Real Time Sign Language Recognition Using Mediapipe Holistic

[1]Prof. (Dr.) Shital Raut
Faculty of Electronics and Telecommunication Engineering Vishwakarma Institute of Technology(VIT),
Pune, India
shital.raut@vit.edu

[2]Aditya Shinde
Department of Electronics and Telecommunication Engineering Vishwakarma Institute of Technology(VIT),
Pune, India
aditya.shinde19@vit.edu

[3]Niranjan Tapasvi
Department of Electronics and Telecommunication Engineering Vishwakarma Institute of Technology(VIT),
Pune, India
niranjan.tapasvi19@vit.edu

***Abstract*— Speech disability refers to a condition that hinders a person's capacity to communicate verbally. In order to address this challenge, sign language is utilized, which is regarded as one of the most structured forms of language. Communication mode of sign language relies on using hand gestures, body language and facial expressions as opposed to spoken words to convey messages in a visual manner. A system that is capable of recognizing language of sign movement, translating it into words can serve as a valuable tool for enabling communication with the deaf and mute community, especially for those who fail to communicate through language of sign. This paper presents a solution to facilitate communication with deaf and mute individuals in a more accessible manner. Proposed model uses Mediapipe Holistic which is a library provided by Google's Mediapipe framework, it processes the video data of hand gestures captured by a camera and provides features which are used to train Long Short Term Memory (LSTM) classifier to recognize the sign language gestures which then produces text. Characters collected are from the American Sign Language(ASL) which are alphabets from A to Z, digits from 0 to 9 and day-to-day words like "Hello" and "Thank you". Model is trained and gives training accuracy of 95% & validation accuracy of 92% respectively with a loss of 0.1229%.**

Keywords - American Sign Language, Mediapipe Holistic, Long Short Term Memory.

I. INTRODUCTION

Effective communication is crucial for connecting with the world around us, and those who lack it can appreciate its significance. An experiment conducted by Orfield Laboratories showed that an average individual could endure being in an entirely silent room for only 45 minutes.[1]. Consider, then, the experience of living in such a world constantly. Deaf and mute individuals face such challenges every day. As stated by the World Health Organization, approximately 466 million people worldwide, or about 5 percent of the population, live with such disabilities, including 35 million children.

Sign language was created to enable individuals with speech disabilities to communicate effectively. It is a highly organized language where each gesture holds a specific meaning and follows its own set of grammatical rules to connect words.

As per the most recent global statistics, approximately 72 million individuals who are deaf use sign language as their main mode to communicate[2]. However, language of sign is only understood by those who have prior knowledge of it. Consequently, communicating with people unfamiliar with sign language can pose challenges for individuals who rely on it. To facilitate communication between people with speech disabilities and the general public, there is a pressing need for a sign language detection method. This paper suggests using a way to detect language of sign movements. The method uses a video of a person's hand movements and analyzes it with a tool called Mediapipe Holistic. This module generates essential features such as hand orientations and positions, which are then utilized to train an LSTM classifier for recognizing sign language gestures. The LSTM classifier captures the temporal dynamics of sign language gestures and produces a recognition result, which is the predicted character text.

II. LITERATURE SURVEY

Paper proposed a system which can detect sign language using a webcam in real time[3]. Model is created using Deep learning and Tensorflow, Training of the model done my CNN. Software identifies and recognizes a range of hand gestures used in American Sign Language and then translates them into spoken or written language.

Research aims at identifying and distinguishing different hand gestures in real time given by American Sign Language data for the Deaf people in the US, Canada's English speaking areas[4]. SIFT algorithm is employed by the model to extract features and translate the gesture video into the corresponding letters of the alphabet. SIFT features are analysed at the edges & these characteristics are independent of image scaling and rotation.,

addition of noise, this enables correct march for keypoints between faces.

Proposed model[5] does real time sign language detection employing Convolutional Neural Network(CNN). System leverages a pre-trained SSD MobileNet v2 architecture that has been trained on a custom dataset to perform transfer learning for the task at hand. System is capable of identifying specific sign language signs with an accuracy ranging from 70% to 80% even in low light conditions and without a controlled background. Model has some limitations such as environmental factors like little light intensity & uncontrolled background which causes reduction in precision for detection.

Proposed system aims to create a two-way conversation with deaf people without the need of a translator so that people can talk to deaf people easily and deaf people can understand the spoken word. System provides speech as well as text output, It uses CNN model for alphabet classification[6]. To change text to speech Pyttsx, a library package for Python is used. For model training Keras, TensorFlow and scikit learn libraries are used. In all 32K images utilized for training, 8K for testing, model gives a training accuracy equal 99.65% , testing accuracy equal 99.62%.

