Hand Gesture Recognition System using Convolutional Neural Networks

Sri Mani Prudhvi Peddi
Computer Science Department
Texas A&M University Corpus Christi
Corpus Christi, Texas
speddi@islander.tamucc.edu

Dr.Longzhuang Li
Computer Science Department
Texas A&M University Corpus Christi
Corpus Christi, Texas
longzhuang.li@islander.tamucc.edu

***Abstract*—Gesture Recognition has the ability to interact with machines efficiently through Human computer interaction and to create easy-to-use interfaces by directly using natural communication.[3] Hand recognition is a very significant recognition for the human computer interaction. Challenging problems are complex background, camera angle, illumination. This system introduces a fast and robust system for gesture recognition. This system is implemented using the deep learning models such as the Convolution Neural Network (CNN) for three-dimensional data. The combined model will effectively recognize both static and dynamic hand gestures. This model needs a camera for capturing data and we can train model to perform interactive tasks on the system. The response time should be faster and should be reliable and work for different people. Index Terms—Convolution Neural Network (CNN), Gesture detection, human computer interface, sign language, deep learning., Gesture detection, human computer interface.**

***Index Terms*—Convolution Neural Network (CNN)**

I. Introduction

Problems faced by Deaf and Dumb People Transportation: The transit system, in India is not planned in an inclusive way. The design of system is not designed to include, people from all sections of society, including differently abled. With given Indian system of transportation, whose major feature is poor dissipation of information, it becomes difficult for even a normal person to commute from one place to another. And, in case of people with low sound sensibility, it is very difficult for them to communicate and commute to their destination [1].

Road Crossing: The quantum of Indian traffic is very weird. People are always in hurry, but never reach their destination in time(which is irony). Due to this, they often cover the zebra crossing, which was meant for pedestrians to cross, due to "their hurry".

Communication: Hearing impaired people cannot telephone [2] and when they watch TV they have to be dependent on hearing people. Even we are deprived of the pleasure of watching TV and films and here also we need the help of a companion because movies and serials on TV do not have sub titles on them and it should be compulsory and mandatory. I may mention here that life has become easy for a few who have access to SMS through cell phones and by using Internet/emails. Not all enjoy these privileges.

Level of Dependency: The disabled person is largely dependent on a family because they get economically and emotional support from family. The disabled person also gets emotional support from friends and relatives and also gets encouragement.

Social Stigma: This is the difficult of all the challenges, faced by them. The physically handicapped face problems as they attempt to adjust the demands of living in social environment. Their problems are not only those caused by their disability but also that of adjustment in a world that has apathetic or hostile attitude towards them magnifies their troubles and threatens their very existence as human beings. They face psychological, educational, employment and social problems. Among these, the most difficult is the adjustment to the hostile social forces in the society (Sharma, 1981), disabled person suffers with the erroneous beliefs, which dry up their day-to-day way of life. It automatically generates a social resistance to accepting means of treating or ameliorating disability.

II Sign Language [6] A sign language is a way of communicating by using the hands and other parts of the body. [4]Sign languages are an important way for deaf people to communicate. Deaf people often use them instead of spoken languages. Spoken languages use sounds from the mouth and are understood with the ears. Sign languages use hands and are understood with the eyes. Deaf people can use sign languages more easily than spoken languages.

[1] Deaf people sometimes learn a sign language from their family, especially if their parents are deaf. But, most deaf children have hearing parents, so they learn a sign language from other deaf people. They may meet other deaf people at school or in the streets. [7] Hearing people may learn to sign directly from deaf people. Or, they may learn a sign language by going to signing classes or by studying a sign language workbook, which can come with an interactive DVD. [Sometimes deaf people do use a spoken language, especially when talking with hearing people. Sometimes hearing people use a sign language with each other, rather than speaking. But, deaf people tend to use sign language, and hearing people tend to use spoken languages.

Some deaf people can also understand spoken words by looking at a speaker's lips. This is known as lip-reading.

It is hard to learn, and few people do it well. Sometimes signing and lip-reading are combined, especially when deaf and hearing people talking.

