Trends in the detection and interpretation of human movements and mime-gesture

Andrei Enachi, Cornel Turcu
Faculty of Electrical Engineering and Computer Science
Ștefan cel Mare University of Suceava
Suceava 720229, Romania
Faculty of Energetics and Computer Science
Vasile Alecsandri University of Bacău
Bacău 600115, Romania
andrei.enachi@ub.ro, cturcu@usm.ro,

George Culea
Faculty of Energetics and Computer Science
Vasile Alecsandri University of Bacău
Bacău 600115, Romania
gculea@ub.ro

***Abstract*— Communication is an important part of our lives. Communication facilitates interaction between people, allowing us to share our ideas and emotions. The research has focused mainly on one side of communication, namely the conversion of sign language into speech. Contrary to traditional approaches, which focus on using static data to convert sign language into speech, our research brings an innovative perspective by integrating artificial intelligence technologies. This research has largely used static data sets such as alphabets, digits and specific gestures taken from a variety of languages. However, addressing this problem is done using artificial intelligence techniques and methods. In this context, detailed research will be presented, focusing on techniques and methods developed to improve the communication using artificial intelligence (Mediapipe, OpenCV, CNN). The contributions presented in this paper represent a significant advance in the field of communication bringing to the forefront cutting-edge methods and techniques in an evolving field.**

Keywords—artificial intelligence, CNN, mime-gesture, OpenCV, Mediapipe

I. Introduction

In this paper is presented the actual state of the hand gesture recognition methods. The objective is to observe each technique and identify the proper one suited for sign language recognition that could easily realize the conversion to text or speech. This approach will not only concentrate on advancing innovative solutions to enhance the interaction of individuals with hearing or speech disabilities with the broader society but also aims to revolutionize the way we perceive and address accessibility challenges. By leveraging cutting-edge technologies and methodologies, the research endeavors to break down communication barriers, empowering individuals with hearing or speech disabilities to fully participate in social, educational, and professional spheres. The anticipated outcomes of these research efforts hold the promise of significant social impact, fostering inclusivity and diversity while enhancing the overall quality of life for individuals with hearing or speech disabilities. Through the implementation of these innovative solutions, we aspire to create a more equitable and accessible environment where every individual, regardless of their abilities, has equal opportunities for engagement, expression, and fulfillment. With their advantages and disadvantages in terms of segmentation and feature extraction (such as the use of YUV, YCbCr, RGB color spaces, histograms, median particle filters, noise filters, thresholding, Gaussian model, HMM model), and arrived at the conclusion that the best performing solution is the use of convolutional neural networks. These convolutional neural networks perform better because they can process a larger amount of information and thus achieve a higher and faster rate of gesture recognition. Neural network-based technologies offer advantages in handling data complexity and in the ability to identify and understand subtle patterns in gestures. Thus, this approach has the potential to reduce delays during the gesture recognition process and improve overall system performance [16-20].

II. Related work

A. Image Capture

Image capture is done directly through the camera, in some accessible way, in Python. The OpenCV (Open Computer Vision) library is used because is powerful enough for the mime-gesture and sign language recognition application, also TensorFlow.

OpenCV, short for Open Source Computer Vision Library, is an open source library developed to provide advanced image processing and computer vision tools. This library, originally developed by Intel in 1999, has rapidly expanded and become an industry standard in the computer vision field. OpenCV provides a wide range of functionality for image analysis and manipulation, as well as for developing computer vision applications. Key features include support for more than 2,500 optimized algorithms covering areas such as object detection, facial recognition, image filtering and correction. OpenCV is written in C++ and has links to various programming languages, including Python and Java. This versatility makes OpenCV accessible to a variety of developers and applications. The modular structure of the library facilitates integration and specific use of functionality, allowing developers to select only the components needed for their project. OpenCV stands out for its high performance, being optimized to work efficiently and quickly even in scenarios with large amounts of data. Support for hardware technologies such as graphics processing units (GPUs) adds an extra level of efficiency in image-intensive applications. The use of OpenCV extends across a wide range of industries, including robotics, medical, automotive, security and more. In the medical industry, for example, OpenCV is used for medical image analysis and computer-aided diagnosis. In the automotive industry, it is involved in driver assistance systems and traffic sign recognition technologies. OpenCV benefits

979-8-3503-4929-0/24/$31.00 ©2024 IEEE

from an active and diverse international community, providing support through online forums, educational resources and detailed documentation. This community contributes to the ongoing development of the library and the continuous improvement of its functionality, thus strengthening OpenCV's position as a leading resource in computer vision. OpenCV remains an essential choice for developers involved in computer vision and image processing projects. With a modular structure, high performance and an active community, OpenCV continues to be a vital tool for the exploration and implementation of computer vision solutions in various industrial and scientific fields.