Research aims to recognize sign gestures in real time using deep learning techniques. Specifically they are using the YOLOv5 algorithm to detect hand signs[7]. Model gives an accuracy of 88.4% with a labelled dataset Roboflow. Using CNN model gives accuracy of 52.98%.

Paper provides a solution for recognition of eleven kinds of hand gestures. System captures these hand gestures through a dynamic video. In order to detect skin color, identify hand contour from complex background, system uses the YCbCr color space transformation. Model[8] gives an accuracy of 95.1%. Model uses Computer Vision extensively in this research to Capture RGB image, translate it to YCbCr and then perform skin and contour detection.

Paper aims at providing real-time computer vision language of sign detection, deep learning techniques[9]. CV is used for image processing in which they pre-process the image and extract various hand gestures from the background. Images obtained are used for forming the data that contains 24 English language alphabets, This doesn't include letters 'J' and 'Z' as it requires movements but the dataset created is an image dataset. Dataset includes a total of 240 images as 10 images are taken per alphabet.After image processing steps data is passed to CNN model which is used for training. Accuracy obtained using this method is 83%.

Research aims at providing a unique solution to hand gesture recognition which is different from existing research. Framework named Mediapipe Hands is used which is a reliable finger and hand tracking device solution[10].It integrates multiple co-working models: It is a palm type acquisition model that works on a complete image and returns a fixed handheld binding box .It uses machine learning to understand 21 3D local hand landmarks from just one frame.It works with a cropped image location defined by palm detector and restores 3D reliable key points. Unlike other approaches which largely depend on powerful desktop this approach benefits real-time performance on mobile phones.To predict they have used KNN algorithm, Model can detect hand and produce coordinators and will be able to recognize letters (A-Z) with an average accuracy of 86 to 91% and achieves 95.7% accuracy for palm discovery.

Research provides a comparative study for gesture recognition[11]. Their dataset contains 22 handshapes which match 26 letters of the English language alphabet, 10 digits. System uses CNN architecture without using transfer learning. They have used their own dataset as well as a premade dataset named NZ ASL dataset to compare dataset performance. It was observed that premade dataset gave higher accuracy than custom dataset they created. The alphabet gesture accuracy of 82.5% and 97% validation set accuracy on digits. Whereas the custom dataset gave an accuracy of 67% for alphabets and 70% for digits.

Proposed system uses SIFT algorithm to detect and recognize various hand gestures corresponding to the alphabets in the English language[12]. System captures and saves real-time images in a designated directory, and extracts features from the most recently captured image to identify which sign gesture is made. Comparison will be made with the existing image previously saved for a particular letter in the directory. Keypoints are taken from input image and then matched with every image in the dataset and stores number of matched keypoints and after checking for each image in dataset,then it finds the highest keypoints matching image and then finally in gesture recognition step the image number is passed to a 1D array of 26 characters and the character correspond to the index value equal to the image number is picked and displayed in the interface.

Paper[13] discusses a sign language interpreter with a broad vocabulary who is capable of real-time continuous gesture recognition utilising a DataGlove. It further explains the statistical analysis that is based on gesture posture, orientation, position, motion. Additionally, it presents implementation of a prototype system with A preliminary system has been developed that has a vocabulary equal two fifty words in Taiwanese Sign Language (TWL). System employs hidden Markov models to recognize fiftyone basic postures, six orientations, eight motion primitives, which can be used to continuously recognize sentences of gestures in a signer-

dependent way in real-time. Average recognition rate is 80.4%, with weighted recognition rate equal 85.7%.

Paper[14] introduces two systems based on real-time HMM that are capable to recognize sentence from American Sign Language with use of one camera. The first experiment demonstrates how to interact with a computer that isn't moving using the system, producing an accuracy of 92 percent in words. The second experiment examines how this method can be employed for a wearable computer as a component of an ASL-English interpreter, achieving 98 percent accuracy.

Proposed system presents an ASL recognition system which is able to detect letters and finger-spelled words in real-time[15]. System utilizes depth sensor with limited resolution which is located in front of user near monitor. Hand and its orientation is segmented by the system through the utilization of the depth data. Letter classification is based on the maximization of the average neighbourhood margin process which is also based on the depth data of the hands. This study was focused on a use-case scenario where segmentation and recognition are only based on depth data.

Paper proposes technique for one of the south Indian languages to recognize sign languages[16]. It defines a group of 32 signs, the five fingers, each with two binary positions. Pre-processing of static images are carried out using a feature point extraction method, 10 photos are used to train the data for each symbol. Results obtained from the test images demonstrate that the sign language recognition model can check images with an accuracy of 98.125% when it is tested with 160 images after training with 320 images.

