# CONVERSATIONAL ASSISTIVE TECHNOLOGY OF VISUALLY IMPAIRED PERSON FOR SOCIAL INTERACTION

K. Ghafoor*, T. Ahmad, M. Hanif and H. Zaheer

Department of Computer Science, University of Engineering and Technology, Lahore, 39161, Punjab, Pakistan

**ABSTRACT:** Assistive technology has been developed to support visually impaired people in their social interactions. Conversation assistive technology is designed to enhance communication skills, facilitate social interaction, and improve the quality of life of visually impaired individuals. This technology includes speech recognition, text-to-speech features, and other communication devices that enable users to communicate with others in real-time. The technology uses natural language processing and machine learning algorithms to analyze spoken language and provide appropriate responses. It also includes features such as voice commands, and audio feedback to provide users with a more immersive experience. These technologies have been shown to increase the confidence and independence of visually impaired individuals in social situations, and have the potential to improve their social skills and relationships with others. Overall, conversation-assistive technology is a promising tool for empowering visually impaired people and improving their social interactions. One of the key benefits of conversation-assistive technology is that it allows visually impaired individuals to overcome communication barriers that they may face in social situations. It can help them to communicate more effectively with friends, family, and colleagues, as well as strangers in public spaces. By providing a more seamless and natural way to communicate, this technology can help to reduce feelings of isolation and improve overall quality of life. The main objective of this research is to give blind users the capability to move around in unfamiliar environments, through a user-friendly device by face, object, and activity recognition system. This model evaluates the accuracy of activity recognition. This device captures the front view of the blind, detects the objects, recognizes the activities, and answers the blind query. It is implemented using the front view of the camera. The local dataset is collected that includes different 1st-person human activities. The results obtained are the identification of the activities that the VGG-16 model was trained on, where Hugging, Shaking Hands, Talking, Walking, Waving video, etc.

**Key words:** Dataset (DS), Visually Impaired person (VIP'S), Natural language process (NLP), human activity recognition (HAR), University of Engineering and Technology (UET), etc.

*(Received 29.09.2023 Accepted 21.11.2023)*

## INTRODUCTION

Information Technology (IT) has served a lot not only for normal people but also have worked for disable community to improve their quality of life. Prominent innovations in the field of IT have overcome the learning disabilities of the handicapped community. People influenced by visual disability needs proper assistance to perform their daily routine task. They are unable to navigate and visualize the unfamiliar environment. It is estimated that around the age of 60, more people face visual disability challenges. According to the World Health Organization (WHO), an estimated 285 million people around the world are visually impaired, with 39 million of them being completely blind ( Keel and Cieza 2021) and the remaining 217 million have moderate to severe visual impairment. These numbers are from the latest available report in 2021.

According to the Pakistan Blindness Survey, which was con- ducted in 2019 by the Pakistan National Eye Health Program in collaboration with the WHO, it is estimated that there are ap- proximately 1.7 million visually impaired people in Pakistan of these, around 225,000 people are completely blind. The

Prevalence of blindness in Pakistan is estimated to be around 1.1%, which is higher than the global average of 0.5%. Typical strategies to overcome the mobility limitation include the use a white cane and service dogs getting training using mobility specialists. Blind people are unaware of surrounding people intentions and their activities. In reality, it is estimated that around 2% to 8% of blind individuals use their cane to navigate. Others rely on their guide dog, their partial sight, or their sighted guide. Blind people have numerous challenges since they are unable to do their tasks independently without any assistance. Blond- ness leads to a lack of self-confidence and constant dependency on some kind of supportive or assistive tool. Due to loss of vision, some blind people are very sensitive to their hearing and touching modalities.

Research has identified several challenges faced by visually impaired persons in social activities, such as finding transportation, accessing venues, communicating with others, and participating in group activities. For instance, a study by (Kurniawan and Zaphiris, 2019) found that the lack of accessibility information, transportation options, and social stigma are the main barriers to social participation among the visually impaired population.

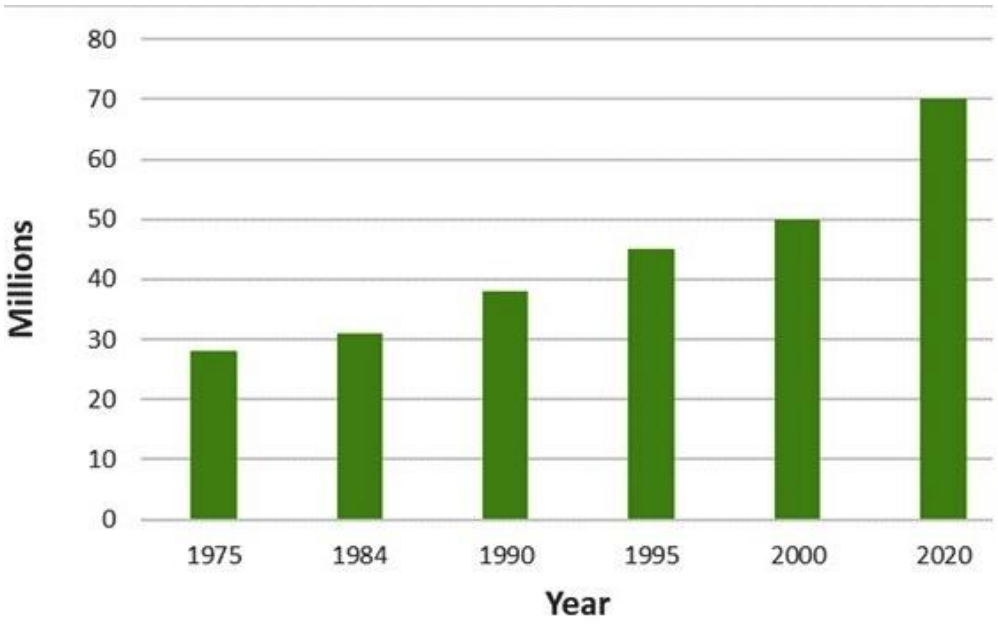


**Fig. 1. Blindness in the World Survey of Ophthalmology**

Furthermore, a study by (Hirano *et al*, 2018) indicated that visual impairments affect communication and social inter- action, leading to social isolation and reduced quality of life. A study by (Mohammed *et al*,2019) developed a conversational agent to assist visually impaired persons in navigation by pro- viding real-time information on obstacles and hazards. (Li *et al* 2021) developed a conversational agent to assist visually impaired persons in accessing social events. The agent en- abled users to search for events, obtain information on venues, transportation options, and ticket prices. The study found that the agent was effective in improving the accessibility and participation of visually impaired persons in social events. A study by (Chandrasekharan *et al*.,2020) developed a conversational agent to assist visually impaired persons in group activities, such as team-building exercises and group decision-making. The agent was designed to facilitate communication, task assignment, and team coordination. The study found that the agent was effective in assisting visually impaired persons in group activities. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. Sample body text. (Starner, R., and Shew, A.2020). Increasing ability, breaking down barriers, and growing communities: assistive technology and the flourishing of people with vision loss through social interaction. Starner and Shew explore the role of assistive technology in promoting social interaction among visually impaired individuals. They argue that communication assistive technology can foster social participation, eliminate barriers, and im- prove the quality of life for visually impaired people. (Law, M., and Petrina, S 2015) conducted a survey to investi- gate the use of assistive technology among visually impaired students and teachers in Hong Kong. The study found that the use of communication assistive technology has improved social interactions among visually impaired students and in- creased their participation in classroom activities (Akkermans and Layard).