B. Training models

TensorFlow, developed by Google, is at the forefront of open-source libraries for developing artificial intelligence (AI) and machine learning (ML) applications. Originally released in 2015, TensorFlow underwent a significant transition with the release of version 2.0, with a focus on simplifying and improving the developer experience. This update introduced the Keras module, providing a high-level interface for building and training models, increasing accessibility for less experienced developers. The notable aspect of TensorFlow 2.0 is the "eager execution" approach, allowing developers to evaluate and modify TensorFlow expressions interactively, eliminating the need for an explicit session. This facilitates development and debugging, thus speeding up the development cycle. TensorFlow retains its graph-based architecture, with developers defining models by constructing a graph describing the relationships between mathematical operations. In TensorFlow 2.0, the Keras module simplifies this process, providing a more accessible interface for defining and training models. The increased modularity of the platform facilitates collaborative development and encourages component reuse, contributing to the efficiency of development efforts. TensorFlow is optimized to take advantage of hardware resources such as graphics processing units (GPUs) and tensor processing units (TPUs). This allows developers to accelerate model training, efficiently tackling large problems. The TensorFlow Serving Platform simplifies model deployment in production, providing a robust infrastructure for serving ML models in a scalable way. This enables easy integration of models developed in TensorFlow into applications and services, providing complete solutions for AI-based application deployment. TensorFlow enjoys an active and growing community, supported by online forums, educational resources and comprehensive documentation. Extensive support for TensorFlow in popular development environments such as Python contributes to the platform's continued popularity and adoption among developers. TensorFlow 2.0 remains a reliable choice for artificial intelligence and machine learning application development. With features such as "eager execution", modularity, and extended support for hardware acceleration, Tensor-Flow continues to provide a powerful and affordable platform for those involved in implementing AI-based solutions.

III. APLICATION, CNN AND MEDIAPIPE

A. Aplication and discussion

The application recognizes hand signs (for starters the peace sign, OK, Like, Rock and Roll or I love you) [1]. The first component of the system that picks up the frames uses OpenCV for this. How do we build this system that picks up these frames (images containing hand gestures)? It starts from the idea that hands are of different sizes and shapes similar to gestures (gestures used with the left/right hand) depending on the context in which they appear, which is almost impossible to implement with a mathematical algorithm. Also, the possibility of specifying the hand with which the gesture is performed leads to the use of artificial intelligence. The Deep Learning approach uses deep neural networks also called the first pixel approach. It uses matrices from the RGB color space (a method of representing images on computers) found in the OpenCV library to capture information from images. This information is then passed through a convolutional neural network which is a special type of neural network optimized to work with images, resulting in gesture classification (first approach) [4]. The human movement and mime-gesture recognition application is based on experimental data obtained through a webcam connected to a desktop and images taken from people with hearing and speech disabilities. The images (dynamic or static, video streams) were captured in the old people's home in Bacău under the coordination and guidance of Father Gabriel Ichim Radu, mime-gesture language interpreter, president of the National Association of Deaf Children in Romania and parish priest of the Church for the hearing impaired, "Saint Andrew of Crete" in Bacău.

Based on the Tensoflow, OpenCV and MediaPipe libraries, an application was developed that can recognize gestures made by a person. The algorithm behind the application is based on convolutional neural networks (three such networks are currently in use) needed to recognize user gestures with high accuracy. To begin with, the training of the algorithm for hand gesture recognition was carried out by dynamically storing frames and image sequences through the camera in order to achieve a classifier, robust enough to recognize the letters A-J. For this, the images used at the beginning (initial database) were mine images while the hand was in front of the camera and entering the gesture storage mode for the subsequent training of the final classifier. The images are stored continuously, this is ensured by implementing a loop, otherwise a black screen is displayed without being able to capture any gesture. Each gesture is labelled with numbers from 1 - 15, corresponding to each letter, these numbers as well as their labels are retrieved in a CSV file.

