International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI-2025), January 16-18, 2025

Visual Understanding and Navigation for the Visually Impaired using Image Captioning

Jitesh Uikey
Department of Information Technology
Vishwakarma Institute Of Information Technology
Pune, India
jitesh.22210369@viit.ac.in

Ishan Undre
Department of Information Technology
Vishwakarma Institute Of Information Technology
Pune, India
ishan.22210696@viit.ac.in

Urvish Vasani
Department of Information Technology
Vishwakarma Institute Of Information Technology
Pune, India
urvish.22210457@viit.ac.in

Aman Shaikh
Department of Information Technology
Vishwakarma Institute Of Information Technology
Pune, India
aman.22210651@viit.ac.in

Riddhi Mirajkar
Department of Information Technology
Vishwakarma Institute Of Information Technology
Pune, India
riddhi.mirajkar@viit.ac.in

Jayashree Bagade
Department of Information Technology
Vishwakarma Institute Of Technology
Pune, India
jayashree.bagade@vit.edu

***Abstract*— Visually impaired people face problems with independent navigation due to limited visual information. Facing significant challenges, they often travel with an assistant or their relative. This project introduces an approach, increasing the independence of a blind user by developing new software solutions. The proposed system employs DenseNet201 for feature extraction and Long Short-Term Memory (LSTM) networks for generating accurate, context-aware captions. These captions are converted into real-time auditory descriptions using the gTTS library, enabling users to interpret and navigate their environment confidently. Evaluated on the Flickr8k dataset, the system achieved a BLEU score of 0.721, demonstrating its ability to generate high-quality captions. The system's architecture is designed to balance accuracy, efficiency, and user accessibility. Additionally, the system incorporates a modular architecture optimized for computational efficiency and scalability. Future work includes exploring wearable technology for continuous real-time feedback, integrating advanced natural language processing (NLP) models for richer contextual understanding, and enhancing its applicability in complex indoor and outdoor environments. This approach represents a significant step toward empowering visually impaired individuals with improved mobility and environmental awareness.**

***Index Terms* - Visually Impaired, Deep learning, Neural Network, Convolutional Neural Network (CNN), DenseNet201, Long-Short Term Memory (LSTM), gTTS, Image Classification, Object Detection, Image Captioning, Flickr-8k, Natural Language Processing (NLP)**

I. INTRODUCTION

Visually impaired people can only sense light and shadows. They cannot see objects in front of them. Furthermore, they move around based on their senses and experiences [1]. The visually challenged depended on simple aids since ancient times to move around the community and read all sorts of written content. Louis Braille invented the Braille alphabet in 1824 as a kind of tactile literacy method that transformed millions of lives into having accessible forms of writing and reading. It is a collection of raised dots representing letters through which readers read with their sense of touch. In a crowded country in Asia or Africa, it becomes quite difficult for visually impaired people to travel from one place to another by just using a traditional white cane [2]. The white cane provides some essential navigational assistance to them by inspecting possible obstacles. It can just be a simple but crucial tool in the moving individual's process of navigation. This meant that technological innovations of the 20th century provided alternative ways through which partially sighted people could receive information and entertainment. The breakthrough was the talking book; that is, literature and educational material came in audio instead of print. There were the screen readers, which would read on-screen text aloud or convert it to Braille, thereby opening up computer and early electronic device-based digital content to the visually impaired. These were the earlier technologies, though such massive strides possessed some limitations regarding interactivity and real-time information processing.

Modern Braille displays, therefore, provide real-time Braille translations of screen text and promote the accessibility of digital content for the blind [3]. They are also often connected to computers or mobile devices using either USB or Bluetooth. This innovation allows blind users to view digital content at an equal pace and scale as that of a visually enabled counterpart.

The other very innovation is OCR software, which enables a user to transform physical printed documents into digital text [4]. OCR software does this by using the strength of powerful image processing and its subsets, called machine

979-8-3315-1852-3/25/$31.00 ©2025 IEEE

learning models, which can pluck characters from scanned images or pictures. This digital text may be read to the user or rendered in Braille, thus allowing visually impaired users greater access to a wide range of documents regardless of size or special format.