(Isaac *et al*. (2019)) give a survey on Communication technologies for individuals with vision and hearing impairments. Isaac provides an overview of various communication assistive technologies available for individuals with vision and hearing Impairments. The chapter highlights the importance of these tech nologies in facilitating social inclusion, communication, and participation in daily activities. (Sharma *et al*. (2014)) proposed a novel assistive reading system tem for visually impaired people based on adaptive thresholding and text-to-speech technology. The system converts printed text into audio, allowing visually impaired individuals to read books, newspapers, and other printed materials. (Lenssen, J. E. M., and Schro¨der, J., 2015).

Communication technology for vi- sually impaired and blind people. Lenssen and Schro¨der review various assistive technologies available for visually impaired and blind people for communication and social interaction. The authors discuss the importance of customizing communication technologies based on individual needs to improve social inter- action, participation and inclusion.

(Padhy *et al*. (2023)), provides a comprehensive survey of various assistive technologies that have been developed for visually impaired individuals from 2018 to 2021, including communication assistive technologies. The authors discuss the design, implementation, and effectiveness of these technologies in enhancing the social interaction of visually impaired individuals. (Blok *et al*. (2020)), Utilizing Technology to Support Social Interaction among Visually Impaired People. Adan and Alhumaidi analyze the impact of technology on enhancing social interaction among visually impaired individuals, specifically focusing on social networks and communication assistive technologies. The study highlights the need to tailor communication assistive technologies to the unique needs and preferences of visually impaired individuals to maximize their effectiveness in promoting social interaction.

(Liu *et al*. (2022)) Researcher provide a comprehensive survey of design and development of a Wearable Device for Social Interaction of Visually Impaired People. This research paper dis- cusses the design and development of a wearable device that uses haptic feedback to improve the social interaction of visually impaired individuals. The device is designed to provide real-time information about the physical presence of people and objects, thereby enhancing social interaction in social settings. (Talanov *et al*. (2022)), propose a system that combines ultra- sonic sensors and speech feedback to provide visually impaired individuals with real-time information about the location of objects in their environment. The system is designed to enhance the social interaction of visually impaired individuals by pro- viding them with information about the physical presence of other people. (Franco *et al*. (2023)), describe the design and implementation of an Android application that utilizes a com- bination of audio, haptic, and touchscreen feedback to enable social interaction among visually impaired individuals. The ap- plication is designed to enhance social interaction by enabling visually impaired individuals to access and share information in a social context.

(Solanki *et al*. (2023)), provide a system that captures the indoor environment, process it and classify the objects exist in images However, the detection becomes very challenging when it is used for recognition of specific objects or unknown obstacles in unfamiliar natural environments. Tele-Guidance systems can be associated with haptic-based cues and spoken instruction from remote caretakers using live video streaming carried by the blind or VIP. Blind persons use Lidars and Vibrotactile Units (LVUs) to sense obstructions. The audio jockey receives the detected image as an audio input. When the impaired individual reaches the obstacle, the audio jockey's vibratory motor and vocal intimation give haptic input.

(Whelan *et al*. (2012)), introduced a wearable system integrated with wearable terminal with an RGBD camera and an ear- phone, a powerful processor particularly for deep learning inference, and a smart- phone for touch-based interaction. (Lin, 2017)Lin *et al*. (2017a), introduced an application that was de veloped to facilitate the visually impaired people while mobility in hospitals, clinics and urgent cases that detect the doors, stairs and sign board with remarkable guidance for indoor navigation. (Qiu *et al*. (2020)), give a comprehensive survey on Simulated gaze and the tactile feedback were significantly effective to enhance the communication quality in the blind and sighted conversation. They also investigate sighted and blind people perceptions and reactions to the interactive Gaze affects the communication quality in blind sighted conversations. Author imposed spoken language intent detection under noisy condition using automatic speech recognition. They deploy confusion 2vec word feature representation to increase the robustness of the spoken language understanding (SLU) and remove the error which is made by (ASR). ATIS benchmark dataset is used to reduce classification error rate 20.84% from (ASR) and increase the robustness rate 37.48% from (SLU).

(Liu *et al*. (2019)), introduced new system which provide dedicated interface by integrated human action recognition (HAR) and sign language recognition (SLR). By integrating both inter- face name is (HASLR). (Narayan *et al*. (2014)), work on trajectories and evaluate the performance of trajectories. (Ozarkar *et al*. (2020)), work on basic three things human -human interaction, human object interaction and human object human interaction. The proposed deep learning model was used to ex- tract deep learning feature from spatial and temporal domain. This model provides accuracy score of 90% in UCI and 87% in WISDM database. Also proposed face recognition system that is comprised of three main modules including Dataset creation, dataset training, and face recognition. Here Haar cascade classifier is used to detect face from a live video stream and then local binary pattern histogram (LBPH) algorithm. This System can detect and recognize multiple people and is also capable of recognizing from both front and side face.

(Shan *et al*. (2020)), distinguishing highlights in each casing prompts computational failure. In a long video transfer, a few casings might have poor quality because of movement obscure, video refocus, impediment, and posture changes. (Husain *et al*. (2016)), used a huge scope dataset named YouTube-Bounding

Boxes (YT-BB) in his research. Which is human-commented on at one edge for each s on video pieces from YouTube with high exactness order names and tight bouncing boxes. YTBB contains around 380,000 video portions with 5.6 million jumping boxes of 23 item classifications. (Sarwar *et al*. (2021)), used KTH dataset which is one of the most standard datasets, which contains six activities: walk, run, run, box, hand-wave, and hand applaud.

To represent execution subtlety, each activity is performed by 25 unique people, and the setting is deliberately changed for each activity per entertainer. Among different exercises, a video dataset was made in this paper. (Zhu *et al*. (2017)), use CAVIAR dataset which incorporates individuals performing 9 exercises. Based on this literature survey from the past five years on conversation assistive technology of visually impaired persons for social interaction, the research gap seems to focus on the customization of assistive technologies for different individuals with visual impairments. While many studies have explored the broad application of assistive technology in promoting social interaction and reducing barriers, there is a need for further research on how assistive technologies can be tailored to meet the specific needs of individuals with varying degrees of visual impairment. And also need to enhance the social interaction specifically for blind people by using real time scenarios.

**Data set:** Designing NRPU system that can be successfully deployed in visually impaired persons. It requires datasets that pose the challenges typical of real-world scenarios. In this paper, we introduce a new UET interaction dataset for visually impaired people. UET interaction dataset cover social interaction modules between blind and sighted person.

