Narrating the Unseen: Real-Time Video Descriptions for Visually Impaired Individuals

Nelson Joseph
College of Engineering, The University of Texas at Arlington, Arlington, Texas
Email: nelsonjoseph123@gmail.com

***Abstract*— This research paper explores a novel system designed to empower visually impaired individuals by narrating their surroundings through spoken language, leveraging the capabilities of a mobile camera. Our study involves a comparative analysis of various pre-trained models in generating descriptive captions. Globally, approximately 2.2 billion people are affected by some form of visual impairment or blindness. Addressing this significant challenge, our research proposes an integrated solution aimed at assisting visually impaired individuals in comprehending their environment. This is achieved through the description of video streams, utilizing advanced Generative AI techniques.**

The cornerstone of our proposed methodology is the use of a pre-trained GPT-4 Vision multimodal model, which has been trained on an extensive dataset comprising 13 million tokens. Additionally, we have engineered a robust Client-Server socket connection framework. This design ensures that intensive computational tasks, particularly video stream preprocessing, are primarily conducted server-side.

A key aspect of our research involves the evaluation of generated captions. These are meticulously compared with standard captions using established metrics such as BLEU and ROUGE scores. Recognizing the semantic limitations inherent in these metrics, we also employ a Semantic Similarity metric for a more nuanced comparison. This comprehensive approach allows for a thorough assessment of the effectiveness of our system in providing accurate and contextually relevant descriptions for the visually impaired.

Keywords— GPT4Vision, VIT, LSTM, RNN, Pipeline

I. INTRODUCTION

In an increasingly visual world, the ability to perceive and interpret one's surroundings is integral to daily life. However, this reality presents a significant challenge for the visually impaired community, which comprises approximately 2.2 billion individuals worldwide as of 2023. The advent of technology, particularly in the realm of artificial intelligence and machine learning, offers unprecedented opportunities to bridge this sensory gap. This paper presents an innovative solution aimed at translating the visual experiences of the environment into auditory information accessible to visually impaired individuals.

One of the critical challenges in developing such a system is the inherent complexity of camera feeds, which often include varying lighting conditions, different angles, motion blur, and real-time changes. These factors are not typically addressed in the training datasets of most image captioning models. Existing datasets such as Flickr 8k [10], Flickr 30k, MS COCO, and Conceptual Captions are limited to a certain number of classes and do not encompass the vast array of objects encountered in the real world. Consequently, models trained on these datasets often produce descriptions that are narrowly focused and not reflective of the actual scene.

To address these challenges, our approach leverages the power of large-scale generative language models, which have been trained on extensive and diverse datasets. The core of our solution is the utilization of a pre-trained GPT-4 Vision multimodal model, trained on a comprehensive dataset of 13 million tokens. This model can generate accurate and contextually relevant descriptions of video streams captured through a mobile camera, significantly enhancing the autonomy of visually impaired individuals and enriching their understanding and interaction with their surroundings.

Our research involves a comparative analysis of various pre-trained models in their efficacy to generate meaningful captions. A novel aspect of our approach is the client-server architecture, where intensive computational processes, especially video stream preprocessing, are offloaded to the server side. This design ensures the efficiency and scalability of our solution.

Furthermore, the paper delves into the intricacies of evaluating the effectiveness of these generated captions. We recognize the limitations of traditional metrics such as BLEU and ROUGE scores in capturing semantic nuances, and therefore, our research also incorporates a Semantic Similarity metric. This comprehensive evaluation framework is pivotal in assessing the real-world applicability of our proposed system.

The significance of this research extends beyond technological innovation; it embodies a step towards creating an inclusive society where visual impairments do not impede access to information and interaction with the environment. As we advance, the integration of AI in assistive technologies heralds a new era of possibilities, bringing us closer to a world where the unseen becomes narrated, and the invisible becomes perceptible for all.

II. RELATED WORK

The field of image captioning has seen significant advancements in recent years, with the development of several models that have set benchmarks in both accuracy and efficiency. These models, integral to numerous applications, have particularly found profound use in the realm of assistive technologies for

visually impaired individuals. In this context, our review of related works is divided into two distinct sections: the advancements in image captioning models and their applications in artificial intelligence for the visually impaired.

A. Related works in the Image Captioning area.

1. Introduction to Image Captioning Models

This section explores the concept of image captioning within the realm of artificial intelligence. Image captioning, pivotal in both academic research and its practical applications, plays a critical role in translating visual content into descriptive text. Its significance is particularly pronounced in the development of assistive technologies, offering substantial benefits to the visually impaired. By converting images into verbal descriptions, these technologies enable visually impaired individuals to gain a better understanding of their surroundings.