Research on assistive technology has benefited from such advances in creating visual substitution for visual impairment [5]. These days, a vast number of modern applications rely upon artificial intelligence (AI) and machine learning (ML) for real-time assistance for a lot of vision-impaired people. Apps like "Seeing AI" use AI-based image recognition regarding descriptive words of objects, reading text, and even facial recognition. Apps apply deep learning models to detect objects and NLP to understand contextual meanings to the users.

Smart glasses are a potential assistive technology for blind and visually impaired people to aid in individual travel and provide social comfort and safety [6]. Such glasses will recognize objects, read text, and also interpret the surroundings of their wearers in real time. Advanced computer vision algorithms in such glasses provide audio descriptions through an earpiece, thereby making users "see" through auditory feedback.

Artificial intelligence has unveiled new facets in helping the visually impaired perceive and understand information from images, natural language, and machine learning. AI models like CNNs recognize the objects and scenes in visual inputs, NLP models can therefore write meaningful descriptions of the environment. These technologies combine to help in creating assistive devices, feeding back information into users in real-time through text-to-speech systems, making navigation as well as understanding of visual information significantly easier. This structure effectively introduces the progression from basic devices to modern AI solutions.

II. Relates Work

Paper reviewed recent developments in applying artificial intelligence, deep learning in particular, to improve the life of the visually challenged. It finds applications in the two spheres: diagnosis of eye diseases as well as development of intelligent visual aids. Early and accurate eye disease diagnosis allows for the treatment that will enhance the outcomes of patients. Authors elaborated explosion of deep learning-powered smart devices in assistive technology towards helping a visually impaired person to accomplish routine tasks [7].

In this paper, authors judged realistic life computer intervention based on machine learning algorithms in PubMed and Scopus. In general, the number of participants in the considered interventions averaged 71, while their follow-up time was reported in days ranged between 3 to 180 days. Health outcomes were obtained in 75% of the considered interventions with statistical significance. However, most of these studies have shortcomings since they have not tested the machine learning methodologies in adequate levels; therefore, rigorous study should be performed to establish clinical potential in healthcare [8].

Paper highlights the significant role of machine learning in computer vision, particularly in biomedical engineering. Focusing on deep learning strategies was a key point, which falls under classification and clustering techniques—these being two broad categories of supervised and unsupervised learning in machine learning tasks. The paper focused more on biomedical data analysis [9].

Authors presented a wearable application designed to provide a visually impaired person with the ability to navigate in an indoor environment. The system will primarily consist of an RGB-D camera and IMU. These will detect objects as well as safe paths through which one could traverse [10].

In the paper, authors presented a new deep architecture along with an encoder-decoder framework where the image captioning incorporates CNN features used from a pre-trained Xception model on ImageNet combined with object features coming from a pre-trained YOLOv4 model pre-trained on MS COCO. The method also introduces a new kind of positional encoding called the "importance factor". While doing so, the object detection features enhance the quality of the captions generated by the model. It was evaluated using the MS COCO and Flickr30k datasets. The huge performance improvements gave the system a 15.04% gain in the CIDEr score compared to similar methods [11].

Paper [12] explored the integration of computer vision with NLP to caption images in Hindi, using a CNN-LSTM neural network model. The performance was perfected by training different models with varying hyperparameters and hidden layers. With extreme performance improvement in results, the BLEU score has gone up to 34.64% (Unigram) and also rose up to 29.13% in the case of the BLEU score (Bigram) when compared to previous work. The model has a wide scope for broad social benefits, especially among Hindi-speaking users.

The paper introduced a description-generating image captioning model that creates descriptive captions and aggregates text extracted from images to better the description accuracy of the visually impaired. Their model uses combinations of CNNs for extraction of features from images and LSTM in sentence generation based on features learned. Capturing the text contained in the image and appending it to the caption, which is also available in audio form. Since most of the standard datasets, such as MS COCO and Flickr contain fewer images with text content, authors have established a new dataset for this task. Therefore, this model can be said to be not only as good as the existing models but also better insights into the matter with the inclusion of text extraction [13].