Research[17] describes system which monitors the uncovered and unmarked hands of a user of sign language. It utilizes skin color to identify the face and hand areas, then uses the Kalman filter to compute blobs and track the location of each hand. To segment skin color regions, dilate and erode processes are utilized. System is also capable of locating the face and hand regions.

Paper states progress of an system that is capable of detecting finger spelling in American Sign Language, Bengali Sign Language by utilizing glove that contains several strategically placed sensors[18]. Data was collected from five people with different sized hands while wearing the same glove. The glove was connected to an Arduino Mega micro-controller board that was wired on a veroboard. System demonstrated a notable accuracy of 96%.

Research provides, a context-sensitive statistical model for Taiwanese Sign Language is proposed, wherein gestures, postures can be accurately identified[19]. A hidden Markov model may quickly identify the postures by decomposing a gesture into a series of them. By combining probability from the hidden markov model with likelihood of every gesture in dictionary, the movement can immediately be identified in semantic manner in actual time. Four different methods are employed statistical classification, Neural network, matching template, time-space curves spline matching models for gesture recognition.

Paper[20] focuses on recognizing Indian signs in accordance with Real time dynamic hand gesture detection algorithms. Preprocessing via converting the captured video to the HSV color space and segmenting it based on skin pixels. Moreover, System incorporated depth information to achieve more accurate outcomes. It extracted Hu-Moments, motion trajectories from image frames, and then used Support Vector Machine to categorize gestures. System was tested using both a webcam and MS Kinect, and 4 gestures were classified with an accuracy rate of 97.5%, which was promising.

Research details the construction of system based on video for detection of language of sign[21]. Beam search employed to reduce computational complexity while recognizing sentences. System designed to detect language of sign sentences from a single signer, and is based on 97 signs from German Sign Language (GSL). Through experimentation, it has been shown that system can attain 94% accuracy when utilizing a lexicon consisting of 52 signs and all available features. However, when using a lexicon with 97 signs, the recognition accuracy decreases to 91.7%.

Proposed model[22] puts forward a 3D convolutional neural network to address issue, which can automatically extract temporal, discriminative spatial characteristics out of without any prior knowledge, from a raw video feed, bypassing designing of characteristics. To increase performance, 3D CNN is supplied with multi channel video feed containing color data, body joint position, depth clue. GMM - HMM achieved an average accuracy rate of 90.8 percent by combining trajectory, hand shape characteristics, It is better than utilising just trajectory or just hand shape features separately.

III. METHODOLOGY

Research suggests a deep learning based method to detect sign language in real time. System block diagram of the system in given in Fig. 1.

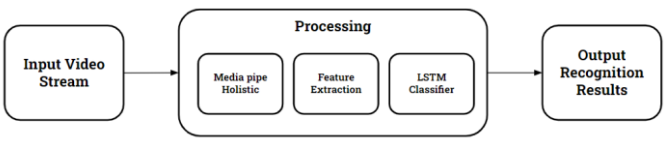

Fig. 1. System Block Diagram

A. Experimental Setup

The recognition of sign language was recorded prominently by 3 sign language character sets. These were sign language alphabets from A - Z, sign language digits from 0 - 9 and some commonly used words like “hello” and “thank you”. In total there were 38 distinct characters in the dataset.

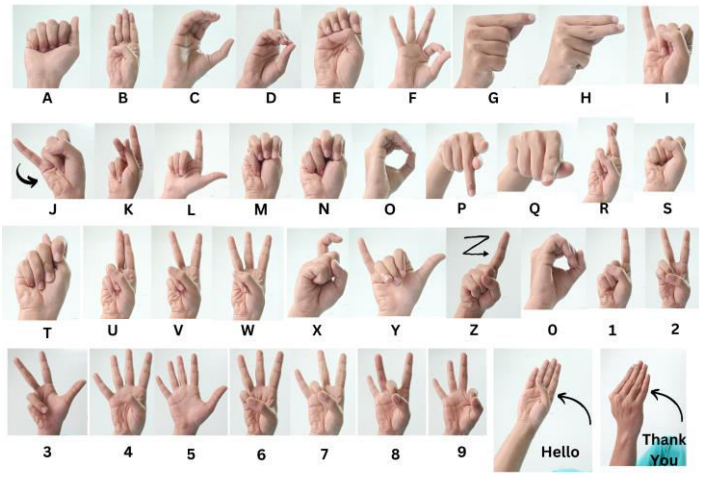

Fig. 2. Sign language gestures for the characters

The dataset contains 38 distinct characters, for each character a separate folder was created inside which the videos there were stored, inside each of 30 video folders there contained another 30 folders to store the frame data And each frame or image contains 1662 keypoints. Hence the dataset contained a total of 1140 videos.