2. Foundational Models and Early Developments

In the early stages of image captioning, foundational models laid the groundwork for what has become a dynamic and evolving field. These models, based on basic principles of computer vision and natural language processing, provided the initial framework for interpreting and describing visual data. However, these early models faced several limitations, including rudimentary object recognition and simplistic descriptive capabilities, which restricted their use in complex real-world scenarios.

3. Progression to Advanced Models: Show and Tell, BLIP, et al.

The field of image captioning experienced significant advancements with the development of models such as Show and Tell [1] and BLIP [12]. These advanced models incorporated more sophisticated algorithms and leveraged larger datasets, leading to notable improvements in accuracy and contextual relevance. Their methodologies, which combined deep learning techniques with innovative language models, represented a substantial leap forward in the field, enabling more nuanced and detailed descriptions of images.

4. Common Challenges in Image Captioning

Despite these advancements, image captioning continues to face several challenges. Key among these are the complexities associated with interpreting intricate scenes, the variability in object representation, and ensuring contextual accuracy. These challenges highlight the ongoing need for further research and development to enhance the reliability and applicability of image captioning models, especially in dynamic and unpredictable environments.

5. Application of Image Captioning in Real-World Scenarios

The adaptation of image captioning models for real-world applications, particularly in assistive technologies for the visually impaired, has seen considerable progress. These technologies have evolved to not only identify objects but also provide detailed information such as the name, distance, and direction relative to the user. The integration of depth sensing and accurate spatial awareness has enabled visually impaired individuals to navigate indoor environments more effectively. Tests have demonstrated that these technologies can adeptly navigate and overcome obstacles, offering a greater degree of freedom and independence to users.

B. Related works associated for visually impaired people

1. Integration of AI in Assistive Technologies

This section delves into the integration of artificial intelligence in assistive technologies, with a focus on innovations specifically designed for visually impaired users. AI has revolutionized assistive technologies, enabling more sophisticated and user-friendly solutions. The incorporation of AI facilitates more accurate interpretations of the physical world, enhancing the independence and quality of life for visually impaired individuals. These technologies range from simple object detection systems to complex navigational aids, demonstrating the broad potential of AI in this field.

2. Techniques in Image/Video Summarization for the Visually Impaired

A comprehensive review of image and video summarization technologies tailored for visually impaired users is presented. This segment highlights various methodologies, particularly object detection models, and their application in providing descriptive scene analysis. These technologies have evolved to not only recognize objects but also provide contextual information about the environment, significantly aiding visually impaired users in understanding their surroundings.

3. Addressing the Continuity Challenge in Scene Description

A significant gap identified in current technologies is the lack of continuity in scene descriptions across dynamic video frames. Our research paper introduces a novel model designed to address this challenge. This model ensures consistent narrative context between consecutive video frames, thereby providing a coherent and continuous understanding of dynamic scenes. This feature is particularly crucial in real-world scenarios where environments are constantly changing, and the need for accurate, real-time descriptions is paramount.

4. The Significance and Future Prospects of Continuous Scene Description

The final section discusses the potential impacts and prospects of continuous scene description in assistive technologies for the visually impaired. Improved continuity in scene descriptions not only enhances the current state of assistive technology but also opens new avenues for research and development. The paper speculates on future advancements in AI-assisted technologies, emphasizing the need for ongoing innovation to further aid the visually impaired community. The integration of more sophisticated AI models increased contextual understanding, and improved user interfaces are identified as key areas for future development.

III. RESEARCH PROBLEM

Recent technological developments have focused on creating practical systems to improve living standards. In order

to help visually impaired people recognize the surrounding, we can use Generative AI capabilities. These people use their senses to learn about their surroundings by touching, hearing, and smelling, but they are limited in how much of their environment they can comprehend because they cannot use their eyes. Therefore, we decided to design an application called "Narrating the Unseen: Real-Time Video Descriptions for Visually Impaired Individuals" to assist them so that they can get a real-time information about their surroundings.

A. Contributions of this paper.

In recent developments, a plethora of electronic aids, including modified walking sticks and various other devices, have been introduced to support those with visual impairments. These existing systems frequently depend on supplementary external hardware and leverage techniques such as image and video summarization, which are rooted in object detection algorithms, to process and extract information from imagery captured through cameras. Diverging from these traditional approaches, our innovative model presents a groundbreaking solution. It is specifically engineered to deliver instantaneous descriptive feedback, eliminating the need for any additional external devices. This breakthrough approach in our model underscores a significant enhancement in assistive technology, offering real-time, efficient assistance that fosters greater autonomy and improved accessibility for individuals with visual impairments.

- Use of Smartphones: Smartphones are equipped with integrated sensors, cameras, and speakers, resulting in a significantly reduced cost.
- Frame-coherence: There is a continuous connection between the descriptions of successive frames.

IV. Framework and System Design

System design involves creating the architecture, components, and interfaces of a system to meet user requirements. Figure 1 presents the proposed model, delineating its distinct phases. The functionality of each phase is detailed in the following sections.

