates a blind intelligent assistant. Given that Braille is one of the methods utilized by the visually impaired to read books or documents, the author's primary objective was to devise an algorithm that could efficiently convert any given document to braille. They developed an intelligent device with a multimodal system that can provide a blind person with an interpreted version of any document. Khan et al. [13] explored the historical perspective to examine the current body of research, which spans from the earliest investigations into electronic travel aids to the application of contemporary artificial vision models in BVIP navigation.

Ashiq et al. [14] developed an intelligent and sophisticated system to facilitate the movement of VIPs and guarantee their security. Real-time navigation is provided by the proposed system via an automated voice. While lacking visual access to their immediate surroundings, VIPs possess the ability to perceive and envision the roaming environment. In order to evaluate the effectiveness of the proposed system, six pilot studies with positive results have been conducted. In order to detect and recognize objects, a deep Convolution Neural Network (CNN) model is utilized, which achieves an accuracy of 83.3%. It is worth noting that the dataset comprises over a thousand distinct categories. Masud et al. [15] designed a versatile, wearable, modest, and secure framework to assist the visually impaired in their daily lives. In order to accomplish this, the objective is to develop a proficient system that can aid individuals with visual impairments by detecting obstacles and classifying scenes. The proposed methodology makes use of an Arduino, Raspberry-Pi 4B, camera, and ultrasonic sensor that are all mounted on the individual's stick. Images of the scene are captured and subsequently subjected to pre-processing using the TensorFlow Object Detection algorithm

and Viola Jones. Rane et al. [16] proposed a new smart spectacle implementation using Image Captioning and OCR to simplify navigation. The system includes a camera in glasses, Image Captioning, OCR, and TTS modules. Text-to-speech is mostly used to convert this data into voice notifications for the visually impaired. The model can generate captions from its surroundings and read text when needed. The model can read English, Hindi, and Marathi, meeting vernacular needs. Bhalla et al. [17] combined in Visual Interpreter of Environment Wizard (VIEW), a computationally efficient system for portable, low-powered devices. VIEW helps visually impaired people navigate by providing natural language descriptions of their surroundings, relative object movement, and navigational tools. The model is trained on the Common Objects in Context (COCO) dataset for multiple object classes and processes the image once during inference, speeding up image processing.

Islam et al. [18] introduces a novel method for augmenting image captioning for visually impaired individuals by combining depth data with RGB images. The proposed model, which is intended to address these issues, has the potential to improve scene comprehension and navigation for the visually impaired. Kornsingha et al. [19] presented a methodology for the development of a voice system capable of reading medication labels. Pharmaceutical product labels include information regarding the medication, the type of medication, the description of the treatment, and the duration of use. A microcontroller and RFID (Radio Frequency Identification) technology are implemented. Information pertaining to an RFID tag-based system is pre-established in the database and is retrieved upon scanning the tag, followed by its conversion into speech.

Optical Character Recognition (OCR) and document layout analysis are two approaches that have been proposed for document image analysis and recognition. However, these methods may not be sufficient for virtually impaired individuals as they require visual input [20-21]. The SIFT method has been used successfully for object recognition. According to the results of the survey, more work was done in the area of preparing devices such as walking sticks, goggles, raspberry pi-based readers, RFID-based labeling devices, and so on. In existing systems, a complex hardware system was used, which was not appropriate in all situations.

In order to address the challenge faced by individuals who are visually impaired or blind in comprehending documents, the authors have developed a system that analyzes a combination of text and images. Individuals who are blind can easily perceive the sound and comprehend the text contained within the sound source. The intelligent framework is necessary for comprehending the page contents, as it distinguishes between text and image content types derived from the document. It may be necessary to implement an approach based on the SIFT Scale-Invariant Feature Transform (SIFT) and OCR (Optical Character Recognition) algorithms in order to identify, classify, and describe the categories of the contents in natural language text. This system works with NVDA (Non-Visual Desktop Access), making the interface smooth and easy to use. This integration makes it easy for users to find their way around and use the system, which improves accessibility and independence. This system is also made to be flexible so that it can be used for many things, from reading printed books to reading text on signs and labels in real time. Because it is flexible, it can be used by people who are blind or visually impaired in a variety of settings. Moreover, this textual description will be rendered audibly, potentially benefiting individuals with visual impairments.

This study aims to conduct research on the development of a document reader system capable of generating voice descriptions of document contents, irrespective of whether those contents consist of text, images, or a combination of both. The model is trained using a sequence-to-sequence approach on a large dataset of images and evaluated using content from academic history books written in English medium for secondary schools and published by the Maharashtra state board. We employ the commonly employed metrics of word error rate (WER), character error rate (CER), and mean opinion score (MOS) to assess the performance of our system. The accuracy and quality of the produced audio are assessed using these metrics; a higher MOS and a decrease in WER and CER signify superior performance.

2. DESIGN OF THE PROPOSED DEVICE

This section provides an elaborate exposition of our model. Authors present a visual tool designed to assist individuals with visual impairments, featuring a built-in reading assistant. The setup offers immediate auditory feedback to the user via the headphones in real time. The diagram in Figure 1 illustrates the workflow of the system. According to the diagram shown in Figure 1, the model is depicted with distinct interfaces for both hardware and software. The Hardware interface gathers data from the environment, utilizing a camera module for capturing images. The software interface consists of distinct procedures for image detection, SIFT (Scale-Invariant Feature Transform) analysis,

image captioning, and text detection using the Tesseract OCR (Optical Character Recognition) engine. The collective outcome of these procedures operates on the acquired image from the camera module and provides auditory information to the user via headphones or speakers.