After entering storage mode by pressing the „ ,'' key, it was possible to enter 15 letters. The code for gesture recognition is written in Python programming language, and it was necessary to process it by implementing an additional line in a „for" loop that allows entering label numbers and store more gestures up to and including the letter J. For the first preliminary tests, it was observed that the algorithm for the first five letters of the alphabet, knows an accuracy rate of up to 95% for each letter, gradually changes when introducing letters up to and including the letter J. After the introduction of all letters, the training of the classifier gives very good results, with an accuracy for most letters of 94%, with very

little occlusion between gestures and similarity between them (letter o with letter c, letter d with letter b, letter f with letter b), these problems being completely reduced by training with a new data set obtained from the old people's home. Also, some letters as well presented a higher degree of difficulty in terms of their recognition by the classifier. As the classifier is trained it considerably increases its accuracy rate for each gesture (for some the accuracy rate is 1:1). Frames and photo/video sequences were taken at the nursing home where they were performed this time by native hearing and speech disability. It was very interesting to discover the small details that make a difference in learning sign language.

The letters performed and captured on camera by the people in the nursing home, were sufficient to improve the accuracy rate of the gesture classifier considerably. At the same time, the branch of this standard was explored for a more complex understanding of the mime-gesture and signs language with small variations of the letters executed by the persons concerned, were established being slightly different and strictly dependent on their area of origin (locality).

The results obtained were very convincing demanding to strive even further into recognizing human motion, sign language and develop application that helps people with hear and speak disability to integrate more easily into today society. The hand gesture recognition methods described above have been analyzed, and were observed that each technique has advantages and disadvantages in terms of segmentation and feature extraction (such as the use of YUV, YCbCr, RGB color spaces, histograms, median particle filters, noise filters, thresholding, Gaussian model, HMM model).

B. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are an advanced branch of artificial intelligence and machine learning, specifically designed to address complex image processing and pattern recognition problems in computer vision.

CNNs have proven to be particularly effective in spatial data analysis, identifying meaningful features and patterns in images, as well as in other areas such as natural language processing. CNNs are composed of specialized layers, including convolution layers, pooling layers and fully connected layers. The convolution layers are the architectural core of CNNs and are responsible for detecting relevant features in images. These layers use convolutional filters to perform convolutions on input images, thus extracting local information such as edges, textures and specific details. The basic principle of how CNNs work is their ability to learn and extract feature hierarchies from input data. As information passes through the different layers, the network becomes able to identify and retain complex features, such as object contours or intricate textures. Pooling layers are often integrated to reduce the spatial dimension of the representation while maintaining meaningful features. This process helps to control the variety of parameters and speed up training time. Fully connected layers at the end of the network interpret the extracted features and use them to perform classification or regression, depending on the nature of the problem.

The second approach, contains several layers, similarly from the camera the images are received and passed through a convolutional neural network other than the one mentioned above. This neural network will not provide results on the hand gesture, but will extract the hand skeleton (key points and joints). The skeleton is then passed through another convolutional network with fewer feed-forward layers. The feed-forward neural network is simpler because no loops/cycles are formed between the connections between two nodes [5]. The information travels in one direction (forward) from the input nodes, passing through the hidden nodes and reaching the output node (result). This neural network provides the final classification through the hand feature extraction step. In this regard, the MediaPipe library is implemented which eliminates the classification problem. It remains to train the neural network based on the hand skeleton to recognize gestures much more easily than training from pixels (anatomical construction of the hand, skin color) [7]. Thus, for the recognition of gesture like, it is not necessary to train with different images, but only with the gesture executed in the same frame only with different orientations. It is also worth mentioning the existence of a necessary step namely the neural network that extracts the hand skeleton. The network that classifies the hand gestures does not require the analysis of all frames (e.g. analysis of each pixel for hand gesture recognition) but focuses on the hand which means, the hand will not be dependent on the background. It will segment and obtain the outline of the hand representing the area from where the information is needed to determine the recognition rate of the collected gesture. MediaPipe will detect the hand gestures until the hand skeleton is obtained and then, in the second part, the feed-forward neural network will be trained and obtained the gesture classifier [2].