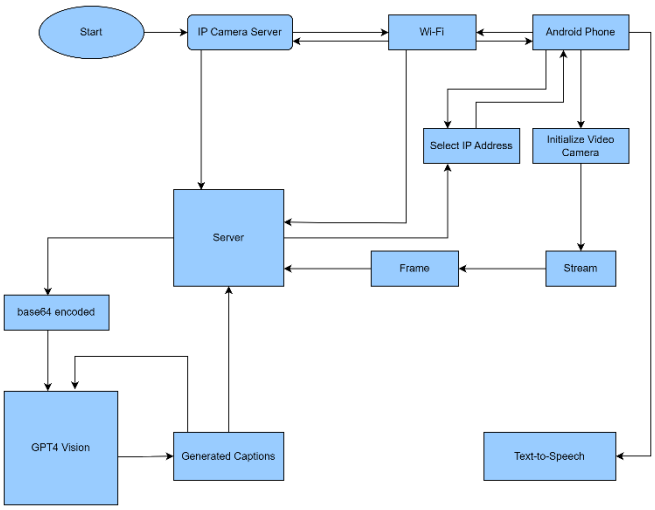

Figure 1 System Design of Proposed Model

- IP Camera Server: This is where the system initializes. We need to start a server in the IP Camera mobile app.
- Server: Here the Computer or a pc act as a local server which does all the processing of video feed and acts as a Hub of the system.
- Wi-Fi: The android device and the Server should be connected to the same Wi-Fi to make the socket connection between the user's phone and the server.
- Android Phone: After the descriptions of the images are generated from the video feed from the android phone. The description is sent back to the phone to be played aloud via socket connections.
- Select IP Address: Both the Android phone and the Computer should be connected to Same IP Address.
- Initialize Video Camera: The video camera will be initialized when we start the IP camera server.
- GPT4 Vision Model: It is a large multimodal model (accepting image and text inputs, emitting text outputs) trained on 13 million tokens.
- Base64 encoded: The frames from the video feed should be converted to base64 encoded format to pass through the GPT4 Vision model to generate captions. Base64-encoded encodes the image data into a string of ASCII characters.
- Generate Captions: The captions are generated using the base64 encoded frames from the video feed.
- Stream: The video feed is streamed from the Android phone using the IP camera server when both the computer and the android phone are connected to the same IP Address.
- Frame: These are the frames extracted from the video feed. The model keeps track of the captions generated from the previous, current frame to create a continuous flow of descriptions.
- Text to Speech: The textual captions obtained from the descriptions are spoken out with the help of Google Text to speech and OpenAI TTS-1-HD model.

V. Results and Discussion

We employ three distinct types of pre-trained model setups in this study. These systems include:

1) Hugging Face image-to-text Pipeline

The ImageToTextPipeline from Hugging Face is engineered to produce textual descriptions from input images. This pipeline finds its utility in tasks such as image captioning and optical character recognition (OCR). It integrates a suite of models, encompassing the Vision Transformer (ViT) for image analysis and a text generation model, like GPT-2, for synthesizing text.

2) ViT Encoder - Decoder Model

This model represents a specific type of Vision-Encoder-Decoder architecture, tailored explicitly for image captioning. It incorporates a Vision Transformer (ViT) as the encoder and utilizes GPT-2 as the decoder. While this model is a particular example within the expansive category of encoder-decoder models designed for image-to-text conversions, like the ImageToTextPipeline previously discussed, it distinguishes itself through its unique and specialized configuration.

3) GPT4 Vision Transformer

GPT-4 with Vision, often referred to as GPT-4V or 'gpt-4-vision-preview' in API contexts, enables the model to process images and respond to queries related to them. Traditionally, language model systems have been constrained to a single modality of input, namely text.

The pre-trained models under consideration have been trained on diverse datasets, resulting in the absence of a uniform dataset for model performance evaluation. To address this, we selected a subset of images from the Flickr-8K [10] dataset and generated captions using various models. We then assessed these models using the BLEU and ROUGE scores based on their caption predictions. However, given that each language model processes predictions differently, relying solely on BLEU and ROUGE scores, which primarily focus on word count rather than semantic similarity, may not suffice for comprehensive comparison. Therefore, we are employing GPT-4 to evaluate the similarity between the predicted and actual captions. This approach involves assigning a similarity score on a scale of 1 to 10, thereby providing a more nuanced measure of model performance.

TABLE I. COMPARISON OF DIFFERENT MODELS ARCHITECTURE

No	Comparison			
	Different Model Architecture	***BlEU***	***ROUGE***	***Semantic Similarity***
1	Hugging Face Image-to-text	0.23	0.22	4.8
2	ViT Encoder - Decoder Model	0.22	0.18	5
3	GPT4 Vision Transformer	0.36	0.34	8.5

Here are some of the results that are obtained from the real-time captions of the GPT4 Vision Transformer.