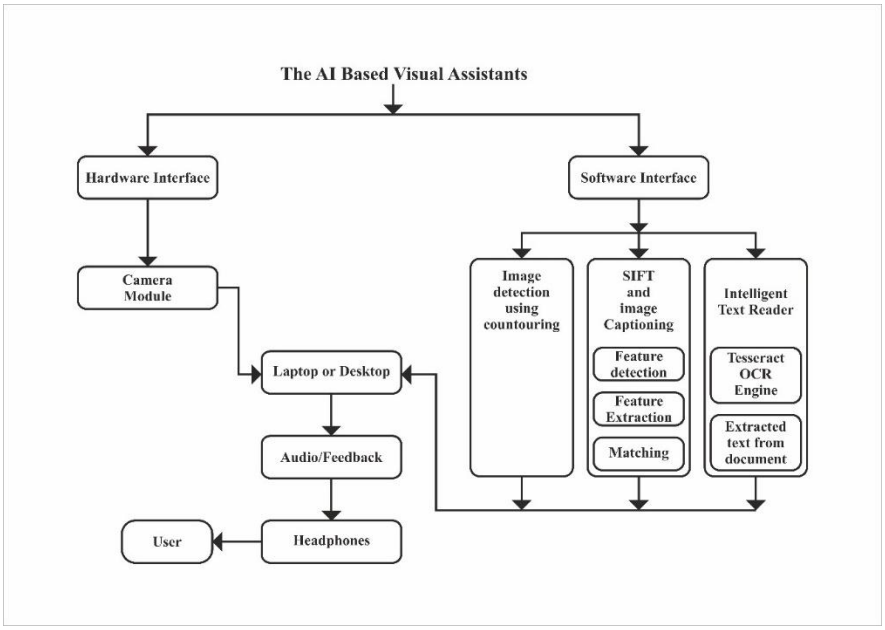

Figure 1. Complete Workflow of proposed System. (need high resolution image)

2.1 Hardware interface and software

The process of acquiring and analysing document images for the purpose of text recognition and speech captioning is an essential element of our system. In order to achieve the utmost level of excellence in our outcomes, we utilize a camera of superior resolution, accompanied by a stand, and implement suitable lighting conditions to enhance the clarity of the images. The document images are acquired using a high-resolution camera that is positioned on a stand. In addition, we ensure the provision of ample illumination to enhance the clarity of captured images for subsequent processing. During this stage, the text's image is acquired using a camera with a minimum resolution of 5 megapixels. The captured image documents exhibit imperfections in terms of their shape and size. During the process of capturing document images, it is important to recognize that they may not always adhere to ideal shapes and sizes. Consequently, we integrate an image processing module to preprocess the captured images, guaranteeing their suitability for subsequent text extraction. Furthermore, the captured image is not in a compatible format for text extraction input. Firstly, the captured image is inputted into the image processing module. The image that was taken is in the .jpg format.

A high-definition camera with ultra-HD resolution of 1920 * 1080 pixels, autofocus is utilized. The camera captures high-definition images with a focal range of 20 cm. High-resolution and autofocus functionalities are essential for capturing precise and crisp document images. The camera's elevated sensor resolution ensures that text is captured accurately, even when taken from short distances. The document image that has been captured can be obtained from the camera and saved in the default .jpeg format within a designated directory. After saving the image in the system, we can use these images for subsequent text recognition from documents.

The primary hardware component of the system is the computer with a fundamental configuration. The system should have a minimum of 4 GB RAM and an Intel i3 6th generation or higher processor. The device should possess multiple USB ports in order to establish connections with various peripherals. A USB cable can be utilized to connect a camera and a laptop. The captured images are being displayed in real-time. Depending on the specific needs of the visually impaired individual, we have the option to select either Bluetooth headphones or speakers. Finally, in order to utilize the as text-to-speech module, it is necessary to have an internet connection available either through a wired LAN or Wifi.

The software interface incorporates a speech captioning model. It primarily depends on four essential subsections. Document text recognition, image recognition using contouring from documents, SIFT (Scale-Invariant Feature

Transform), and captioning, as well as text-to-speech conversion. SIFT improves the system's capacity to identify and compare image characteristics, which can be advantageous for tasks like aligning documents and conducting image-based searches. Text-to-Speech serves as a conduit that converts extracted text into an audible format, allowing visually impaired users to access it through headphones or speakers.

Figure 2. Gives hardware setup of the proposed device

2.2 Text Recognition

An Optical character recognition (OCR) based approach is used for normal text recognition. This approach is quite easy to use the thing for generating description voice in document reader concept. The text recognition technique process includes the use of Python-tesseract, which is an optical character recognition (OCR) tool for python. it will recognize and “read” the text implanted in images. Overall, this Python-tesseract is a wrapper for Google's Tesseract-OCR Engine. in addition to this, we need to add getting a grayscale image, noise removal, and thresholding processes. Once the Text recognition process is complete then the result is stored in a folder with the name MyFile.txt

2.3 Image Recognition using countouring

Contouring, in the context of image processing, pertains to the task of detecting and delineating curves that connect adjacent points with identical color or intensity along a boundary. These contours are useful for a range of applications, such as shape analysis, object detection, and recognition. Contouring refers to the creation of a smooth curve that connects all the adjacent points with the same color or intensity along the boundary. Contours are a valuable tool for analyzing shapes and detecting and recognizing objects. To enhance precision, utilize binary images.