Authors gave a novel contribution to multimodal learning. The paper proposes an encoder-decoder pipeline unifying joint image-text embeddings with neural language models. It creates a multimodal joint embedding space for images and text and introduces a novel structure-content neural language model that decouples sentence structure from content, conditioned on the encoder's representations. The encoder allows ranking of images and sentences while the decoder generates new descriptions from scratch. It achieves the best

performance on Flickr8K and Flickr30K datasets when using LSTM to encode sentences, while improved results with the 19-layer Oxford convolutional network. Besides, the learned embedding space has multimodal regularities such as vector arithmetic with image features and text. Sample captions for 800 images are made available for comparison [14].

The paper suggested an association Long Short-Term Memory network framework specifically for video object detection and thus tackles the challenge of object association between consecutive frames. Unlike the traditional image-based object detection approaches, which do not know how to exploit temporal information, the model in this paper directly regresses object locations and categories and produces association features to track objects across frames. Aggregating with the decrease in object regression and association errors, the proposed technique yields online videos that are more accurate at video object detection in real-time. Association LSTM performs better than the conventional methods in a set of video datasets. The feature updates are not given as a feedback from LSTM outputs; however, future work is targeted to include the feedback loop in optimally extracting the feature by considering the timely context followed by modules of RNN [15].

The paper proposed facial recognition system using VGG16 model to achieve the high accuracy. Authors utilized 5-step process data collection, cleaning, fine-tuning, training, and evaluation. The system can effectively identify individuals in various applications, including surveillance, access control, and law enforcement [16].

III. Proposed System

A. Data Collection and Pre-Processing

In the experiment, we used Flickr8k datasets [17]. These datasets contain 8,000 images. For all images, there are five descriptive captions. It is very famous in the field of image captioning because it normally contains any variety of scenes and objects or activities [18]. This is a natural description of each image, containing specific details about the content of the images-thus ideal for deep models to use this in training and get their own descriptive captions. The dataset is publicly available.

The captions preprocessing included many important steps that helped the model learn from the data. We ensured that all captions get converted into lowercase so that it becomes uniform. Using regular expressions, we took out all the special characters and unnecessary spaces in the text so that it becomes more simplified. We introduced two special tokens, <startseq> and <Endseq> to mark the caption's beginning and end while effectively preparing for training the model. This would make the model realize that what has been generated is up to what limit. We further tokenized the captions using Keras' Tokenizer. This created a mapping of words to unique integer indices, realizing vocabulary size that spelled out the amount of linguistic complexity of the dataset.

For the images, we applied specific pre-processing steps to ensure uniformity and enhance the model's performance. All the images were resized to 224x224 [19] pixels so that all input images were of the same size; this is usually a requirement if CNNs are going to be utilized. Normalizing the pixel range to the range 0 to 1 improved convergence in training. This time, we chose to settle for an existing pre-trained model DenseNet201 since it extracts meaningful features from images. It was trained on learning feature vectors, which would project to the same images whilst compressing and reducing dimensionality. As a step towards correct model evaluation, the data set is divided into a training and a validation set. We have mainly assigned 85% of the image for training purposes and the remaining 15% for validation where we test our model on unseen data.

B. System Architecture

1) Image Captioning Module: The image captioning mod- ule developed in this research utilizes an elaborate fusion of deep learning techniques to build image descriptions that are spot on and contextual. The module is designed based on the dual architecture of a Convolutional Neural Network (CNN), the design being dedicated to extracting visual features, and a Recurrent Neural Network (RNN), equipment of the LSTM units with a language-generating capability [20]. The fundamental aim of this module is to present visual data as meaningful textual descriptions. This improves the retrieval of the contents in the images.

The feature extraction component employs DenseNet201 architecture with densely connected layers for convolution. This is the reason why good feature reuse is ensured and it also eludes the vanishing gradient problem mainly because of deep networks. Preprocessed the input images to 224x224 pixels and normalized their pixel values to [0, 1] using a division operation by 255 [21]. DenseNet201 model pre-trained on ImageNet was used as a feature extractor. Elimination of the top classification layers outputs a feature vector of 1920 dimensions-the output of the penultimate layer. These feature vectors will encapsulate relevant visual information such as objects, textures, and spatial hierarchies that are visible in images.