The classifier is initially set up to recognize the gesture performed with one hand only, later it will be trained to recognize the gestures of both hands. Whether or not the hand is rotated, the gesture will be recognized. It can identify whether the user is right-handed or left-handed, depending on which hand is performing the gesture. In the application window running the hand gesture recognition model and the hand skeleton detection model, a filter for checking the recorded frames (FPS) has been added. A maximum of 20 frames per second are recorded due to the simultaneous running of the two models [6]. Also, if both hands are captured by the camera, the frame rate decreases to 17 frames per second as the processing is done by the computer processor [9].

C. Mediapipe

The model for the hand skeleton was made by MediaPipe from a portfolio of 30,000 images and 21 three-dimensional key points (joints) (x, y and z parameters).

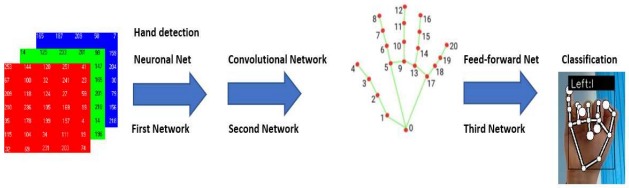

Fig. 1. Architecture of the recognition model with three different networks

For the estimation of several positions of the hand, a synthetic hand model is made regardless of background, lighting conditions and skin color corresponding to the three-

dimensional coordinates. Within the OpenCV library, the camera is declared in the application code and in order for the camera to capture continuously, a wait loop is implemented via a key in this case ESC, otherwise a black screen is displayed without the possibility of capturing images [10]. It is observed that the hand-skeleton based approach for the feed-forward model (used for gesture classification) solves the problems of hand position, shape, skin color, scaling by adding the preprocessing step. Hand skeleton values are defined as absolute pixel values. The distance or value for each of the 21 joints is calculated as the difference between the starting point of the joint and the wrist [8].

These values are converted to a one-dimensional list, representing the input data of the feed-forward neural network (one-dimensional vector). The vector values are found to be ambiguous, unclear and a normalization step is introduced. The maximum pixel value is taken and an absolute (normalized) pixel value is obtained regardless of negative values [11-15]. The final value being obtained by dividing the pixel value by the maximum value. The skeleton preprocessing is normalized with values between -1 and 1 respectively (how far the joints are positioned from the wrist). Recording a gesture is done directly from the application by saving the labelled frames from A-J [3].

IV. Results

A. *First Dataset of gestures with ages between 30 and 40*

For the first set, we used five gesture to familiarize the algorithm (we used 5 train gestures) with the environment constraint, occlusion like intensity of the light, skin color, distance between hand and camera. The accuracy of the model register 95% from the original stream, for all five gestures the recognition was above 92%.

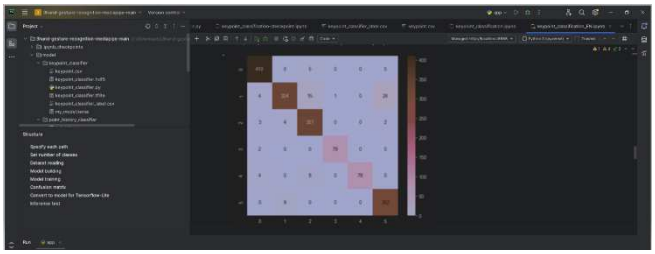

Fig. 2. Histogram of 5 gestures from first dataset

We dived deep and so we consolidated the initial database and in total we obtained 15 gestures collected this time from nine different individuals on three certain age categories (first 30 – 40, second 40 – 50 and third 50+).

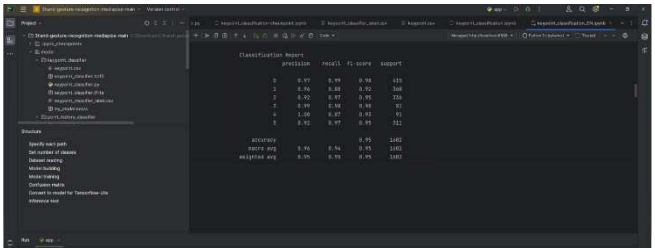

Fig. 3. Clasification report for five train gestures

We divided the results into three different datasets.