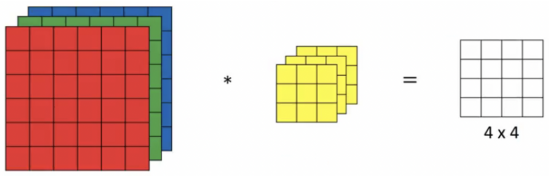
Some of the Pros and Cons of this concept are, Speed and sufficient reliable for recognition system. Good performance system with complex background. The radial form division and boundary histogram for each extracted region, overcome the chain shifting problem and variant rotation problem. Another pros will be the exact shape of the hand obtained let to good feature extraction, fast and powerful results from proposed algorithm. The system successfully recognizes static and dynamic gesture[11].

The Irrelevant object might overlap with the hand, wrong object extraction appears if object is larger than the hand performance recognition algorithm decreases when the distance is greater than 1.5 meters between the user and the camera. the proposed method is susceptible to errors especially in shapes like square and circular. System limitations restrict the applications such as: Gestures are made with right hand only, the arm must be vertical, the palm is facing the camera, background is plain and uniform. The system does not reflect the dynamic gesture characteristics.

1) Interpretation: In order to facilitate communication between deaf and hearing people, sign language interpreters are often used. Such activities involve considerable effort on the part of the interpreter, since sign languages are distinct natural languages with their own syntax, different from any spoken language.

2) Remote interpreting: Interpreters may be physically present with both parties to the conversation but, since the technological advancements in the early 2000s, provision of interpreters in remote locations has become available. In video remote interpreting (VRI), the two clients (a sign language user and a hearing person who wish to communicate with each other) are in one location, and the interpreter is in another. The interpreter communicates with the sign language user via a video telecommunications link, and with the hearing person by an audio link. VRI can be used for situations in which no on-site interpreters are available.

[5] However, VRI cannot be used for situations in which all parties are speaking via telephone alone. With video relay service (VRS), the sign language user, the interpreter, and the hearing person are in three separate locations, thus allowing the two clients to talk to each other on the phone through interpreter.



The RGB dimensions [1] above represent the height, width and channels in the input and filter. Keep in mind that the number of channels in the input and filter should be same. This will result in an output of 4 X 4.

[11] Since there are three channels in the input, the filter will consequently also have three channels. After convolution, the output shape is a 4 X 4 matrix. So, the first element of the output is the sum of the element-wise product of the first 27 values from the input (9 values from each channel) and the 27 values from the filter. After that we convolve over the entire image.

• Input: n X n X nc • Filter: f X f X nc • Padding: p • Stride: s • Output: [(n+2p-f)/s+1] X [(n+2p-f)/s+1] X nc' Here, nc is the number of channels in the input and filter, while nc' is the number of filters.

II. Related Works

In a world of almost 7 billion people more than 500 million suffer from some physical, sensory or mental disability. Their lives are often impeded by such deformities which bar them from full participation in society and the enjoyment of equal rights and opportunities. Sign language is common for the deaf and the dumb. Sign language is an efficient alternative to talking, where the former is replaced by hand gestures. Hand gestures are combination of hand shapes, orientations and movement of the hands, alignments of the fingers and positioning of the palm which are used to express fluidly a conveyor's thoughts. Signs are used to communicate words and sentences to audience. The objective of this paper is to optimize an algorithm for recognition of hand gestures with reasonable accuracy, where the input to the pattern recognition system will be given from the hand.

A. Blind/Deaf Communication API for Assisted Translated Educational Digital Content

With the rise of usage in digital content in education, deaf and blind communities face communication barriers which as a result makes education less inclusive. These barriers do not allow them to integrate within the larger scholarly communities as most tools used for information dissemination remain inaccessible to them. This paper presents BDC-API (Blind/Deaf Communications API), a free-to-use modular toolkit that will ease accessibility for the blind and deaf communities to digital education content. This content includes the use cases of Massive Online Open Courses and Serious Games used in education. BDC-API incorporates the use of state of the art technologies such as,[12] 3D sign language

translator, grammar translation, voice recognition and text-to-speech. This paper demonstrates in greater detail, how these technologies culminate in the creation of an API ready to use for any educational digital content and how the BDC-API can ensure higher quality of digital content.