Captions are then generated from the elicited features. This is accomplished using an LSTM-based decoder. A decoder, based on LSTM, works extremely well on sequential data and generate context for any number of variable time steps [22]. The captions are further tokenized, and using a trained tokenizer, the tokens were converted into sequences of integers. An embedding layer maps such integer sequences into dense vector representations where the model can capture semantic relationships between words. The LSTM is thus designed to iteratively predict the next word in the caption, conditioned upon on the extracted image features and the previously generated words.

The architecture showed in Fig 1. consists of two primary inputs: the image feature vector and the word sequence. The reshaped feature vector feeds into a fully connected dense layer with 256 units and ReLU activation. This will give an intermediate representation that can capture both visual and textual data. Then, the concatenation of this representation with the embedding of the current word forms the input to

the LSTM cell, thereby allowing it to make use of both visual and linguistic contexts in producing the next word. This continues until a defined maximum length is attained or until an end-of-sequence token is predicted.

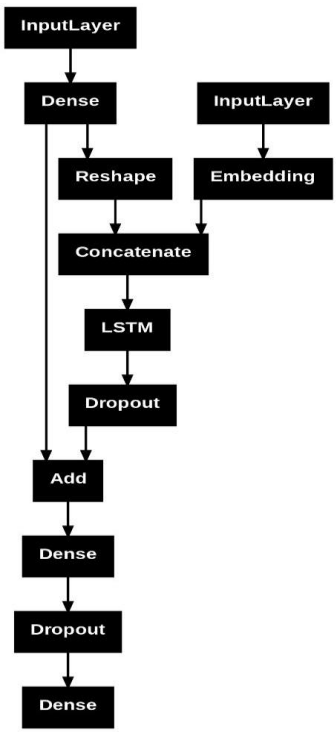

Fig. 1. Layered Architecture

2) Audio Module: The audio module utilizes the Google Text-to-Speech(gTTS) library to read the captions generated from the image captioning system in the audio format for listening [23]. This facility is complementary to the said captions viewable to the user, making the aspect a more natural and pleasantly experiential advancement for the user. gTTS offers a very simple API that reads input text processed using advanced neural network-based TTS models by Google, trained on various linguistic datasets. These models pick up every nuance, including intonation and rhythm, and produce speech that is almost indistinguishable from human articulation. Then there is the process of passing this formatted text to gTTS along with parameters for the language and the audio output, which results in an audio file that can be played within the application or saved for later offline use.

The audio module supports real-time caption-to-speech conversion, making the application more accessible to users who have impaired vision and also other users who process information better by hearing [24]. The image descriptions by the gTTS library open a way for the user to understand visual content without ever having to see it. This helps in enhancing inclusivity and ensuring that all users can access the application efficaciously. This audio feedback makes the inter- action much more appealing and adds greater complexity to the overall user experience while harmoniously incorporating both visual and audio information. The inclusion of the gTTS audio module fits well into the image captioning framework in that it provides an alternative way of content consumption and further justifies the multimodal approach taken by the proposed system. This integration encourages the simplicity of access and user interaction by the application to enable the blind to attain the maximal informative value offered within the images.

C. Training and Optimization

The model is trained using categorical cross-entropy loss, which quantifies the difference between the actual word and the probability that was likely predicted for the next word within the caption. The optimization is performed using Adam optimizer by adjusting the learning rate dynamically based on the first and second moments of the gradients to improve the effectiveness of convergence. Dropout at 0.5 is further applied after the LSTM layer to help it generalize and not overfit to the output layer, which is a dense layer. This custom data generator prepares batches of features with their captions during training preparation hence enhancing the training data. For each epoch of continuous updates of the model through weights using backpropagation through time (BPTT), this backpropagation through time propagates errors backward in the LSTM network with consideration of dependencies arising in sequences of time.

Several callbacks are used to optimize the performance and achieve a good train. These include *ModelCheckpoint*- saving the best model according to validation loss, *EarlyStopping* - where the training will be topped once validation loss does not improve, and *ReduceLROnPlateau*- in which the model's performance has reached a plateau such that the learning rate needs to be dynamically adjusted. All these mechanisms improve efficiency in training without overfitting; therefore, more stable and accurate image captioning outcomes are triggered.