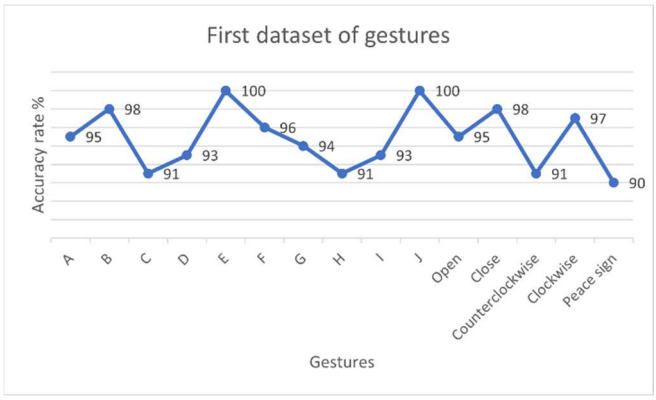

Fig. 4. Accuracy rate for the first dataset of gestures

B. *Second Dataset of gestures with ages between 40 and 50*

So the second and also the third dataset will be trained from a total of 15 gestures of the same individuals. We started with a budget camera, the lowest and cheapest to observe the capability of the algorithm in severe conditions like low resolution, dark pixels in broad light, occlusion from intensity of the light. Success in these less favorable conditions highlights the algorithm's potential to adapt to variability and suggests possibilities for implementation in low-budget devices or resource-constrained environments.

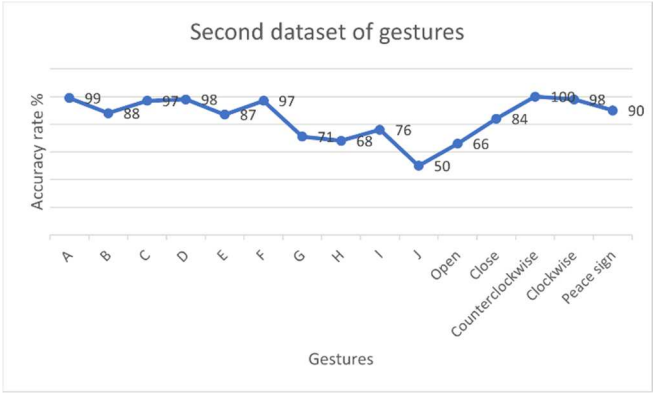

Fig. 5. Accuracy rate for the second dataset of gestures

The results obtained were beyond our expectation, almost all the occlusion from the light were fixed by training of the algorithm on certain gestures proving an accuracy of 94%.

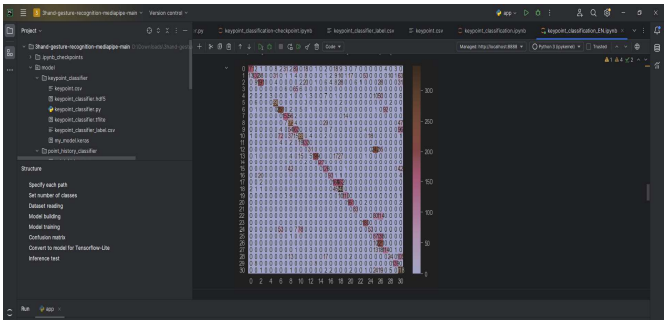

Fig. 6. Histogram for the trained gestures on all datasets

The only problem in recognizing the gestures were the fact that some letters were hard to recognize.

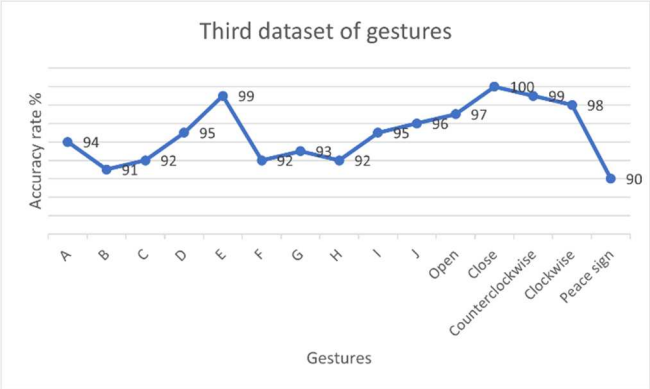

Fig. 7. Accuracy rate for the third dataset of gestures

The research conducted highlights the effectiveness of the algorithm in a diverse range of conditions, and the identification of specific challenges in recognizing certain gestures suggests the need for further investigation to improve performance and adaptability in varied contexts.