A contour refers to a delineation surrounding an object with clearly defined boundaries, enabling a machine to measure variations in gradient and create a distinguishable shape. It is employed for the purpose of identifying objects within images. To select the contour of objects, the hierarchy is as follows. The provided figure displays the outcomes of contouring, demonstrating a contrast between the original images and the segmented images. A contour is a fundamental concept that delineates the perimeter of an object characterized by distinct boundaries. Contouring is a crucial technique in image object recognition, as it enables machines to compute changes in gradient and create distinguishable shapes. By tracing these outlines, a machine can recognize and distinguish between different objects depicted in an image. This visual comparison illustrates the efficacy of contouring in accurately outlining objects and shapes in images, highlighting its importance in image analysis and object detection.

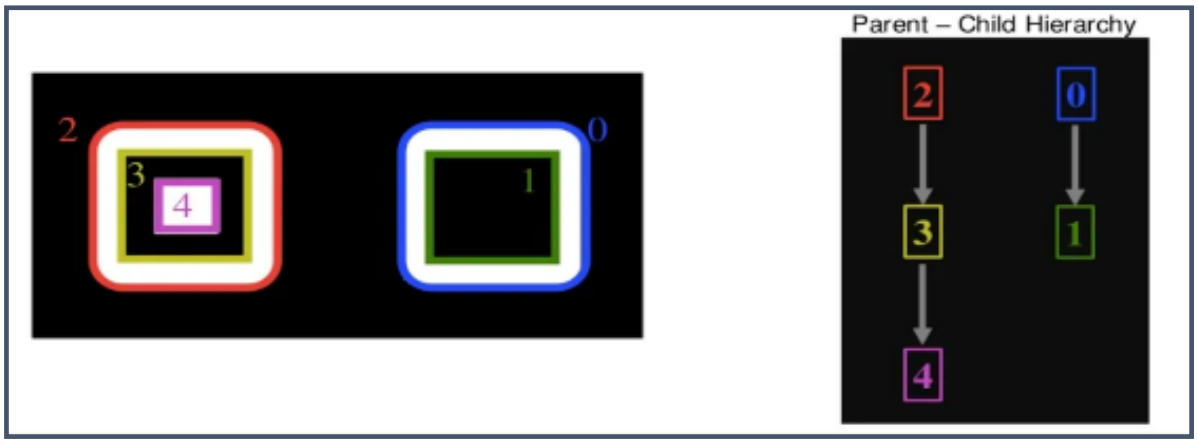

Figure 3 Contour Hierarchy

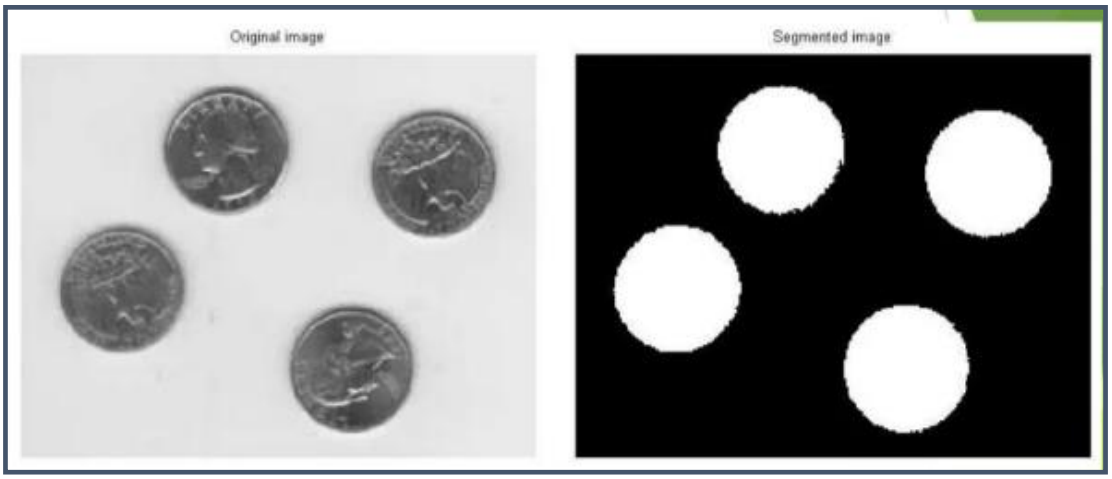

Figure 4 Contouring results

3. SIFT (SCALE-INVARIANT FEATURE TRANSFORM)

The SIFT algorithm is a prevalent technique in computer vision for object recognition, specifically for describing local features in images. The Scale-Invariant Feature Transform (SIFT) algorithm is a well-established method for detecting and describing local features in images. Our research leverages SIFT's robustness and accuracy in varied and cluttered environments to enhance text and image recognition within our reading assistance system. For detailed information on the SIFT algorithm, readers are referred to David Lowe's seminal paper 'Distinctive Image Features from Scale-Invariant Keypoints.