A Visual Recognition of Static Hand Gestures in Indian Sign Language based on Kohonen Self- Organizing Map Deepika Tewari, Sanjay Kumar Srivastava in their paper, presented an algorithm for hand gesture recognition which is further used in Indian sign language recognition system. In their system, self-organizing map (SOM) is used for classification purpose. In their approach, first of all the hand image is segmented by utilizing the pixel value for skin and background. Background is represented by the black color. Kohonen self-organizing map is commonly used for this very purpose. In their method, They have used single layer SOM which is feed forward [10] neural network. Since it does not require any prior information about the data therefore it can be used for clustering the data without number of class.

B. Hand Gesture Recognition Using K-Means Clustering and Support Vector Machine

Human-Robot Interaction (HRI) requires[5] media for communication which can be both understood by robot and easily done by human. Usually, human using oral language to communicate but there are some situations that require.

Performing non-verbal activities such as deaf people, patient, and old people, therefore gesture recognition as communication media is needed to give order to Robot. This paper discusses hand gesture recognition as input command for Bioloid Premium Robot using two methods, K-Means clustering and Support.

Vector Machine (SVM) with [5]directed acyclic graph (DAG) decision. Four gestures (forward, right, left and stop) were recognized using Kinect v2. The testing was done 6 peoples for three distances (2m, 3m, and 4m) and three slopes position (45, 0, -45). The SVM required 10ms recognition time with accuracy reached 95.15%, while K-Means needed 4.45ms recognition time with 77.42% accuracy. This study resulted in Multiclass SVM with DAG decision performs better than K-Means clustering method.

C. Sign-Voice Bidirectional Communication System for Normal

”Deaf/dumb” and blind people face problems in communicating with others with difficulties in dealing with the communication technology. The goal of this paper is to design a desktop human computer interface application that is used to facilitate communication between normal, ”deaf/dumb” and blind people. SVBiComm system helps blind person to hear voice saying the word gestured by the ”deaf/dumb” while the deaf will receive a gesture representing the word said by the blind. SVBiComm works in two directions, the first direction is processing from video to speech. The animated word gestures are mapped with language knowledge base into text. Then, the relevant audio is generated using Text-to-Speech (TTS) API. The second direction is processing from speech to video. The voice from blind is converted into its corresponding text using Speechto-Text (STT) API. Then, the natural language is mapped from the database to ”deaf/dumb” in a relevant sign language form by using a 3D graphical model. The system was evaluated using a set of 113 sentences with 244 signs. In voice recognition; system recognized words with a percentage of 90% from 19 different persons. For Image recognition the system recognized images with a percentage of 84% for 21 different persons. SVBiComm system provides many facilities with low cost that could be used in many areas.

D. Sign Language Recognition Using Intrinsic-Mode Sample Entropy on sEMG and Accelerometer Data

Vasiliki E. Kosmidou, and Leontios J. Hadjileontiadis, in their paper, suggested electromayogram and 3-D accelerometer based sign language recognition system. In their method, five-channel surface electromayogram and 3-D accelerometer are used to get the data from the dominant hand of signer‘s. These data is then analysed using intrinsic mode entropy(IMEn) for recognizing the sign language. They have used this method for Greek sign language. Isolated signs are used for recognizing the sign language. Since gesture of the body is directly related to the hand movement therefore measurement of hand movement reflects the gesture of the body. This fact is the motivation behind exploring the possibility of using electromayo gram and 3-D accelerometer here is used to capture the hand movement while the wrist and finger motion is obtained by corresponding muscles in the arm.

Novel Segmentation Algorithm for Hand Gesture Recognition Dhruva N. and Sudhir Rao Rupanagudi, [1] suggested in hand gloves worn to recognize the sign language. They have used white color woollen hand gloves for better accuracy. Since in sign language each and every fingers represent unique message therefore it is very important that each and every fingers must be segmented out clearly.