D. Evaluation

In evaluating the quality of the captions generated, this module works on BLEU scores by comparing generated captions with reference captions. The range for the scores yielded falls between 0 and 1 [25], and the better the score, the closer it will be to a human-created caption. The image captioning module can generate many coherent, meaningful descriptions summarizing visual content; what it does is expose the capabilities in combined CNN and LSTMs.

IV. Results

While running some of the tests, we have come across some of the loss values at some preliminary runs. Table 1 shows the Training Loss and Validation Loss over 14 epoch which stopped early.

Epoch	Training Loss	Validation Loss
0	4.25	3.50
1	3.75	3.25
2	3.50	3.10
3	3.25	3.05
4	3.15	3.00
5	3.10	3.00
6	3.00	3.00
7	2.90	3.00
8	2.85	3.00

9	2.80	3.05
10	2.75	3.05
11	2.70	3.10
12	2.65	3.10
13	2.60	3.10
14	2.60	3.05

Table 1: Loss over Epoch

Fig. 2. Captured Test Image for Caption Generation

high-level features of the image. The output of the DenseNet201 is a feature vector summarizing the essential visual information from the image. After extracting the image feature, an LSTM network takes over the generation of the caption for the image. Caption generation is a sequential process. The model starts from the <Start> token and makes one word prediction after another to generate the caption. This is all repeated until this results in a <EndSeq> token at the end for the completion of the caption generation.

The model effectively processed the image, shown as a reference in Figure 2, and identified key features such as "Man,", "Sitting," and "Laptop," which were tokenized and sequentially passed to the LSTM network for word prediction via Softmax function. This combination of visual feature extraction and sequential language modeling allowed the system to generate a meaningful and contextually accurate caption that described the image content.

Generated Caption: *"a man sitting at a desk in front of a laptop computer."*

Our model was evaluated and got the BLEU score 0.7210778932010424. For comparison, we also trained our dataset with same parameters using InceptionV3 and ResNet50 layered architectures. Former gave the BLUE scope of 0.475887330964125, while ResNet50 model stopped training in between proving our model's efficiency better.