During the feature description process, the algorithm assigns an orientation to keywords and generates a histogram based on the gradient direction. The keypoints are subsequently characterized using a collection of local features that remain unaffected by changes in scale, rotation, and translation. The SIFT algorithm converts the image into a set of local feature vectors. The purpose of these feature vectors is to possess distinct characteristics and remain unaffected by any changes in scaling, rotation, or translation of the image. SIFT, as indicated by its name, possesses the characteristic of scale invariance. By initially detecting key points in an image, which are points that exhibit significant contrast in relation to the pixels surrounding them. The key points are subsequently characterized using a collection of local features that remain unchanged regardless of changes in scale, rotation, or translation. Descriptors are employed to align keypoints in various images, enabling precise object recognition. Employing SIFT, our system effectively recognizes and matches key points in images, ensuring robustness against various transformations. This capability is integral to our system's performance in real-world conditions. The SIFT algorithm encompasses several steps, namely Difference of Gaussians (DoG), Space Generation, Key points Detection, and Feature Description. SIFT exhibits invariance to changes in image scale and rotation. During the implementation of the algorithm, SIFT identifies specific key points and provides them with quantitative information, known as descriptors, which can be utilized as an example for object recognition.

3.1 Feature Detection

In this process, feature point (also called keypoint) detection is takes place. The SIFT algorithm transforms the image into a collection of local feature vectors. These feature vectors are aimed to be distinctive and invariant to any scaling, rotation or translation of the image. As its name shows, SIFT has the property of *scale invariance*. SIFT combines the pyramids and different σ-space to detect blobs under different scales.

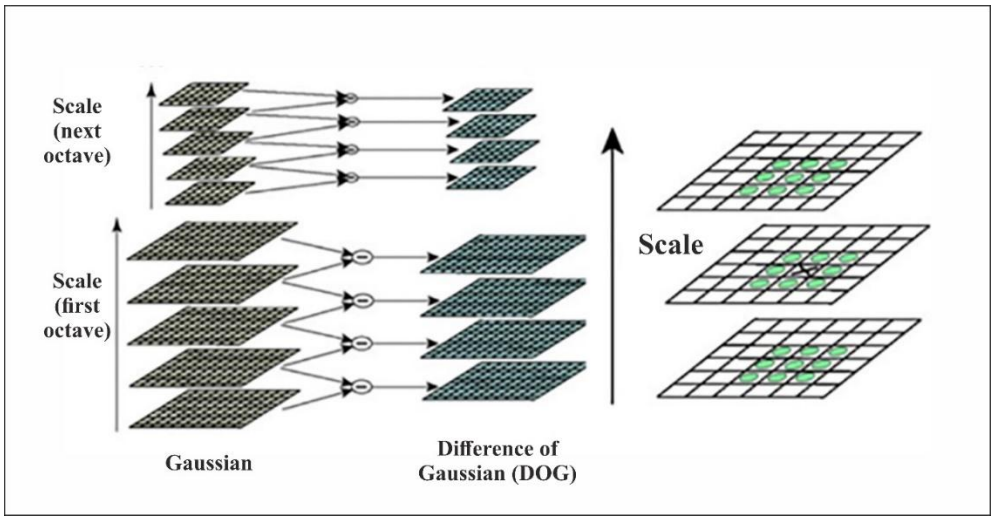

Figure 5 Feature Detection Process in SIFT [20]

After completion of detection process , the feature point localization process will be carried out . here we detect some key-points which are coarsely localized, at best to the nearest pixel, dependent upon where the features were found in the scale-space. They are also poorly localized in scale since σ is quantized into relatively few steps in the scale-space. The second stage in the SIFT algorithm refines the location of these feature points to sub-pixel accuracy whilst simultaneously removing any poor features. This process majorly focus on the Octave and gaussian.

3.2 DOG (Difference of Gaussian)

In image processing, the Difference of Gaussians (DoG) algorithm is a potent feature enhancement technique. It works by taking the difference between two Gaussian-blurred versions of the original image, where the second version is less blurry than the first. This entails convolving the original grayscale image in the case of grayscale images using Gaussian kernels with different widths (standard deviations). In the simplest scenario, grayscale images are convolved with Gaussian kernels of varying width (standard deviations) to produce blurred images from the original grayscale images. When a Gaussian kernel is used to blur an image, only high-frequency spatial information is suppressed. Spatial information lying between the range of frequencies preserved in the two blurred images is preserved when one image is subtracted from the other. Thus, the DoG functions as a spatial band-pass filter, attenuating frequencies that are distant from the band center in the original grayscale image. The subtraction step is where the DoG algorithm shines. Spatial information between the range of frequencies retained in the two blurred images can be preserved by subtracting one blurry image from the other. Stated differently, the DoG functions as a spatial band-pass filter. It attenuates or reduces information present at frequencies far from the band center while enhancing the image's features that fall within a particular range of spatial frequencies. Difference of Gaussians (DoG) is a powerful feature enhancement algorithm used in image processing. It operates by subtracting one Gaussian blurred version of an original image from another, with the second version being less blurred than the first. In the case of grayscale images, this involves convolving the original grayscale image with Gaussian kernels that have varying widths (standard deviations). In the simple case of grayscale images , the blurred images are obtained by convolving the original grayscale images with Gaussian kernels having differing width (standard deviations). Blurring an image using a Gaussian kernel suppresses only high-frequency spatial information. Subtracting one image from the other preserves spatial information that lies between the range of frequencies that are preserved in the two blurred images. Thus, the DoG is a spatial band-pass filter that attenuates frequencies in the original grayscale image that are far from the band centre. The brilliance of the DoG algorithm lies in the subtraction step. By subtracting one blurred image from another, you preserve spatial information that lies between the range of