V. SUMMARY AND CONCLUSIONS

We concluded that the most effective solution is to use convolutional neural networks. These convolutional neural networks perform better, because they can process a larger amount of information and thus achieve a higher and faster rate of gesture recognition. CNNs have found extensive applications in a wide range of fields, including object recognition, image classification, semantic segmentation and even in medicine for computer-aided diagnosis.

Recent advances have led to the development of specialized architectures, such as ResNet, Inception and MobileNet, which optimize network performance and efficiency. Transfer learning technologies have also played a significant role, allowing pre-training models on large datasets and subsequent adaptation to specific problems with smaller datasets. This approach has been instrumental in reducing the need for massive training data for each individual problem. Convolutional neural networks have evolved significantly and have become an essential tool in computer vision. With their ability to extract complex features and solve complex pattern recognition tasks, CNNs continue to be a central component in the development of artificial intelligence-based solutions for visual data analysis.

Neural network-based technologies offer advantages in handling data complexity and in the ability to identify and understand subtle patterns in gestures. Thus, this approach has the potential to reduce delays during the gesture recognition process and improve overall system performance. Convolutional neural networks stand out for their remarkable efficiency in image manipulation, surpassing traditional methods. Their ability to identify and extract complex features from images makes them fundamental tools in a diverse range of applications, from object recognition to semantic segmentation. CNNs have become fundamental to solving machine vision challenges, especially in terms of pattern recognition in images. They are crucial to advances in technologies such as facial recognition, image classification and object detection. The applications of CNNs are expanding into various fields, including medicine, automotive, security and others.

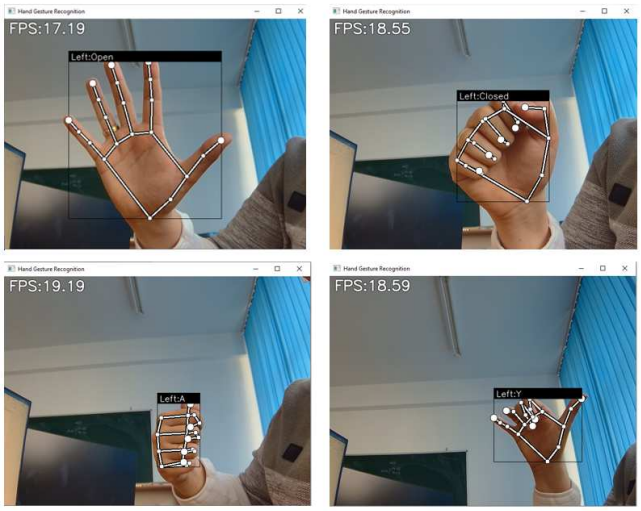

Fig. 8. Gestures from the training dataset

Above is the result of a pre-trained model, combining MediaPipe library and 3 ConvNets where a successful recognition rate of 96% were obtained for the presented gestures. Their ability to learn hierarchical and abstract representations makes them useful in a variety of contexts and industries. The development of specialized architectures such as ResNet, In-ception and MobileNet has led to significant improvements in the performance and efficiency of CNNs. These developments have addressed issues such as network degradation and facilitated their integration into resource-constrained devices. Transfer learning technology has made significant advances, allowing models to capitalize on their pre-existing knowledge to solve specific problems. This has reduced the reliance on massive datasets, which is crucial in situations where data is limited. Although CNNs have been remarkably successful, the field continues to evolve. Researchers and developers are exploring ways to improve existing architectures, optimize performance and extend applications into new areas.

Finally, convolutional neural networks are a central pillar in the development of artificial intelligence, paving the way for innovative applications in computer vision and representing a key tool in addressing the complexity of visual data analysis.

REFERENCES

[1] A. L. Sluÿters, S. & Vanderdonckt, Jean & Vatavu, Radu-Daniel, RadarSense: Accurate Recognition of Mid-Air Hand Gestures with Radar Sensing and Few Training Examples.," ACM Transactions on Interactive Intelligent Systems, 2023, doi: 10.1145/3589645.