Following are the steps taken to collect the data

Steps:

1. First, the model looped through the "actions" array, which contained all the actions to be detected
2. Next, for each action, the model iterated through 30 videos.
3. For each video, it further looped through 30 different frames.
4. Upon reaching the first frame of every video, the model displayed a message saying "STARTING COLLECTION”, indicating the start of a new video and a new hand gesture.
5. For each subsequent frame, the model displayed information about the identified action and the video number associated with that action, simplifying the tracking of video numbers.
6. Meanwhile, the model extracted key points from each frame and saved them into an array, which was then stored in the appropriate folder. This process happened simultaneously for each frame.

B. Data Collection

In this study, we collected our own data using the built-in webcam of a laptop, 30 videos were collected for each action and each video consisted of 30 frames. This input is a video stream of the signer's hand gestures which is processed by the Mediapipe Holistic module, it is a specific module within the Mediapipe framework that provides a holistic approach to pose estimation, including full-body tracking, hand tracking, and facial landmark detection. This module uses a multi-stage machine learning pipeline to estimate 3D coordinates of the different body parts in real-time. The module is useful for sign language recognition because sign language involves hand gestures, facial expressions, body movements, other non-manual features

The 38 sign language characters had their own folders with 30 subfolders, each containing videos of that particular action. Each video folder had 30 arrays, each containing 30 frames of data. Therefore, the dataset consisted of 1140 video files in total. Given that the dataset contains 1140 video files, and each of these video files has 30 frames, and each frame has 1662 keypoints, total keypoints in dataset are 56,840,400. Fig. 3. shows the detected keypoints comprising face landmarks, pose landmarks and right hand landmarks.

Fig. 3. Keypoint detection using mediapipe Holistic

To identify sign language characters, dataset used is the American Sign Language data. This particular dataset is considered to be highly effective and easily accessible in comparison to other sign languages used around the world.

C. Feature Extraction

Upon collecting and storing data in the designated folders according to the performed action, a label map is constructed with 38 distinct characters, labeled from 0 to 37. Subsequently, all frames associated with a particular action are assigned the same label. For instance, all 90 frames related to the action 'A' are labeled as 0, and this is repeated for all actions. The resulting data is then saved in the sequence array, which contains frames for all actions, and the labels array, which holds the corresponding labels for each frame.

D. Classification

Long Short-Term Memory (LSTM) classifier is employed to recognize sign language gestures, which are a kind of movement that happens over time. LSTM can remember how the hands move and use this information to predict what sign is being performed.

The neural network contains three LSTM layers, which are specialized recurrent neural network layers capable of capturing temporal dependencies in sequential data. The LSTM layers are configured with different numbers of units from the input sequences, significant features should be extracted. First layer contains 64 units, Second layer contains 128 units, third layer contains 64 units. For the first two layers model retains sequential nature of data, which is essential for subsequent layers to process input effectively. Rectified linear unit (ReLU) activation function is applied to all LSTM layers, allowing the network to learn complex patterns and nonlinear relationships within the data.

After the LSTM layers, three dense layers are added to the model. These layers consist of fully connected neurons and contribute to the final classification process. The first dense layer contains 64 units, followed by a layer with 32 units, and the The final layer contains as many units as there are distinct sign language movements in dataset. Activation function employed is ReLU to these dense layers to introduce non-linearity into network.

IV. RESULTS

Model is compiled and tested in real time, for training the model optimizer employed is Adam with a rate of learning equal to 0.0001 and categorical accuracy as well as validation accuracy is calculated. Model trained for 100 epoch, gives an training accuracy equal 95 percent, validation accuracy equal 92 percent respectively with a loss equal 0.1229 percent. Fig. 4. shows comparison between training, validation accuracy, Fig. 5. shows comparison between training loss, validation loss