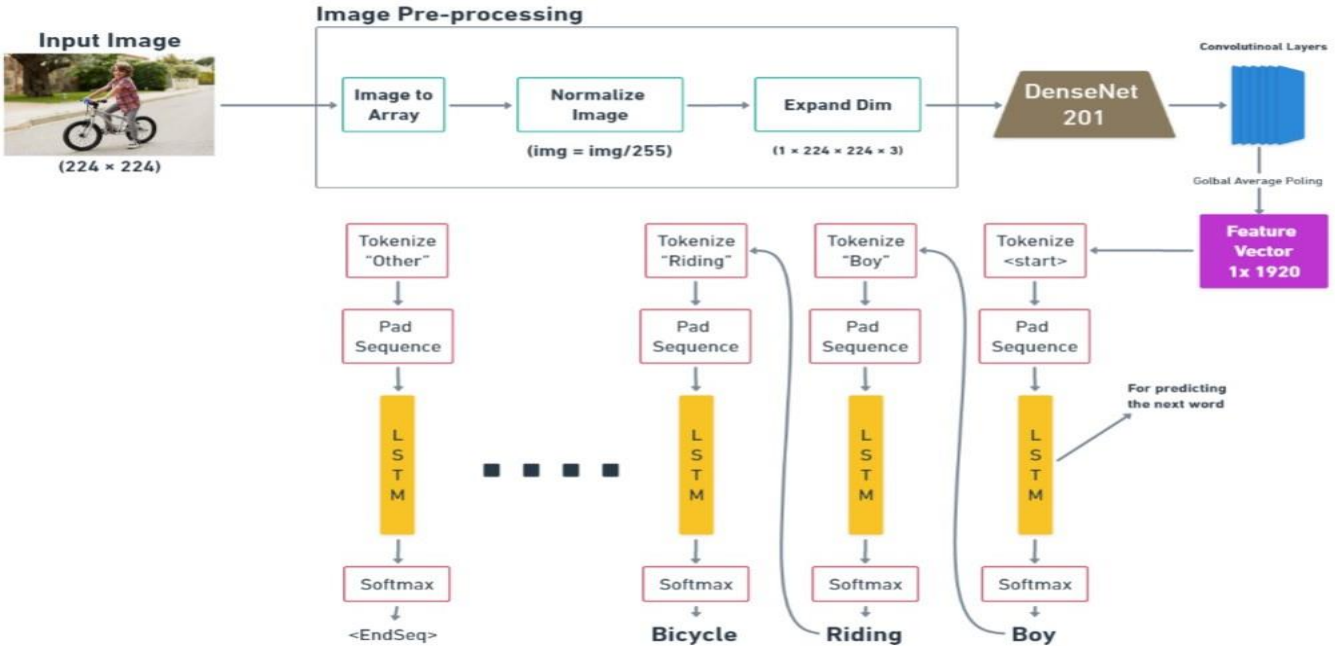

Fig. 3. System Architecture

The Fig 3. describes the image captioning model, an integration of a CNN merged with an LSTM network. Its early stage includes reading an input image-most preferably of size 224 x 224 pixels-then preprocesses the input image into a numerical array by normalizing the pixel values and expanding its dimension so that it can have batch processing. The preprocessed image is then fed into DenseNet201, one of the CNN architectures optimized for efficient extraction of

V. Future Scope

That proposed system in reference to the visual understanding and navigation through the visually impaired brings great potential for an upgrade. Some of the most critical upgradations that are needed for future inclusion are advanced NLP models that produce more elaborate captions, and those captions are contextually aware. In this manner, the system will not only describe the objects but, through

inference, be able to perceive actions, emotions, or interactions within the scene. The real-time GPS added to it, coupled with the mapping technology, makes further enhancement towards understanding directions outside while indoors by using visual and auditory cues that promote safe travel in unknown territories.

Some of the promising directions involve wearables such as smart glasses that will be constantly giving real-time feedback. Such a system can be included in lightest portable devices so that users can easily get information concerning their surroundings without necessarily capturing the images on screen. Haptic feedback systems may also include the provision where users receive tactile signals concerning the different objects or hazards.

In addition to these, sophisticated algorithms in the quest for object detection and classification might make the system more comfortable to be accurate in complicated scenarios like a busy or changing scene. The system may be supported to communicate in several languages, thus reaching diverse people from different countries. The presence of user feedback together with machine learning mechanisms empowers the system to be adaptive towards the preferences of different individuals, therefore becoming more intuitive and effective with time.

VI. Conclusion

This is the first approach toward assistance in assistive technology for the visually impaired, a development in the visual understanding and navigation system. The developed system relies on deep learning models like DenseNet201 for feature extraction and LSTM for image captioning. It provides an solid framework providing efficient solution for interpretation of visual information. Integrating with text-to-speech conversion via gTTS, the utility allows visually impaired users to achieve real-time auditory descriptions thus highly boosting their independence and mobility. It therefore not only smoothes navigation but connects to the rich experience of the world for sighted people to that of a blind person.

This system solved all the problems that can be integrated into something promising. Future refinement of such a system might end up getting more invested in the enhancement of contextual understanding from images, wearable technology with features of general-purpose computing in future wearables, and multi-sensory feedback systems. This project will, therefore become an ideal base for many intelligent and adaptive systems to be built as research in AI moves forward in the future, potentially leading to a higher quality of life for many with visual impairments

References

[1] Saleh, Shadi, Saleh, Hadi, Nazari, Mohammad and Hardt, Wolfram. (2019). Outdoor Navigation for Visually Impaired based on Deep Learning

[2] Islam RB, Akhter S, Iqbal F, Saif Ur Rahman M, Khan R. Deep learning based object detection and surrounding environment description for visually impaired people. Heliyon. 2023 Jun

[3] Latif, G., Brahim, G. B., Abdelhamid, S. E., Alghazo, R., Alhabib, G., & Alnujaidi, K. (2023). Learning at Your Fingertips: An Innovative IoT-Based AI-Powered Braille Learning System. *Applied System Innovation*, *6*(5), 91.

[4] Reddy, K. K., Badam, R., Alam, S., & Shuaib, M. (2024). IoT-driven accessibility: A refreshable OCR-Braille solution for visually impaired and deaf-blind users through WSN. *Journal of Economy and Technology*, *2*, 128-137.