frequencies that were retained in the two blurred images. In other words, the DoG acts as a spatial band-pass filter. It enhances the features in the image that fall within a specific range of spatial frequencies, while attenuating or reducing information that exists at frequencies far from the band center.

3.3 Feature Description

In the feature description process, one essential step is assigning orientations to keywords or visual elements within an image and subsequently creating histograms based on the gradient direction. This technique is often referred to as Histogram of Oriented Gradients (HOG), and it plays a pivotal role in image feature extraction, particularly in computer vision and object detection tasks. To make the HOG descriptor robust to changes in illumination and contrast, the histograms within each block (a grouping of cells) are normalized.

3.4 Feature descriptor generation

The SIFT algorithm's final stage is to generate the descriptor, which is a normalized 128-dimensional vector. At this point in the algorithm, we are given a list of feature points with descriptions in terms of location, scale, and orientation. This allows us to build a local coordinate system around the feature point that should be consistent across multiple views of the same feature. The descriptor is a histogram formed from the grayscale image's gradient. A 44-dimensional spatial grid of gradient angle histograms is employed. The grid's dimensions are determined by the scale of the feature point, and the grid is centered on the feature point and rotated to the orientation determined for the keypoint. Each spatial bin contains an angle histogram divided into 8 segments. (128=4×4×8). The magnitude and angle of the image gradient are generated once more from the scale-space. The gradient orientations of sample points within a region around the keypoint are used to create an orientation histogram. The orientation histogram is divided into 36 bins that cover the entire 360-degree range of orientations. Each sample added to the histogram is weighted by the magnitude of its gradient as well as a Gaussian-weighted circular window with a that is 1.5 times the scale of the keypoint.

3.5 Feature Matching

Figure 6 Feature Matching- for each feature in A , Find nearest neighbour in B

SIFT feature matching in image stitching can be used to fully automate panorama reconstruction from non-panoramic images. The SIFT features extracted from the input images are compared to find the k nearest neighbors for each feature. Feature matching is the process of locating corresponding features in two similar images using a search distance algorithm. The feature matching technique is used to either find or derive and transfer attributes from the source to the target image, with one image serving as the source and the other as the target..

3.6 Captioning

Image caption generation is a process of recognizing the context of an image and annotating it with relevant caption. It includes the labelling of an image with English keywords with the help of datasets. It can be done on HSSHB Model Dataset. The dataset, consisting of images from HSSHB model documents and their corresponding captions, serves as the foundational resource. These images may contain historical illustrations, maps, diagrams, or other visual content.

4. EXPERIMENTS & RESULTS

4.1 Dataset

The authors of the study chose to use Higher Secondary School History Books (HSSHB) as a primary source for their research. The authors chose Higher Secondary School History Books (HSSHB) as a source for implementing our research. The dataset in this HSSHB Model is divided into two categories. The first category includes all images captured of Maharashtra state board assigned higher secondary school history book pages. Each page is considered

a document, which consists of text and images. We gathered over 300-page documents here. These documents are critical to the research because they are the primary source of textual and visual content. The second category of the dataset differs from the first in that it focuses on separated photos found within the documents. This category's dataset contains over 500 image images, each with its own caption. These images and captions are an important part of the research because they allow for the development and testing of various image processing, captioning, and recognition techniques.

4.2 Experimental Results

The research results indicate that the analysis of sample documents involves three key components: text recognition, image recognition, and the integration of these outputs with the HSSHB model. The findings of the study suggest that the examination of exemplar documents necessitates the participation of three fundamental elements: text recognition, image recognition, and the amalgamation of these results with the HSSHB model. The outcomes obtained from analyzing two sample images indicate that the output was divided into three distinct components: text recognition, image recognition, and output integration with our HSSHB model. Output 1 identifies the text contained within the sample document shown in figure 7, Output 2 (figure 8) identifies the images contained within the sample document, and Output3 (figure 9) provide total number of words & images contained within sample document.

Sample document 1 comprised 436 words and two images in total.

Sample document 2 comprises 538 words in total, with one image included.

The obtained results are saved in a designated folder as distinct output files in the English language. The results of the image and text recognition processes are saved in files that end in ".txt." The outcomes of the text recognition process are specifically stored in a file denoted as "MyFile.txt" within the module or system. The results of this study underscore the efficacy of the text recognition and image recognition methods utilized in the research. These processes significantly enhance the comprehension and analysis of the content contained in the sample documents.