III. Proposed Work

In existing, they tackled background problems, variation in lighting condition problems. And edge detection-based hand gesture recognition has the issues such as false edge detection, broken edges, high computation time and less recognition accuracy.To solve the background and lighting problems. Use effective method for reduce the computation time and increase the recognition accuracy. The proposed system consists of modules that capture the real time data using the web camera. The captured image is further preprocessed to make the constant image size, converting the image into gray scale image and binary image etc. The binary image is further background subtracted using the reference image through Gradient mixture modeling (GMM). The object that is semantic with the capture image is highlighted. The semantic area of the binary image is blob. The blob image is further transferred to the proposed CNN architecture. The proposed model is trained with certain types of image blobs. The highly correlated pattern is detected

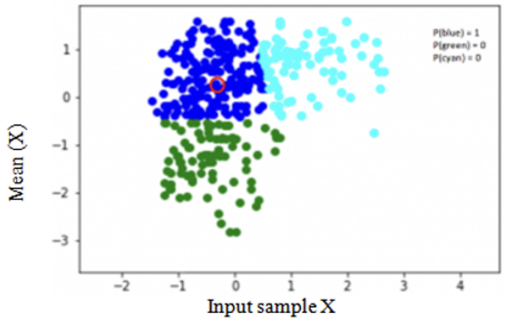
at the outcome of the proposed CNN-Res-Net architecture. The overall performance of the network is measured through accuracy, precision and recall.

1) Feature Extractions: In machine learning, pattern recognition and in image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.

When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meters, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

A. Proposed Methods

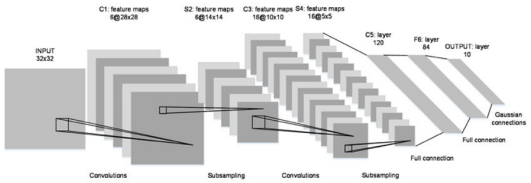
1) Gaussian Mixture modeling (GMM): Gaussian mixture model (GMM) is a [9] kind of clustering algorithm. Each clustered semantic object is generated with unique Gaussian distribution. Each Gaussian model is grouped together with the relevant data. If the given image input consists of three unique Gaussian distribution points say, g1, g2, g3, then equivalent mean values will be u1, u2, u3 and the variance of the given g1, g2, g3 will be v1, v2, v3 etc. Gaussian mixture models use the soft clustering approach for segregating the unique clusters and their equivalent distribution factors.



Representation of GMM data of few inputs

2) Multi-layer Convolution Neural Network (MLCNN): Convolution neural networks are a deep learning algorithm [9] which considers the input image and learn the unique information from the image through weight or bias assignment in terms of various aspects. The weight assignment is based on image pattern, shape and size to show the differentiation. The architecture of MLCNN is shown above. The MLCNN structure is analogues and resembles the pattern of neurons that is organized in the virtual cortex. The CNN inputs the image of NxN pixels as two-dimensional patterns. The MLCNN capture the spatial and temporal dependencies of the input image by using the features of the image. These features are nothing but the partial parts of the input image. These feature points are unique for certain images. From the input image, the convolution filter is selected as MxM array. These parts of the filter is called as Kernels. The system consists of RELU layer, POOLING layer and combination of both. The fully connected layer flattens and accumulates the feature mappings of the all inputs as a feature vector and finally the highest correlated points act as the decision score.

CNN Architecture.



3) SYSTEM DESIGN DATA FLOW DIAGRAM: The DFD is also called as bubble chart. It is a simple b. graphical formalism that can be used to represent a system in t. terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.

1. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

2. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.

3. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

Convolution Layer

The values of matrix elements in the given test images are updated by optimum values as it goes through the training session of the MLCNN. Convolution means, dot product of pixels with every matrix value, The filter matrix, slide the kernel window over the image and find a new matrix called convolved image. The element wise operation is continued till the entire image matrix. The similar operation is repeated and applied to multiple kernels in one image. Multiple kernels being applied to one image will result in multiple convolved matrices.