**UET interaction dataset:** UET interaction dataset specially design for blind people, and try to cover all possible objects, and activities regarding blinds need in one dataset in a spontaneous manner. The dataset contains Bounding Boxes, Polygonal Segmentation, points, an- notations including activities and involving interactions with objects. We provide a real base environment and its challenges featured by our dataset, highlighting the open issues for object detection and recognition algorithms. Therefore, we propose a new HAR method for human activity recognition to tackle the novel challenges provided by our dataset. HAR methods can be divided into two categories based on the types of sensors: wearable devices, smart phone cameras.

We collect a data from video sequence which is still a big challenging task. This method provides us an effective state of communication regarding human-human interaction, and human object interaction. This is particularly beneficial to detect social activities, poses of sighted person who stand in front of blind. We show that the method we propose UET interaction dataset is another popular challenging dataset. Building such a surveillance system requires a long-term, well under- stood understanding. In recent years, numerous datasets for activity recognition, object recognition, in trimmed videos have been proposed. Dataset for social interaction proposed in (Husain *et al*. (2016). One more researcher provide a dataset on social interaction Shadi *et al*. (2019). Object activity and text reading dataset provide in Kang and Wildes (2016). Very Little work has been done on activities and object recognition for blinds. By activity detection, we mean predicting activity labels as well as temporal boundaries within an input video.

**This UET interaction dataset has to cope with important open challenges:**

- Handling the combinations explosion of activity proposals while detecting navigation path, pose estimation, and social interaction activities in along one video sequences.
- Managing concurrent activities.
- Distinguishing between indoor and outdoor activities.

In this work, we focus on videos of Activities and objects recognition for blinds. For social interaction dataset collection, out- door as well as indoor environments were considered. The dataset consists of 50 recorded videos at a resolution of (640 x 352) using a front view camera which was held by the observer in a landscape position at the angle of the eyes. The camera was not stationary but moved along with the eye movement of the observer. The videos were not recorded in controlled environments and hence provide real-world conditions. The videos range in length from 7 to 13 seconds, and each one includes different types of activities and facial expressions.

Motivated by the shortcomings of current datasets, we introduce UET interaction dataset for blinds. UET interaction pro- vides realistic videos with diverse spontaneous human activities and real-world environment. We invited some students of com puter science department and librarian for the recording videos in a university. The volunteers are senior people in the age range of 40 to 50 years. Each volunteer was recorded for 12 to 15 seconds for one video. The resulting data consists of 536 long RGB videos with 50 annotated activities. Overview of the challenges in UET interaction. On the left part, we present challenges. Our goal is to create a large-scale dataset for visually impaired person where all module is present in only one dataset.

We record a video from various perspectives, we employ surveillance cameras and smart phones for recording videos, with the camera positioned front and center in a landscape orientation, a smart phone is used to record first-person activities from the perspective of a blind person. This extensive interpretation process has

resulted in a wide variety of activities. Figure 1 presents the diversity of activities. In this dataset. Activities are classified into single or multiple activities. Single activities are primarily conducted between individuals. And multiple activities are performed between multiple actors in the same videos. UET interaction has several activities which are relatively in Figure -2.

## MATERIALS AND METHODS

A system takes input from visually impaired person in the form of voice and video. The user query is converted into text using Speech to Text module. Intent Detection Module recognizes the intention of the visual impaired person regarding visual environment.

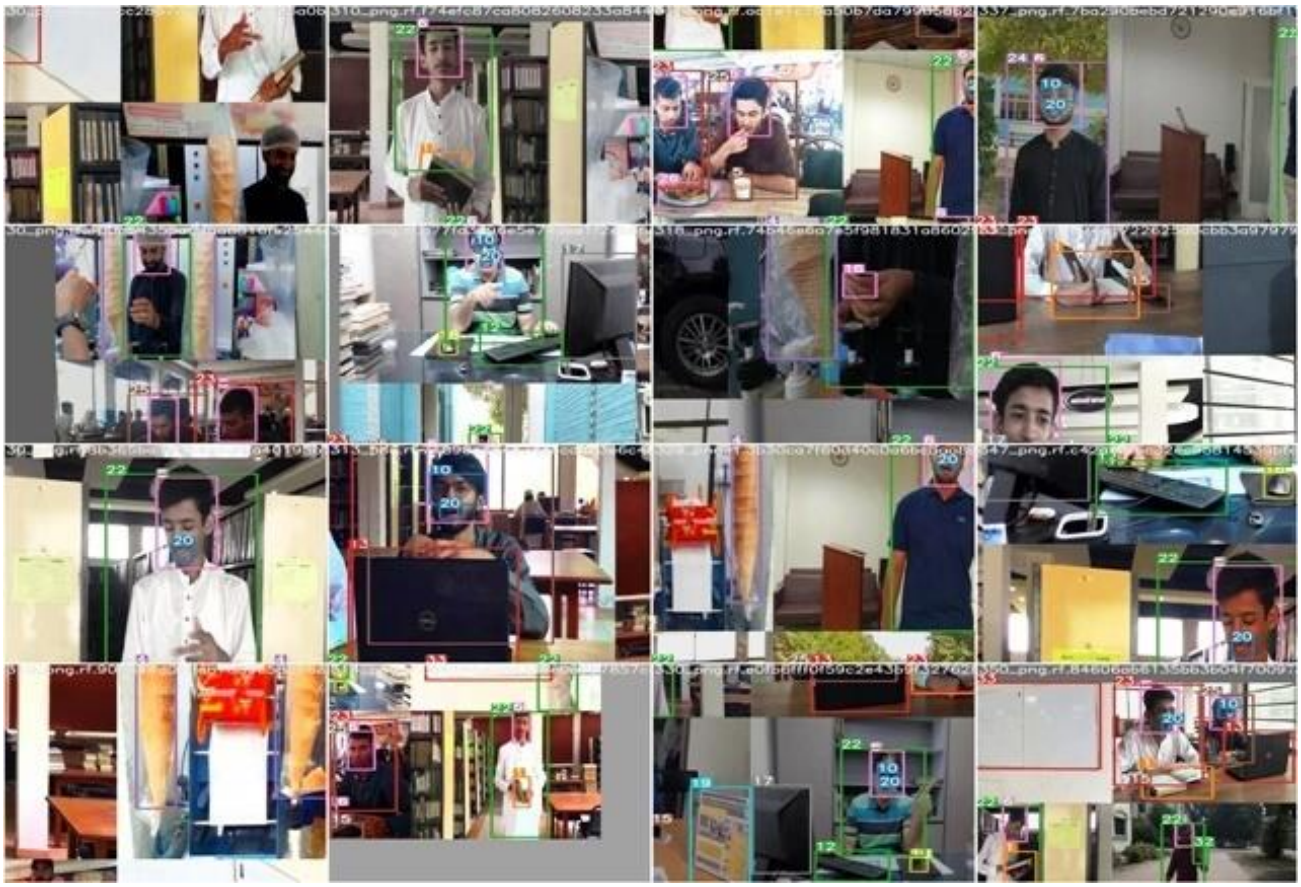

**Fig.2. Exhibits a diverse human activity, exemplifying handshake, walking, and standing.**

Video stream goes to three different modules; Face Detection, Object Detection and Activity Recognition. Face Detection and Object Detection works on the static image as shown in figure 4. In our research I used machine learning and deep learning algorithms, Yolov5, Yolo V7 and Yolo V8 used for object detection module.