The process concludes with the generation of an audio output suitable for listening on via headphones or speakers. The audio output presented here serves as the result of the aforementioned operations, which comprised text and image recognition, as well as the HSSHB model integration. The synthesized audio was produced by integrating information, recognized text, and captions. The purpose of this auditory feedback is to present the information contained in the images and documents in a manner that is straightforward for the user to comprehend. By providing an audio output that is operable through headphones or speakers, this system guarantees inclusivity and accessibility for people with diverse preferences and requirements. It facilitates effective user interaction with the Higher Secondary School History Books (HSSHB) for individuals with visual impairments or other disabilities.

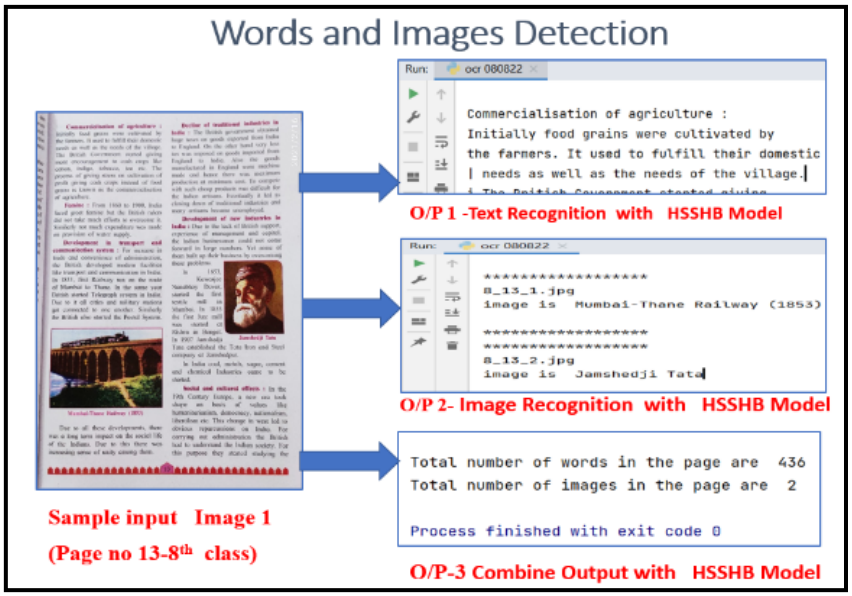

Figure 7 Input sample image 1 and its corresponding output results

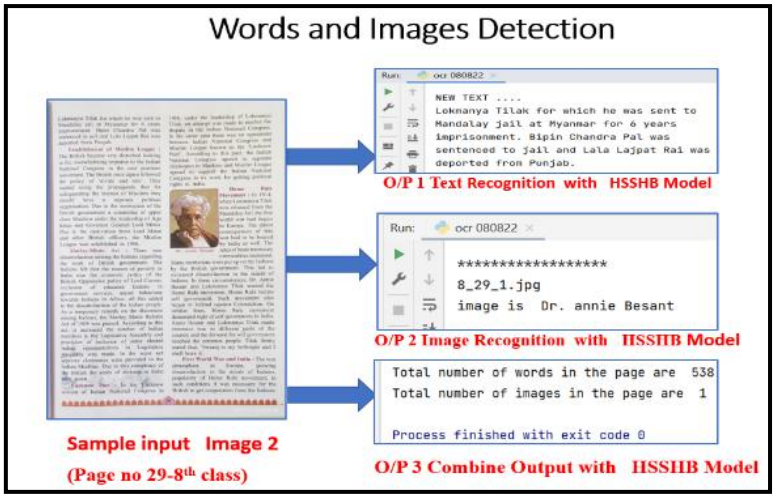

Figure 8 Input sample image 2 and its corresponding output results

Figure 9 Speech Output results of input samples

4.3 Performance Evaluation and result analysis

The performance of our HSSHB model can be evaluated with separate detection of word and images.

For detection of word

This performance evaluation for word detection is observed with graphical analysis of word (Text) detection results (fig 10) and With Quality check test for words. In fig 10, The line chart graph shows the comparision between manual word count and detected word count for each selected document page in book. Authors observe that in many cases we find approximately correct match of word count, which gives good accuracy.