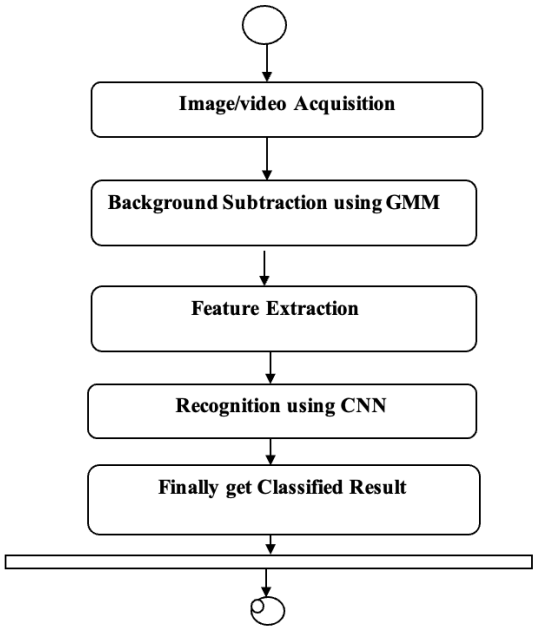
Max-Pooling Layer

Max-pooling layer is used to perform dimensionality reduction. Max pooling identifies the highest values that is unique in the input image and assign the equivalent pixel value to it. It converts the higher data size of input image and converts into smaller image size with unique parameters. These values after the kernel point we call as feature maps.

Fully connected layer

The fully connected layer captures the high quality unique data from the given values and provides the computationally robust parameter values in each kernel. These values are statistically applied to find mean, median and variance. The unique data at the end provides the metric score of the given input.

4) ACTIVITY DIAGRAM: Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



B. Dataset and Evaluation Metrics

Gesture images are dynamic in nature. The hand gesture data varies with changes in external lighting, intensity etc.[7] The dynamic changes enable the object gesture vary with different scales. For the proposed work, Real time live images are captured using web camera. The unique gesture points and the landmarks of the gesture, shape and size etc. The image acquisition is collected with even light intensity and gathered as the dataset. For training purpose same gesture is captured N number of times as samples. Each gesture is formulated with the labels as well.

The evaluation metrics of the proposed system is achieved using the four factors namely, true positive rate, true negative rate, false positive rate, false negative rate etc.

False Positives: pixels that belong to the background that were misclassified as belonging to lesions;

False Negatives (FN): pixels that belong to lesions that were misclassified as belonging to the background;

True Positive: pixels that belong to lesions that were correctly classified as belonging to lesions;

True Negative (TN): pixels that belong to the background that were correctly classified as belonging to the background.

The global accuracy is measured using the following formula

Foundations of Convolutional Neural Networks
The objectives behind the first module of the course 4 are: • To understand the convolution operation • To understand the pooling operation • Remembering the vocabulary used in convolutional neural networks (padding, stride, filter, etc.) Building a convolutional neural network for multi-class classification in images

C. Difference in methodologies

In the existing system, MLCNN architecture with primitive feature extraction process is done. The feature extraction part in the existing system extract the alignment angle of the input image as texture. These parameters are fetched into the proposed CNN for classification. The drawback found in the existing architecture is computationally complex and overfitting is the issue.

In the proposed model keeping these points in mind a residual network with CNN architecture is framed. The proposed architecture is nothing but a customized network that adopt the layer performance through tunable factors such as weight adjustment, loop adjustment classifies the input image with less complexity and less processing time. The over fitting issue is not found with the proposed one due to the tunable factors available.

Parameters Value	Existing CNN	Proposed RESNET
Hidden Layer	7	7
Dropout Layer	1	0
Number of filters	3264	3264
Batch size	32	64
Activation Function	reLu	reLu
Optimizer	Adam	Xception
Epochs	30	50
Kernal Size	(3, 3)	(5*5)
Size of pool	(2, 2)	(3*3)
Strides	(2, 2)	(2X)

D. Performance of Existing methods

Dataset	Real time images	Real time images
Algorithm	Multi-layer CNN	CNN-Residual Network
Optimizer	Adam	xception
Epochs	30	50
Accuracy	99.13	99.25

IV. CONCLUSION

Gesture detection becomes one of the interesting area of research since many AI-Artificial intelligence models depends on the gesture inputs. Effective recognition of hand gesture is the goal of computer vision to interpret and manage the classification with high accuracy. The proposed system is evaluated with such criteria, using Gaussian mixture model (GMM) based Convolution neural networks (CNN). The ensemble version of GMM-CNN enables the system to interpret the hand

gestures effectively. The system considers both static and dynamic input modeling and hence the performance achieved is comparatively effective. The proposed GMM-CNN enhanced gesture detection system achieved 99.25% accuracy with real-time images using Xception model and 99.13% accuracy with Multi-layer CNN Adam optimizer.