[2] A. P. T. B. V. Chowdary, A. Sreeja, K. N. Reddy and K. S. Chandana, Sign Language Detection and Recognition using CNN, presented at the 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), 2023.

[3] S. D. S. De, S. K. Sabut, S. Biswas and S. Kar, Detection of Sign Language Alphabets Using a Weighted Ensemble of Neural Networks, presented at the 2023 International Conference on Communication, Circuits, and Systems (IC3S), 2023.

[4] N. Singla, American Sign Language Letter Recognition from Images Using CNN, presented at the 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2023.

[5] A. K. a. H. Canbolat, Prediction of Turkish Sign Language Alphabets Utilizing Deep Learning Method, presented at the 2023 5th

International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2023.

[6] A. J. S. Kurundkar, A. Thaploo, S. Auti and A. Awalgaonkar, Real-Time Sign Language Detection, presented at the 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies, 2023.

[7] S. P. S. J. P. Sahoo, S. Ari and S. K. Patra, Hand Gesture Recognition Using Densely Connected Deep Residual Network and Channel Attention Module for Mobile Robot Control, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, vol. 72, pp. 1-11, 2023, doi: 10.1109/TIM.2023.3246488.

[8] M. G. Y. N. Rajasekhar, C. Vedantam, K. Pellakuru and C. Navapete, Sign Language Recognition using Machine Learning Algorithm, presented at the Sign Language Recognition using Machine Learning Algorithm, 2023.

[9] A. K. M. N. K. Pandey, D. Singh, A. Jaraut and A. Bisht, "An Improved Hand Sign Recognition using Deep Learning," presented at the 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, 2023.

[10] P. A. R. K, P. K S, S. Sasikala and S. Arunkumar, "Hardware Implementation of Two Way Sign Language Conversion System," presented at the 2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP), BHOPAL, India, 2023.

[11] A. G. a. K. I. M. Priyankara, "Sign Language Translation Techniques Using Artificial Intelligence for the Hearing Impaired Community in Sri Lanka: A Review," presented at the 2023 7th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI), Kuliyapitiya, Sri Lanka, 2023.

[12] M. G. Y. N. Rajasekhar, C. Vedantam, K. Pellakuru and C. Navapete, "Sign Language Recognition using Machine Learning Algorithm," presented at the 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023.

[13] A. J. S. Kurundkar, A. Thaploo, S. Auti and A. Awalgaonkar, "Real-Time Sign Language Detection," presented at the Real-Time Sign Language Detection," 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), Vellore, India, 2023.

[14] S. R. G. V. Kandukuri, V. Kamble and V. Satpute, "Deaf and Mute Sign Language Translator on Static Alphabets Gestures using MobileNet," presented at the 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS), Nagpur, India, 2023.

[15] J. R. S. a. V. Vanitha, "Sign Language Detection Using Faster RCNN Resnet," presented at the 2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2023.

[16] N. S. S. Amutha, A. V. S. R. P. Naidu, P. V. Kumar and G. S. S. Narayana, "Real-Time Sign Language Recognition using a Multimodal Deep Learning Approach," presented at the 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023.

[17] M. E. J. B. A. Dabwan, Y. A. Ali and F. A. Olayah, "Arabic Sign Language Recognition Using EfficientnetB1 and Transfer Learning Technique," presented at the 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 2023.

[18] S. B. a. S. Bhamare, "Translating the unspoken Deep learning approaches to Indian Sign Language recognition using CNN and LSTM networks," presented at the 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS), Kanjirapally, India, 2023.

[19] C. S. U. P. D. Cerna, R. S. Evangelista, A. T. Darkis, M. S. Asiri and J. A. Muallam-Darkis, "An IOT-based Language Recognition System for Indigenous Languages using Integrated CNN and RNN," presented at the 2023 3rd International Conference on Smart Data Intelligence (ICSMDI), Trichy, India, 2023.

[20] N. V. S. K. B. Surya, A. S. SankarReddy, B. V. Prudhvi, P. Neeraj and V. H. Deepthi, "An Efficient Real-Time Indian Sign Language (ISL) Detection using Deep Learning," presented at the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023.