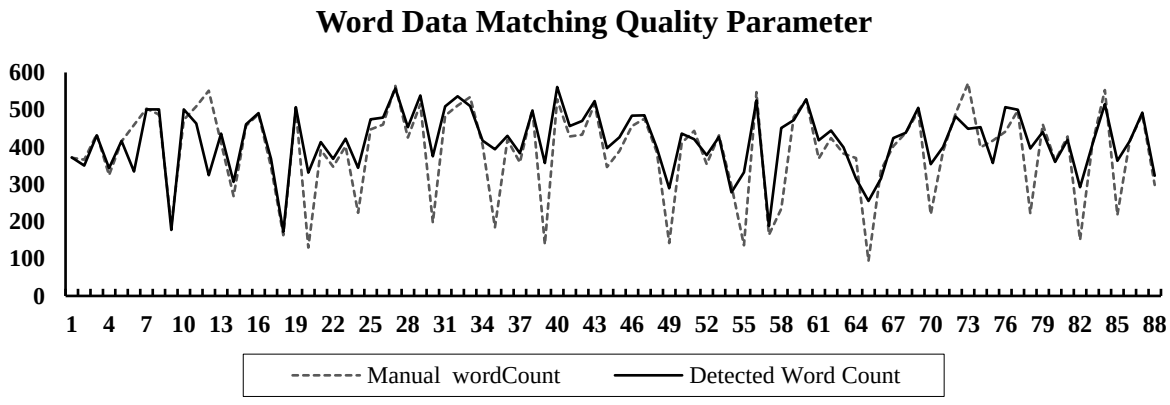

Figure 10 Performance evaluation Graph for Text (word) Detection

4.4 Quality Check Test for detection of words

Once the word detection process for a given document has been finalized, the performance is assessed using the Quality Check test for words. The results of this assessment are classified into three discrete quality grading categories: "Good," "Moderate," and "Worst." The determination of these categories is predicated on the proportion of errors detected in the word count, whereby each category signifies a distinct degree of precision and excellence. Defined category parameters are determined by the percentage of errors discovered in the word count. This classification comprises outcomes that have an error rate that is below 7%. It denotes a considerable degree of precision and accuracy in the process of word detection. The error rate for results falling within this category varies between 8% and 15%. This suggests a moderate degree of precision, albeit with potential for further enhancement. The category labeled "Worst" comprises outcomes that exhibit an error rate surpassing 15%. The evaluation of the word detection process was conducted on 88 documents, which constituted the complete dataset. A substantial majority of the 88 documents that were assessed are classified as "Good," with 72 documents (82%) falling into this category. Nine documents (10%) are categorized as Moderate, while seven documents (8%) are classified as Worst. A significant proportion of the documents attained a commendable degree of precision (categorized as "Good"), which signifies the system's ability to identify words with precision.

For Detection of images

The performance evaluation of image detection is conducted using an analysis of graphs representing the results of tested image detection (see Figure 11) and a quality check test for images. The line chart graph in Figure 11 illustrates the comparison between the manually counted and detected image counts for each selected document page in the book. It is observed that with good precision, an approximate correct match of the image count is obtained in the majority of cases.

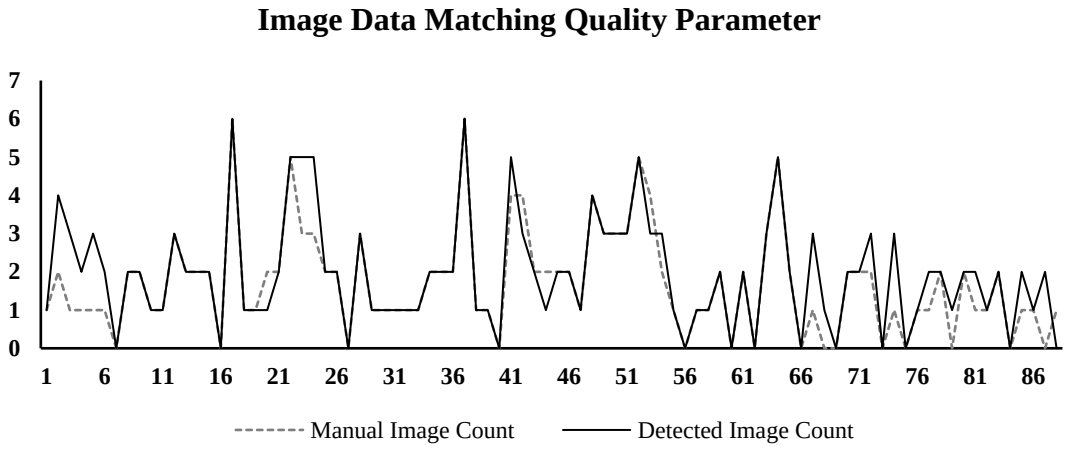

Figure 11 Performance evaluation Graph for Image Detection

4.6 Quality Check Test for detection of Images

Assessing the system's capability to accurately identify and count images within documents is a crucial aspect of the research, known as the evaluation of image detection performance. Authors assess the efficacy of image detection using a Quality Check test. The evaluation results for the test are categorized into three levels of data matching quality: Good, Moderate, and Worst. The parameters used to define these categories are determined by the discrepancies in the number of images between the manually and automatically detected image counts in documents. Image counts of 0, 1, and >1 are classified as Good, Moderate, and Worst, respectively. We conducted a comprehensive analysis on a total of 88 documents, examining their image detection capabilities. Our findings revealed that 69 documents (78%) fell into the Good category, 11 documents (13%) were classified as Moderate, and 8 documents (9%) were categorized as the worst.

4.7 Discussion

The evaluation of the proposed system on a dataset of Higher Secondary School History Books has yielded promising results, demonstrating the system's effectiveness in recognizing both text and images within documents. Authors has

been evaluated the proposed system on a dataset of higher secondary school history books. The system achieved approximately an accuracy of 92% in recognizing the text and 91% in detection of images from document with consideration of combination of Good and Moderate category results. The system was able to recognize and read printed text and images accurately and efficiently.

5. CONCLUSION AND FUTURE WORK

The proposed system offers a highly effective and accurate approach for visually impaired individuals to comprehend printed text, thereby enhancing their autonomy and overall well-being. The system utilizes OCR, SIFT algorithm, and Image captioning to identify and interpret printed documents containing both text and images. Image captions are substituted for images, and the combined text will be conveyed through audio feedback. The generated audio exhibits clarity and is readily comprehensible. The system demonstrated an impressive accuracy rate of around 92% in identifying printed text within the documents. This indicates a significant degree of accuracy and dependability in transcribing written content. Within the domain of image detection, the system exhibited a remarkable accuracy rate of approximately 91%. This level of accuracy includes the combined outcomes from the "Good" and "Moderate" classifications, emphasizing its expertise in recognizing and tallying images within documents. 82% of the documents were classified as "Good," demonstrating the system's high level of accuracy in detecting words. The research findings illustrate the system's capacity to improve accessibility and understanding for users, especially in educational settings.