[5] Ganesan, J.; Azar, A.T.; Alsenan, S.; Kamal, N.A.; Qureshi, B.; Hassanien, A.E. Deep Learning Reader for Visually Impaired. Electronics 2022

[6] Mukhiddinov, Mukhriddin and Jinsoo Cho. “Smart Glass System Using Deep Learning for the Blind and Visually Impaired.” Electronics (2021)

[7] Wang, Jiaji, Wang, Shuihua and Zhang, Yudong. (2023). Artificial Intel- ligence for Visually Impaired. Displays. 10.1016/j.displa.2023.102391

[8] Triantafyllidis AK, Tsanas A, Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature J Med Internet Res 2019;21(4):e12286

[9] Park, C., Took, C.C. and Seong, JK. Machine learning in biomedical engineering. Biomed. Eng. Lett. 8, 1–3 (2018)

[10] Z. Li, F. Song, B. C. Clark, D. R. Grooms and C. Liu, ”A Wearable Device for Indoor Imminent Danger Detection and Avoidance With Region-Based Ground Segmentation,” in IEEE Access, vol. 8, pp. 184808-184821, 2020

[11] Al-Malla, M.A., Jafar, A. and Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. J Big Data 9, 20 (2022)

[12] Ayush Kumar Poddar, Dr. Rajneesh Rani, Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language, Procedia Computer Science, Volume 218, 2023

[13] Bhalekar, M. and Bedekar, M. 2022. D-CNN: A New model for Generating Image Captions with Text Extraction Using Deep Learning for Visually Challenged Individuals. Engineering, Technology and Applied Science Research. 12, 2 (Apr. 2022)

[14] Kiros, Ryan, Salakhutdinov, Ruslan and Zemel, Richard. (2014). Uni- fying Visual-Semantic Embeddings with Multimodal Neural Language Models. 31st International Conference on Machine Learning, ICML 2014. 3.

[15] Y. Lu, C. Lu and C. -K. Tang, ”Online Video Object Detection Using Association LSTM,” 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017

[16] Mishra, R., Wadekar, S., Warbhe, S., Dalal, S., Mirajkar, R., Sathe, S. (2023). Facial Recognition System Using Transfer Learning with the Help of VGG16. In: Neustein, A., Mahalle, P.N., Joshi, P., Shinde, G.R. (eds) AI, IoT, Big Data and Cloud Computing for Industry 4.0. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-29713-7_9

[17] Verma, A., Yadav, A.K., Kumar, M. et al. Automatic image caption generation using deep learning. Multimed Tools Appl 83, 5309–5325 (2024)

[18] Al-Shamayleh, A.S., Adwan, O., Alsharaiah, M.A. et al. A comprehensive literature review on image captioning methods and metrics based on deep learning technique. Multimed Tools Appl 83, 34219–34268 (2024)

[19] S. Abinaya, M. Deepak and A. Sherly Alphonse,” Enhanced Image Captioning Using Bahdanau Attention Mechanism and Heuristic Beam Search Algorithm,” in IEEE Access, vol. 12, pp. 100991-101003, 2024

[20] H. S. Khan, R. Muzaffar, S. Y. Arafat and Z. Irshad,” Deep Learning- Based Urdu Image Captioning,” 2024 International Conference on Engineering & Computing Technologies (ICECT), Islamabad, Pakistan, 2024

[21] Khubchandani, V. (2024). Image caption generator using DenseNet201 and ResNet50. SSRN

[22] Liu, T., Cai, Q., Xu, C., Zhou, Z., Xiong, J., Qiao, Y., & Yang,
T. (2024). Image Captioning in news report scenario. arXiv preprint arXiv:2403.16209

[23] Kurlekar, S., Deshpande, O., Kamble, A., Omanna, A., & Patil, D. (2020). Reading Device for Blind People using Python OCR and GTTS. International Journal of Science and Engineering Applications, 9(4), 049-052

[24] Katakam, N. P., Mynedi, H., Imambi, S., & Nikhileswar, C. Deep learning techniques for Real time image to voice for the Visually impaired: A review.

[25] Faurina, R., Jelita, A., Vatresia, A., & Agustian, I. Image captioning to aid blind and visually impaired outdoor navigation. Int J Artif Intell ISSN, 2252(8938), 1105.