In the object detection module, the system gets the prepossessed image frames from image processing module. These are pro- vided to the object detection module. The model returns the class category including tree, sidewalk, road, person etc. The corresponding bounding box positions. There can be multiple objects in single scene. All the detections are then sent to the activity recognition. It gives better accuracy and AP (average precision) score than the previous ones i.e., YOLOv5, VGG16, R-CNN etc. It has 6 different weights including YOLOv7, YOLOv7-X, and YOLOv7- V6 etc. which give different AP score for datasets. YOLOv7 is pretrained on COCO dataset and gives 51.4% AP score. The same model weight was trained on custom dataset which gave 56%mAP score (mean average precision). Activity Recognition module takes the sequence of images as input. Hidden Markov the model (HMM) is crucial to the activity recognition system.

It is employed for pattern recognition, gesture recognition, and speech recognition. This study classifies the data and time instances while demonstrating the activity recognition algorithms. In order to determine the features in each model layer can distinguish between activities and objects a support vector machine is used for classification of an images. It was trained on the output of each layer to classify different activities. A

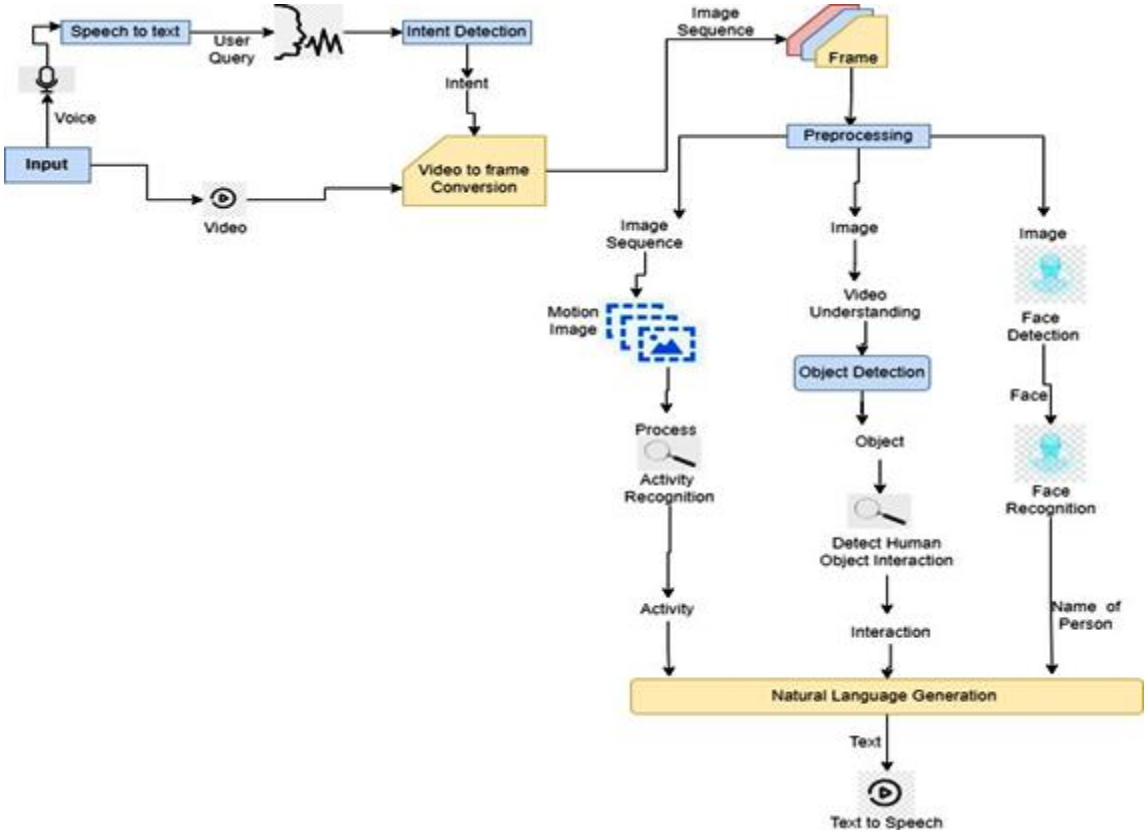


**Fig.3. System architecture for social interaction of visually impaired.**

Linear SVM is known as best method for this task. Data points should be linearly separated in order to be read out by down- stream population of neurons Negri and Lotito (2012) (Hui- juan Xu *et al.*, 2017), ICCV introduced a new model, Region Convolutional 3D Network (R-C3D), which encodes the video streams using a three-dimensional fully convolutional network, then generate candidate temporal regions. Consists of activities, and finally classifies selected areas. Accounts are saved due to certain activities.

A sharing of concentration characteristics between the proposition and Classification of pipelines. The entire model is trained. End-to-end with improved localization and classification losses combined. RC3D is faster than existing methods (569 frames per second). JPL first person interaction datasetryoo2013firs is a large-scale dataset utilized for human activities samples on a large number of categories and provides good within-class variability to capture different transformations. This dataset includes 150,000 images with 1,000 activities for validation and test data. UET interaction dataset consists of 50 videos that includes various activities. The extracted frames from the custom dataset were trained on VGG16 model. The dataset was spilt into two parts: 80 percent for training and 20 percent for testing the model. The training accuracy of the locally trained model is 0.88 percent while Val accuracy is 0.05. The loss of our trained model is 0.5 and Val loss is found to be 1.62.

Following Convolutional Neural Networks (CNN) models are used for the image and object classification. In our proposed research we discuss different architectures ResNet 50, Residual network (RN), and VGG 16. These models evaluate how neural network interact with social activities for visually impaired person. The UET interaction dataset of social activities is input to the CNN, VGG 16, and ResNet50. Principle Component Analysis (PCA) algorithm is used for input layers to evaluate the performance of models, and support Vector Machine (SVM) is used for classification and feature extraction from an image. SVM is trained on the output of the model for detect social activities regarding visually impaired persons.

Four different models used to compare on different task of social activities. These four models consist of two neural network architecture trained on different task. The first two models utilize VGG 16 network. One model was trained on object recognition and other one is used to trained on activity recognition. Second two models utilize a much deeper ResNet 50 network and trained on same object and activities recognition tasks. The set of residual models is ResNet 50 which is proposed by (he *et al.*, 2017) in Sarfraz *et al.* (2017).

ResNet 50 have great success in image recognition by forming deep graphs with implementation containing 150 hidden layers (hidden layers depends on dataset). The key insight of the ResNet architecture in the preservation of data representation. CNN get deeper the risk of extra layers mistakenly altering on already correct representation of the data also increase. Lower layers can form a very good representation of the data, high layers have very complex structure and it attempt to correct the residual error across convolutions, or copy of the lower layers as they were.

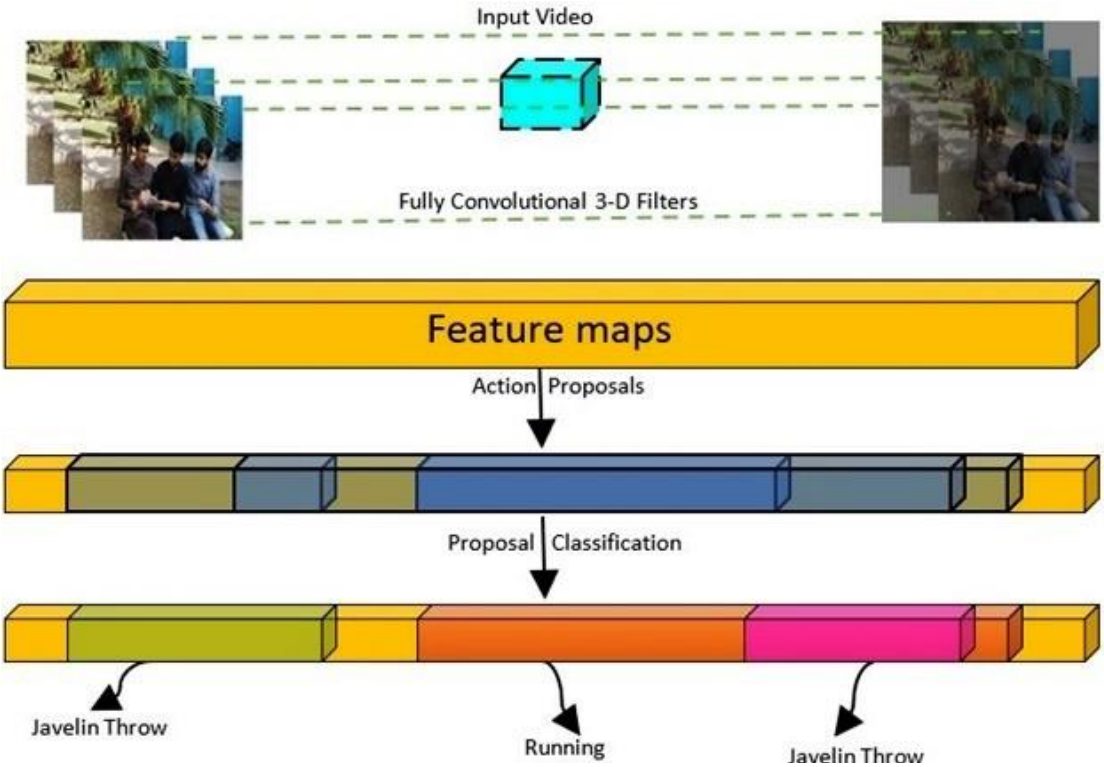


**Fig.4. Feature extraction from video**

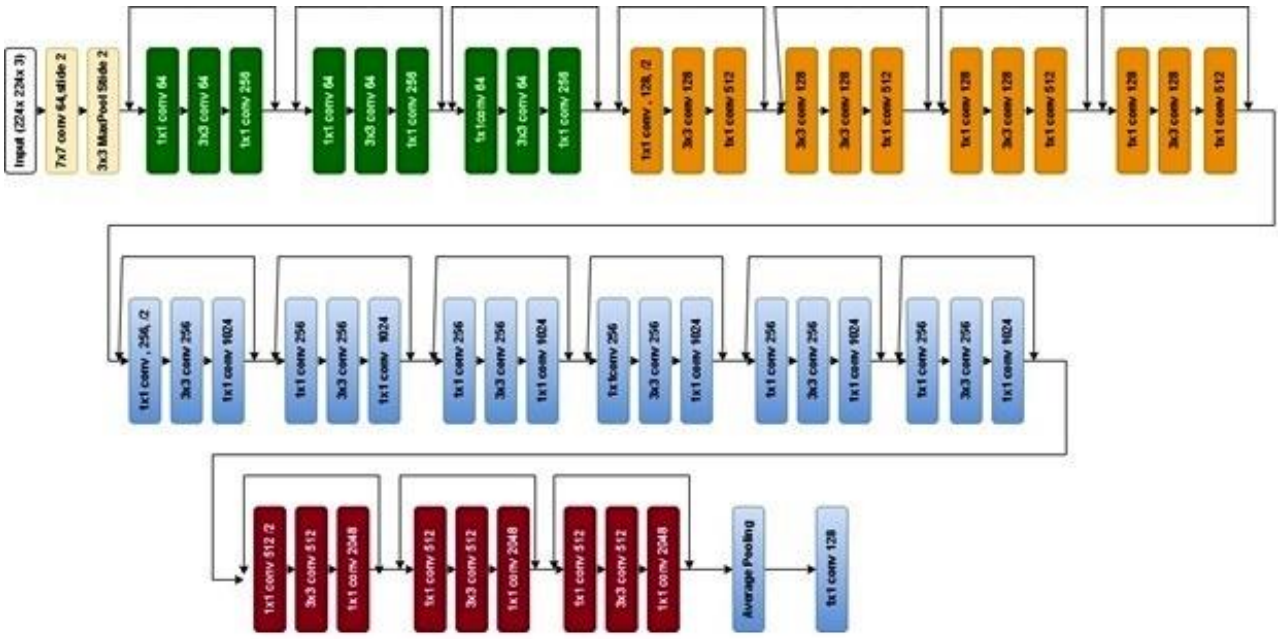

**Fig.5. RN consists of 34 blocks. Each block represents a series of convolutions with the same dimension.**

## RESULTS AND DISCUSSION

The frame interval is set to 5 and 16 on JPL and Activity Net 1.3 separately. On JPL, we set the length of observation window l‰ to 128 and the maximum duration length L to 64, which can cover length of 99% action instances. On Activity Net, following Real *et al*. (2017) Sarfraz *et al*. (2017). We rescale each feature sequence to the length of the observation window l‰ = 100 using linear interpolation, and the duration of corresponding annotations to range [0,1]. The time complexity of Hidden Markov Models (HMMs) algorithms is typically on the order of $O(T * N^2)$ , where T represents the length of the sequence being analyzed and N denotes the number of states within the HMM.

The maximum duration length L is set to 100, which can cover length of all action instances. To train BMN from scratch, we set learning rate of 0.0001, batch size of 18 and epoch number of 50 for both datasets. T. Lin, X. Zhao, and Z. Shou, present the validation set of Activity Net 1.3 by using AR@100 (Val) dataset and provide 72.01% accuracy, and T. Lin, X. Zhao, H. Su, C. Wang also present 75.0% accuracy and: J. Gao, K. Chen, and R. Nevatia present 72.17% on the other hand UET Interaction dataset provide us 76.23% accuracy on same dataset. From the labeled photos in yolov7 format, atrained and valid set was created. The model is trained using Google Colab. For model training, a desktop CD is mounted in a Colab file. The Github repository of Yolov7 is downloaded to the computer's hard disc. The necessary software and model weights are downloaded before beginning the model training process. YAML files are then used to specify the location of the dataset for training the model.

This is the testing phase of the model in which a given video detects the different activities of persons. Each frame of the video detects the activities like waving hands, moving hands, talking, walking legs, hugging hands, shaking hand and other objects like a person, face, and so on. Yolov7 model used for the activation detection. Local data-set used for the training purpose. Total 6 activities specify in this model training i.e., giving hand, waving hand, hugging hand, shaking hand, walk-walking and open mouth. The accuracy of the activity detection was evaluated using F1 curve, recall and precision. The results showed that the activity detection model was able to accurately recognized activity in the dataset with a precision of 0.87, and recall of 0.75. The results demonstrate that the activity detection model is a effective model for activity detection. The high precision and recall scores indicate that the model is able to accurately detect activity with low false positives and false negatives.

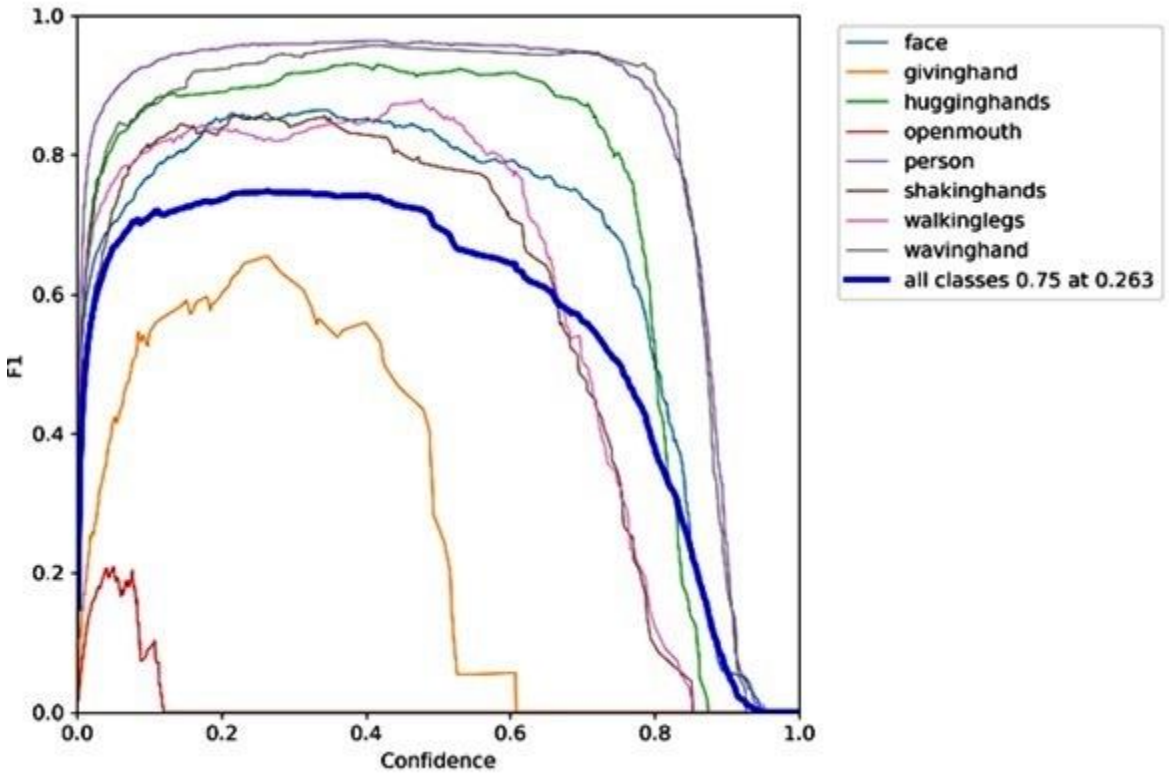


**Fig. 6. Striking the optimal balance: exploring the F1 score.**

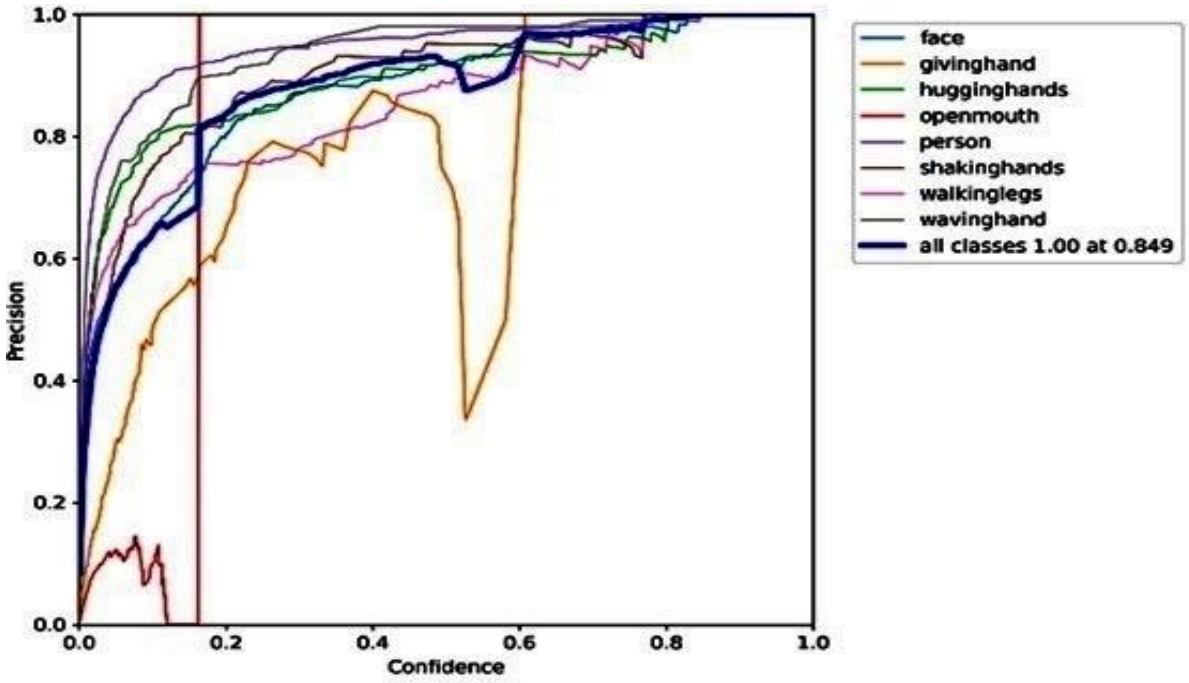


**Fig.7. Mapping the path of precision: visualizing accurate classification.**

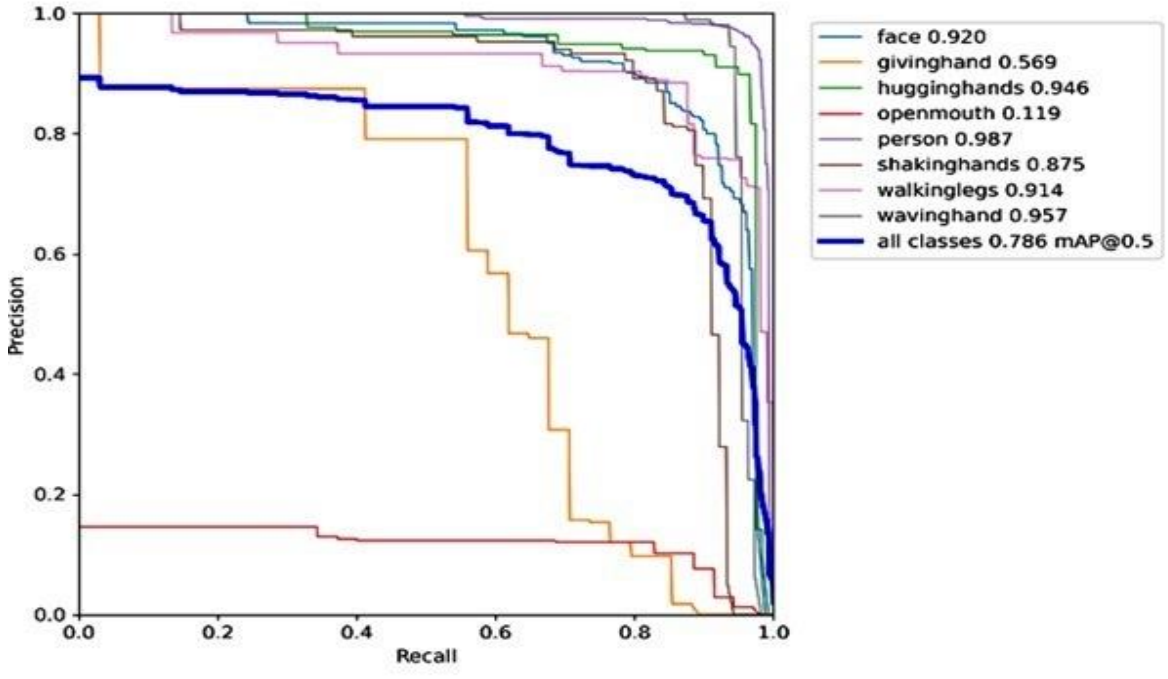


**Fig. 8. Mapping the performance: unveiling the mean average precision curve.**

**Table 1: Comparison between our method and other state-of-theart temporal action proposal generation methods on validation set of ActivityNet-1.3 dataset in terms of AR@AN and AUC.**

| **Method** | **Dai *et al.* (2017)** | **Ghanem *et al.* (2017)** | **Lin *et al.* (2017b)** | **Gao *et al.* (2018)** | **Lin *et al.* (2018)** | **UET interaction** |
|---|---|---|---|---|---|---|
| AR@100(val) | —- | —- | 72.01 | 72.17 | 75.00 | 76.23 |
| AUC(val) | 58.58 | 60.12 | 63.80 | 63.80 | 63.80 | 63.80 |
| AUC(test) | 62.56 | 66.18 | 61.80 | —- | 62.00 | 67.28 |

**Table 2: Comparison between our method with state-of-the-art proposal generation methods JPLRyoo and Matthies (2013), COCOLin *et al.* (2014), SSTBuch *et al.* (2017),SCNNShou *et al.* (2016), SCANDKarnan *et al.* (2022), TAGZhao *et al.* (2017), CTAPGao *et al.* (2018), BSNLin *et al.* (2018) on UET interaction dataset in terms of AR@AN, where SNMS stands for Soft-NMS.**

| | | | | | | |
|---|---|---|---|---|---|---|
| C3D | SCNN-Prop | 18.22 | 25.17 | 36.03 | 52.17 | 59.30 |
| C3D | SST | 19.00 | 20.11 | 36.90 | 52.51 | 60.26 |
| C3D | TURN | 18.63 | 27.68 | 38.43 | 53.57 | 61.27 |
| C3D | BSN+NMS | 28.19 | 35.48 | 38.16 | 57.23 | 60.89 |
| C3D | BSN+SNMS | 29.12 | 36.38 | 45.55 | 59.39 | 61.48 |
| C3D | UET interaction+NMS | 29.48 | 36.90 | 47.79 | 60.23 | 62.96 |
| C3D | UETinteracttion+SNMS | 33.73 | 42.68 | 47.89 | 61.73 | 62.57 |
| 2 Stream | TAG | 17.55 | 29.73 | 40.61 | — | — |
| Flow | TURN | 20.65 | 32.21 | 42.79 | 58.69 | 64.89 |
| 2 Stream | CTAP | 31.49 | 43.89 | 53.22 | — | — |
| 2 Stream | BSN+SNMS | 35.47 | 44.79 | 42.34 | 63.35 | 65.00 |
| 2 Stream | BSN+SNMS | 37.48 | 46.03 | 50.69 | 62.64 | 65.49 |
| 2 Stream | UET interaction+NMS | 38.36 | 47.75 | 55.84 | 63.19 | 67.23 |
| 2 Stream | UET interaction+SNMS | 40.15 | 49.72 | 56.70 | 63.07 | 68.49 |

**Conclusion and Future work:** In summary, this study aims to develop a conversational assistive agent for visually impaired people in Pakistan. The objective of the model is to enable blind users to interact with the social environment independently. The model focuses on face, object, and activity identification to provide blind individuals with the capability to navigate unfamiliar environments. It utilizes a front-facing camera to capture the blind person's view, detect objects, recognize activities, and respond to user queries. This research contributes to the development of assistive technology for the visually impaired in Pakistan. The proposed conversational support agent offers a promising solution to enhance the social interaction and mobility of the visually impaired, empowering them to navigate and engage with their environment more freely. The study emphasizes the importance of developing a user-friendly tool that integrates natural language processing (NLP) and machine learning techniques. By taking advantage of these technologies, visually impaired people can access information about their surroundings and improve their social interactions. The effectiveness of the model is evaluated through activity recognition, and the results show the correct identification of activities such as hugging, shaking hands, talking, walking, and waving. Further developments in this area could improve the quality of life for the visually impaired by providing better access and support. Communication assistive technology has been designed to eliminate barriers and support social participation, increasing the independence and quality of life of individuals with visual impairments. The continued exploration and implementation of communication assistive technology can lead to further innovations, ultimately enhancing the social integration of people with visual impairments.

A promising avenue for future development involves the implementation of a mobile version aimed at enhancing the assistance provided to visually impaired individuals. By crafting a mobile-friendly adaptation, we can offer a more seamless and efficient user experience. This mobile iteration would be designed with an emphasis on user-friendliness, ease of adaptation, and increased accessibility. Leveraging the ubiquity of smartphones. This evolution would ensure that the visually impaired community can access vital assistance effortlessly, promoting greater independence and inclusion. As we embark on this path, attention to intuitive design, enhanced features, and streamlined user interaction will be essential, enabling us to further empower visually impaired individuals in their daily lives.

## REFERENCES

Akkermans, B., Layard, A. The editors have brought together an impress sive and diverse group of authors from across the globe. the book deals with property in the context of law and society and therefore illustrates how prop erty comes to life in the real world whilst at the same time providing a rich source of state-of-the-art research for property scholars.

Blok, M., van Ingen, E., de Boer, A.H., Slootman, M., 2020. The use of information and communication technologies by older people with cognitive impairments: from barriers to benefits. Computers in Human Behavior 104, 106173.

Buch, S., Escorcia, V., Shen, C., Ghanem, B., Carlos Niebles, J., 2017. Sst: Single-stream temporal action proposals, in: Proceedings of the IEEE

conference on Computer Vision and Pattern Recognition, pp. 2911–2920.

Dai, X., Singh, B., Zhang, G., Davis, L.S., Qiu Chen, Y., 2017. Temporal context network for activity localization in videos, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5793–5802.

Franco, M., Gaggi, O., Merzougui, S.E., Palazzi, C.E., 2023. Accessible wayfinding for the visually impaired through sustainable smartphone-based sensing, in: 2023 IEEE 20th Consumer Communications & Networking Conference (CCNC), IEEE. pp. 1–6.

Gafni, H., Zeevi, Y., 1977. A model for separation of spatial and temporal information in the visual system. Biological Cybernetics 28, 73–82.

Gao, J., Chen, K., Nevatia, R., 2018. Ctap: Complementary temporal action proposal generation, in: Proceedings of the European conference on computer vision (ECCV), pp. 68–83.

Ghanem, B., Niebles, J.C., Snoek, C., Heilbron, F.C., Alwassel, H., Khrisna, R., Escorcia, V., Hata, K., Buch, S., 2017. Activitynet challenge 2017 summary. ArXiv preprint arXiv: 1710.08011.

Husain, F., Schulz, H., Dellen, B., Torras, C., Behnke, S., 2016. Combining semantic and geometric features for object class segmentation of indoor scenes. IEEE Robotics and Automation Letters 2, 49–55.

Isaac, E.R., Elias, S., Rajagopalan, S., Easwarakumar, K., 2019. Trait of gait: A survey on gait biometrics. ArXiv preprint arXiv: 1903.10744.

Kang, S.M., Wildes, R.P., 2016. Review of action recognition and detection methods. ArXiv preprint arXiv: 1610.06906.

Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P., 2022. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. IEEE Robotics and Automation Letters.

Lin, B.S., Lee, C.C., Chiang, P.Y., 2017a. Simple smartphone-based guiding system for visually impaired people. Sensors 17, 1371.

Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. CoRR abs/1405.0312. URL: http://arxiv. org/abs/1405.0312, arXiv: 1405.0312.

Lin, T., Zhao, X., Shou, Z., 2017b. Temporal convolution-based action pro- posal: Submission to Activitynet 2017. arXiv preprint arXiv: 1707.06750.

Lin, T., Zhao, X., Su, H., Wang, C., Yang, M., 2018. Bsn: Boundary sensitive network for temporal action proposal generation, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Liu, C., Lu, J., Yang, H., Guo, K., 2022. Current state of robotics in hand rehabilitation after stroke: a systematic review. Applied Sciences 12, 4540.

Liu, J., Li, Y., Lin, M., 2019. Review of intent detection methods in the human machine dialogue system, in: Journal of Physics: Conference Series , IOP Publishing , P, 012059.

Narayan, S., Kankanhalli, M.S., Ramakrishnan, K.R., 2014. Action and interaction recognition in first-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 512–518.

Negri, P., Lotito, P., 2012. Pedestrian detection on caviar dataset using a movement feature space, in: XIII Argentine Symposium on Technology (AST 2012) (XLII JAIIO, La Plata, 27 y 28 degosto de 2012).

Ozarkar, S., Chetwani, R., Devare, S., Haryani, S., Giri, N., 2020. Ai for accessibility: virtual assistant for hearing impaired, in: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE. pp. 1–7.

Keel, S., Cieza, A., 2021. Rising to the challenge: estimates of the magnitude and causes of vision impairment and blindness. The Lancet Global Health 9, e100–e101.

Padhy, S., Alowaidi, M., Dash, S., Alshehri, M., Malla, P.P., Routray, S., Al- humyani, H., 2023. Agrisecure: A fog computing-based security framework for agriculture 4.0 via blockchain. Processes 11, 757.

Qiu, S., Hu, J., Han, T., Osawa, H., Rauterberg, M., 2020. Social glasses: simulating interactive gaze for visually impaired people in face-to-face communication. International Journal of Human–Computer Interaction 36, 839–855

Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V., 2017. Youtube bounding boxes: A large high-precision human-annotated data set for object detection in video, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5296–5305.

Ryoo, M.S., Matthies, L., 2013. First-person activity recognition: What are they doing to me? in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2730–2737.

Sarfraz, M.S., Constantinescu, A., Zuzej, M., Stiefelhagen, R., 2017. A multi- modal assistive

system for helping visually impaired in social interactions. Informatik-Spektrum 40, 540–545.

Sarwar, M.G., Dey, A., Das, A., 2021. Developing a lbph-based face recognition system for visually impaired people, in: 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), IEEE. pp. 286–289.

Shadi, S., Hadi, S., Nazari, M.A., Hardt, W., 2019. Outdoor navigation for visually impaired based on deep learning, in: Proc. CEUR Workshop Proc, pp. 97–406.

Shan, C.Y., Han, P.Y., Yin, O.S., 2020. Deep analysis for smartphone-based human activity recognition, in: 2020 8th International Conference on Information and Communication Technology (ICoICT), IEEE. pp. 1–5.

Sharma, A., Srivastava, A., Vashishth, A., 2014. An assistive reading system for visually impaired using ocr and tts. International Journal of Computer Applications 95.

Shou, Z., Wang, D., Chang, S.F., 2016. Temporal action localization in untrimmed videos via multi-stage cnns, in: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 1049–1058.

Solanki, R., Shankar, A., Modi, U., Patel, S., 2023. Materials today chemistry.

Materials Today 29, 101478.

Talanov, M., Vallverdu, J., Adamatzky, A., Toschev, A., Suleimanova, A., Leukhin, A., Posdeeva, A., Mikhailova, Y., Rodionova, A., Mikhaylov, A., *et al*., 2022. Neuropunk revolution. hacking cognitive systems towards cyborgs 3.0. arXiv preprint arXiv:2205.06538 .

Whelan, T., Johannsson, H., Kaess, M., Leonard, J.J., McDonald, J., 2012.

Robust tracking for real-time dense rgb-d mapping with continuous.

Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D., 2017. Temporal action detection with structured segment networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2914–2923.

Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y., 2017. Flow-guided feature aggregation for video object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 408–417.