Conflict of Interest

There is no conflict of interest from author's side.

Data availability statements

The data can be made available on the request to the corresponding authos.

Funding

No Funding has been received for this research work

REFERENCES

[1] Y. Zhu, C. Yao, X. Bai, Scene text detection and recognition: Recent advances and future trends, Front. Comp. Sci. 10 (1) (2016) 19–36

[2] Kuriakose, B., Shrestha, R., & Sandnes, F. E. (2022). Tools and technologies for blind and visually impaired navigation support: a review. *IETE Technical Review*, *39*(1), 3-18.

[3] R. Lienhart, Automatic text recognition in digital videos, in: Proceedings SPIE, Image and Video Processing IV, 1996, pp. 2666 –2675

[4] X.-C. Yin, Z.-Y. Zuo, S. Tian, C.-L. Liu, Text detection, tracking and recognition in video: a comprehensive survey, IEEE Trans. Image Process. 25 (6) (2016) 2752–2773

[5] A. Coates, B. Carpenter, C. Case, et al., Text detection and character recognition in scene images with unsupervised feature learning, International conference on document analysis and recognition 2011 (2011) 440–445.

[6] S. Lee, J.H. Kim, Integrating multiple character proposals for robust scene text extraction, Image Vis. Comput. 31 (11) (2013) 823–840.

[7] P. Shivakumara, S. Bhowmick, B. Su, et al., A new gradient based character segmentation method for video text recognition, International conference on document analysis and recognition 2011 (2011) 126–130.

[8] Reyes Leiva, K. M., Jaén-Vargas, M., Codina, B., & Serrano Olmedo, J. J. (2021). Inertial measurement unit sensors in assistive technologies for visually impaired people, a review. *Sensors*, *21*(14), 4767.

[9] Naz, S., Hayat, K., Razzak, M. I., Anwar, M. W., Madani, S. A., & Khan, S. U. (2014). The optical character recognition of Urdu-like cursive scripts. Pattern Recognition, 47(3), 1229-1248.

[10] Perera, M., Farook, C., & Madurapperuma, A. P. (2017, September). Automatic video descriptor for human action recognition. In *2017 National Information Technology Conference (NITC)* (pp. 61-67). IEEE.

[11] Musale, S., & Ghiye, V. R. (2018). Smart Reader for Visually Impaired. 2nd International Conference on Inventive Systems and Control. doi:10.1109/ICISC.2018.8399091

[12] Sabab, S. A., & Ashmafee, M. H. (2016, December). Blind reader: An intelligent assistant for blind. In *2016 19th International Conference on Computer and Information Technology (ICCIT)* (pp. 229-234). IEEE.

[13] Khan, S., Nazir, S., & Khan, H. U. (2021). Analysis of navigation assistants for blind and visually impaired people: A systematic review. *IEEE access*, *9*, 26712-26734.

[14] Ashiq, F., Asif, M., Ahmad, M. B., Zafar, S., Masood, K., Mahmood, T., ... & Lee, I. H. (2022). CNN-based object recognition and tracking system to assist visually impaired people. *IEEE Access*, *10*, 14819-14834.

[15] Masud, U., Saeed, T., Malaikah, H. M., Islam, F. U., & Abbas, G. (2022). Smart assistive system for visually impaired people obstruction avoidance through object detection and classification. *IEEE Access*, *10*, 13428-13441.

[16] Rane, C., Lashkare, A., Karande, A., & Rao, Y. S. (2021, June). Image captioning based smart navigation system for visually impaired. In *2021 International Conference on Communication information and Computing Technology (ICCICT)* (pp. 1-5). IEEE.

[17] Bhalla, A., Goutham, S., Prakash, K., & Sanjana, T. (2021, December). VIEW: Optimization of Image Captioning and Facial Recognition on Embedded Systems to Aid the Visually Impaired. In *2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)* (pp. 1-6). IEEE.

[18] Islam, A., Tee, M. K. T., Lau, B. T., & Chong Foh-Zin, K. (2023, October). A Preliminary Study on the Possibility of Scene Captioning Model Integration as an Improvement in Assisted Navigation for Visually Impaired Users. In *Asia Simulation Conference* (pp. 352-361). Singapore: Springer Nature Singapore.

[19] Kornsingha, T., & Punyathep, P. (2011, May). A voice system, reading medicament label for visually impaired people. In RFID SysTech 2011 7th European Workshop on Smart Objects: Systems, Technologies, and Applications (pp. 1-6). VDE.

[20] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*, 91-110.

[21] Asad, S., Mooney, B., Ahmad, I., Huber, M., & Clark, A. (2020, June). Object detection and sensory feedback techniques in building smart cane for the visually impaired: An overview. In *Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 1-7).

[22] Paiva, S., & Gupta, N. (2020). Technologies and systems to improve mobility of visually impaired people: A state of the art. *Technological Trends in Improved Mobility of the Visually Impaired*, 105-123.