Further the hand gesture model need to get improvised using smart devices through AI enabled applications, with improved accuracy using YOLO learning models.

V. Future Scope

Macpi board can be used it has 3 GB inbuilt ram and an inbuilt GPU, hence the speed will improve. Real images can be captured easily. The new method should involve both hands. Deep learning method can be used for greater accuracy but deep learning cannot be implemented in raspberry pi.

VI. REFERENCES

[1] Yiwen He, Jianyu Yang, Zhanpeng Shao, Youfu Li, "Salient feature point selection for real time RGB-D hand gesture recognition", IEE International Conference on real-time Computing and Robotics, 2017.

[2] Rania A. Elsayed, Mohammed S. Sayed, Mahmoud I. Abdalla, "Hand gesture recognition based on dimensionality reduction of histogram of oriented gradients", Japan-Africa Conference on Electronics, Communication and Computers, 2017.

[3] HimadriNathSaha, SayanTapadar, Shinjini Ray, Suhrid Krishna Chatterjee, "A Machine Learning based approach for Hand Gesture Recognition using distinctive feature extraction", IEEE 8th Annual Conference on Computing and Communication, 2018.

[4] M. Al-Hammadi et al., "Deep Learning-Based Approach for Sign Language Gesture Recognition With Efficient Hand Gesture Representation," in IEEE Access, vol. 8, pp. 192527-192542, 2020, doi: 10.1109/ACCESS.2020.3032140.

[5] S. Tam, M. Boukadoum, A. Campeau-Lecours and B. Gosselin, "A Fully Embedded Adaptive Real-Time Hand Gesture Classifier Leveraging HD-sEMG and Deep Learning," in IEEE Transactions on Biomedical Circuits and Systems, vol. 14, no. 2, pp. 232- 243, April 2020, doi: 10.1109/TBCAS.2019.2955641.

[6] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang and L. Yuan, "Hand Gesture Recognition Using Deep Feature Fusion Network Based on Wearable Sensors," in IEEE Sensors Journal, vol. 21, no. 1, pp. 539-547, 1 Jan.1, 2021, doi:10.1109/JSEN.2020.3014276.

[7] C. Liu, Y. Yang, X. Liu, L. Fang and W. Kang, "Dynamic- Hand-Gesture Authentication Dataset and Benchmark," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1550-1562, 2021, doi: 10.1109/TIFS.2020.3036218.

[8] N. Siddiqui and R. H. M. Chan, "Hand Gesture Recognition Using Multiple Acoustic Measurements at Wrist," in IEEE Transactions on Human-Machine Systems, vol. 51, no. 1, pp. 56-62, Feb. 2021, doi: 10.1109/THMS.2020.3041201.

[9] S . Shin and W. Kim, "Skeleton-Based Dynamic Hand Gesture Recognition Using a Part-Based GRU-RNN for Gesture-Based Interface," in IEEE Access, vol. 8, pp. 50236-50243, 2020, doi: 10.1109/ACCESS.2020.2980128.

[10] W. Zhang, J. Wang and F. Lan, "Dynamic hand gesture recognition based on short-term sampling neural networks," in IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 1, pp. 110-120, January 2021, doi: 10.1109/JAS.2020.1003465.

[11] S. Skaria, A. Al-Hourani and R. J. Evans, "Deep-Learning Methods for Hand-Gesture Recognition Using Ultra-Wideband Radar," in IEEE Access, vol. 8, pp. 203580-203590, 2020, doi: 10.1109/ACCESS.2020.3037062.

[12] B. Qiang et al., "SqueezeNet and Fusion Network-Based Accurate Fast Fully Convolutional Network for Hand Detection and Gesture Recognition," in IEEE Access, vol. 9, pp. 77661- 77674, 2021, doi: 10.1109/ACCESS.2021.3079337.

[13] Dursun, C., Erdei, T. and Husi, G., 2020. Artificial Intelligence Applications in Autonomous Vehicles: Training Algorithm for Traffic Signs Recognition. IOP Conference Series: Materials Science and Engineering, 898, p.012035.

[14] Edward, M., 2021. African sign languages are not American product: Indigenous African Deaf People and indigenous African Sign Languages. Academia Letters.