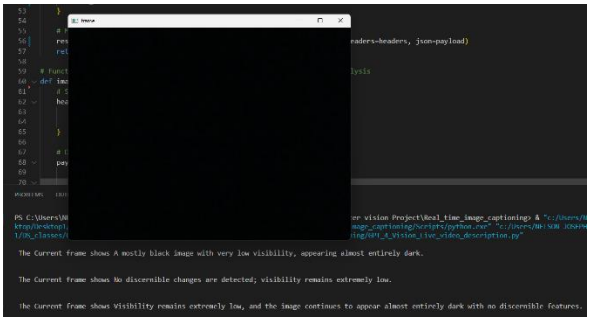

Figure 2 Blank image on the camera is being continuously detected.

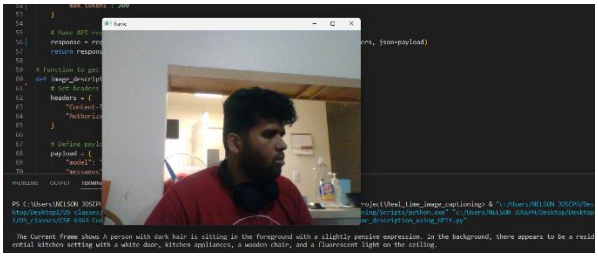

Figure 3 A person in red t-shirt and the surrounding is being - detected.

VI. CONCLUSION AND FUTURE SCOPE

In this research, we presented a novel system designed to assist visually impaired individuals by enhancing their perception of their environment. This system focuses on multimodal features of the GPT4 Vision, converting images into descriptive text and speech to facilitate the surroundings. Such technology not only clarifies what has been achieved in supporting the visually impaired but also significantly improves their quality of life. It achieves this by employing image and video summarization applications, effectively substituting sight in environments of visual limitation.

Looking ahead, our future work will leverage the multimodal capabilities of Google DeepMind's Gemini framework to develop an interactive camera application. This app is envisioned to allow users to interact with a virtual assistant, asking questions and receiving real-time responses, thereby making their navigation through unfamiliar environments safer and more intuitive.

There is ample room for enhancement in the system design. A key objective is to ensure the application's complete functionality within a smartphone, eliminating the need for external devices or interactions. Additionally, personalizing the assistant based on the user's familiar surroundings could further facilitate safer and more efficient navigation.

REFERENCES

[1] OriolVinyals, AlexanderToshev, SamyBengio, DumitruErhan. ShowandTell:ANeuralImageCaptionGenerator. arXiv:1411.4555v2 [cs.CV] 20 Apr 2015.

[2] Alexey Dosovitskiy , Lucas Beyer , Alexander Kolesnikov , Dirk Weissenborn , Xiaohua Zhai , Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE. arXiv:2010.11929v2 [cs.CV] 3 Jun 2021

[3] Md Milon Islam, Muhammad Sheikh Sadi, Kamal Z Zamli, and MdManjur Ahmed. Developing walking assistants for visually impaired people: A review. IEEE Sensors Journal, 19(8):2814–2828, 2019.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.K. Elissa, “Title of paper if known,” unpublished.

[5] Sunit Vaidya, Naisha Shah, Niti Shah, and Radha Shankarmani. Real time object detection for visually challenged people. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pages 311–316. IEEE, 2020.

[6] Sadia Zafar, Muhammad Asif, Maaz Bin Ahmad, Taher M Ghazal, Tauqeer Faiz, Munir Ahmad, and Muhammad Adnan Khan. Assistive

devices analysis for visually impaired persons: a review on taxonomy. IEEE Access, 2022.

[7] Jawaid Nasreen, Warsi Arif, Asad Ali Shaikh, Yahya Muhammad, and Monaisha Abdullah. Object detection and narrator for visually impaired people. In 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pages 1–4. IEEE, 2019.

[8] Chaitra C1, Chennamma1, Vethanayagi R1, Manoj Kumar M V1, Prashanth B S1, Snehah H R1, Likewin Thomas3, Shiva Darshan S L2. Image/Video Summarization in Text/Speech for Visually Impaired People. 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon).

[9] Sahil Takkar, Anshul Jain, Piyush Adlakha. Comparative Study of Different Image Captioning Models. Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021) IEEE Xplore Part Number: CFP21K25-ART

[10] "Flickr 8k Data | Illinois." https://forms.illinois.edu/sec/1713398 (accessed Mar. 05, 2021).

[11] Derek Tam, Colin Raffel, Mohit Bansal. Simple Weakly-Supervised Image Captioning via CLIP's Multimodal Embeddings. UNCChapel Hill {dtredsox, craffel, mbansal}@cs.unc.edu

[12] Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597v3 [cs.CV] 15 Jun 2023