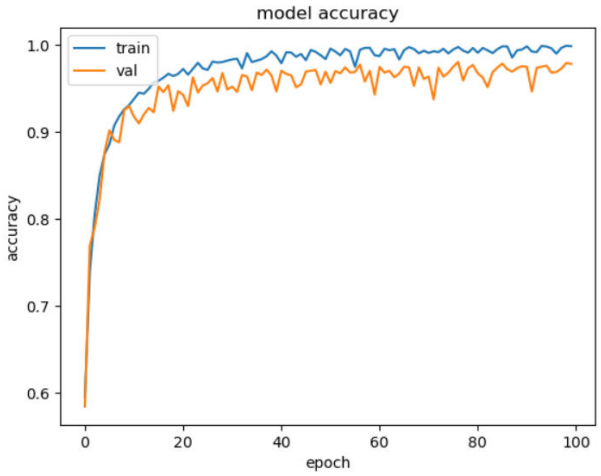

Fig. 4. Comparison between training and validation accuracy

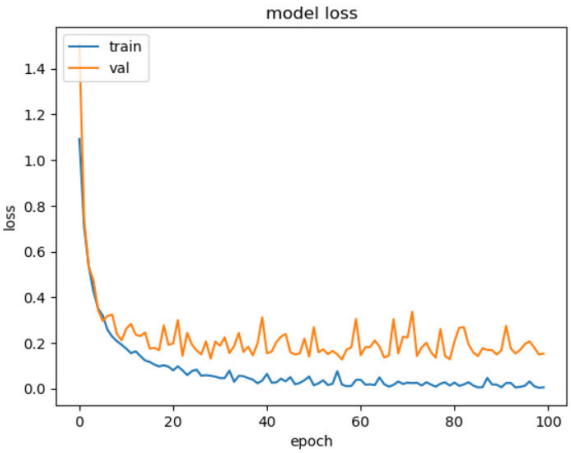

Fig. 5. Comparison between training and validation loss

Fig. 6 shows real time prediction of sign language gesture

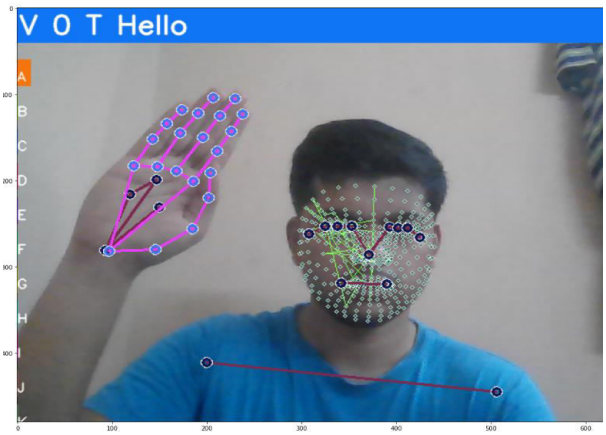

Fig. 6. Real time detection of sign language

V. CONCLUSION

This paper presented a method for decoding sign language detection models to aid individuals who are deaf or mute and cannot communicate effectively with those unfamiliar with sign language. Model is built using American Sign Language (ASL) dataset and is capable of converting sign language gestures into easily understandable text. With this solution, individuals who were previously unable to communicate effectively with others can now do so with greater ease and understanding, enhancing their ability to connect and communicate with the world around them. Model performs well with a training accuracy equal 95 percent, validation accuracy equal 92 percent respectively with a loss equal 0.1229 percent. It is able to detect and predict hand gestures even in low light conditions which makes it extremely resilient Further the system can be used to create interactive video lessons, which can help learners to practice sign language in real-time, virtual sign language interpreter for public events or conferences.

REFERENCES

[1] Online Resource : Orfiled Laboratories, https://www.orfieldlabs.com/news (2012)

[2] Online Resource : World Health Organization(WHO) Deafness and Hearing Loss, https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss (2023)

[3] Sanket Dhobale, Aniket Kandrikar, Sumeet Manapure, Aniket Zullarwar, Ankush Temburnikar, "Real Time Sign Language Detector Using Deep Learning", in International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 1, 2022.

[4] Pallavi Gurjal, Kiran Kunnur, "REAL TIME HAND GESTURE RECOGNITION USING SIFT", in International Journal of Electronics and Electrical Engineering, vol. 2, 2012.

[5] Aman Pathak, AvinashKumar, Priyam, Priyanshu Gupta, Gunjan Chugh, "Real Time Sign Language Detection" in International Journal for Modern Trends in Science and Technology, pp. 32-37, 2022.

[6] Amrita Thakur, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha, Subarna Shakya, "Real Time Sign Language Recognition and Speech Generation", in Journal of Innovative Image Processing (JIIP), vol. 02, no. 02, pp. 65-76, 2020.

[7] Sanket Bankar, Tushar Kadam, Vedant Korhale, A. A. Kulkarni, "Real Time Sign Language Recognition Using Deep Learning, in International Research Journal of Engineering and Technology (IRJET), vol. 09, 2022.

[8] H. Y. Lai, H. J. Lai, "Real-Time Dynamic Hand Gesture Recognition," in International Symposium on Computer, Consumer and Control, pp. 658-661, 2014.

[9] J. J. