WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

SpeakSight

Lakshmana Kannan K[1], Mithran B[2], Pradeep M K[3], Sachin M[4], Dr. S. Prakash[5]

Sri Shakthi Intitute Of Engineering And Technology,Coimbatore

A B S T R A C T :

SpeakSight is a assistive AI platform that makes visual content more accessible by transforming images into sense-making auditory descriptions. Drawing on computer vision and natural language processing, the platform examines visual content—recognizing objects, contexts, emotions, and spatial relationships—and produces accurate, descriptive captions. The captions are then translated into speech using text-to-speech technology, enabling blind and visually impaired individuals to see and envision visual content through sound. SpeakSight is an example of the social value of AI multimodal technologies in bridging digital gaps to foster digital inclusion, to support navigation, education, media accessibility, and assistive communication.

1.INTRODUCTION

SpeakSight is a pioneering assistive technology created to empower visually impaired users by allowing them to comprehend and visualize the content of images in terms of audio descriptions. The project integrates the best of computer vision, natural language processing (NLP), and text-to-speech (TTS) synthesis in order to achieve a smooth user experience that converts visual information into speech. The central operation is processing an uploaded picture, creating a descriptive caption based on sophisticated deep learning models, and reading it out to the user. This multidisciplinary framework is designed to meet the essential accessibility barrier for the blind community to access visual content in digital media. Building on the development of image captioning and generative AI systems from classic convolutional neural networks (CNNs) to current transformer-based models. SpeakSight demonstrates the power of AI for social good. Not only does SpeakSight increase accessibility but also sets the stage for more accessible human-computer interaction, making visual content accessible by means of sound.

2. REVIEW OF LITERATURE

The technological basis of SpeakSight is based on the development of artificial intelligence models, from initial generative adversarial networks (GANs) to current transformer-based architectures such as DALL-E, CLIP, and Vision Transformers. These developments have made tremendous strides in creating descriptive captions from images, a task essential for accessibility tools.

Visual feature extraction methods like YOLO, ResNet, and ViT (Vision Transformers) are employed to identify and analyse visual elements such as objects, actions, and scenes. Semantic labelling applies meaningful textual labels (e.g., "a woman riding a bicycle on a street") based on pre-trained classification and segmentation models. This is followed by Natural Language Generation (NLG), where models such as BERT or GPT generate coherent, context-sensitive, and emotionally engaging captions from the visual information.

To guarantee access for blind users, user feedback incorporation is utilized to enhance descriptions and contextualize them more accurately. In addition, multimodal alignment models such as CLIP assist in associating image features with descriptive text in a common embedding space, guaranteeing semantic precision and consistency. SpeakSight, by virtue of its multi-layered design, interprets silent images into rich soundscapes, assisting visually impaired users in "seeing" through sound.

3. FIELD OF THE INVENTION

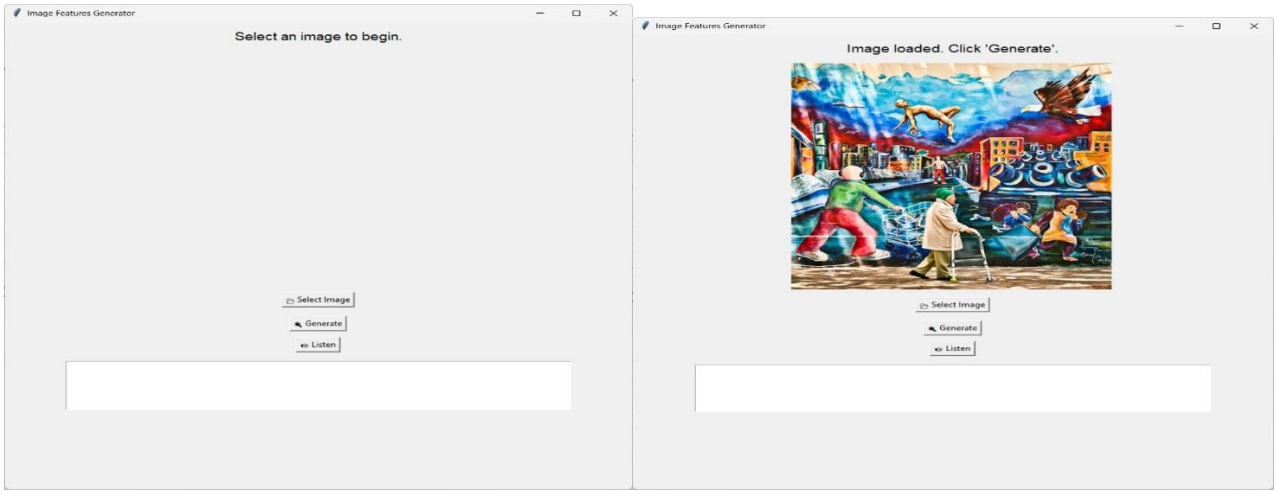
The invention falls within the assistive technologies category, specifically those that promote visually impaired people's accessibility. It is more specifically about an automatic image interpretation and auditory description system and method. Through computer vision, natural language processing, and text-to-speech synthesis, the invention provides users who are blind or have low vision with access to visual content through auditory feedback. This invention intersects at artificial intelligence, human-computer interaction, and inclusive design with the objective to fill the gap of information between visual media and visually impaired consumers.

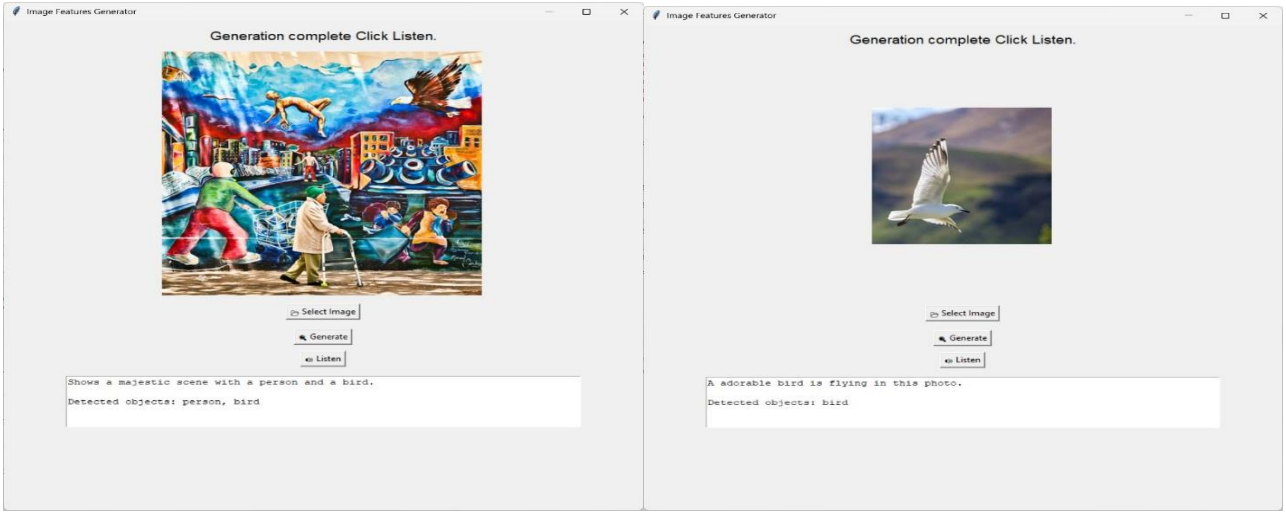
SOFTWARE DESCRIPTION

- PYTHON (Tensorflow , Pickle , NumPy , tqdm , OpenCV)
- Tkinter

- Flickr30k (Data set)

SCREENSHOTS





CONCLUSION

SpeakSight is a new technology that uses computer vision, natural language processing, and text-to-speech to augment visual accessibility. By transforming images into faithful and descriptive audio descriptions, SpeakSight allows the blind and visually impaired to comprehend and visualize visual information through sound. It enhances their independence and facilitates digital inclusion. As an important advance in the application of artificial intelligence for the benefit of society, SpeakSight solves an important accessibility problem. With ongoing development and feedback from users, it has the potential for more extensive applications in education, wayfinding, media use, and assistive communication.

Acknowledgements

I would like to express my sincere gratitude to my mentor, the department staff, and the Head of Department (HoD) for their invaluable guidance, support, and encouragement throughout the course of this project. Their expertise and constant assistance have been instrumental in the successful completion of this work.

REFERENCES

1. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
2. Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR), 51(6), 1–36.
3. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). arXiv preprint.
4. Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going deeper with convolutions (GoogleNet). In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 1–9.
5. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
6. Microsoft Azure, IBM Watson, and Google Cloud APIs – Documentation for text-to-speech and computer vision services.
7. World Health Organization (WHO). World report on vision. (2019).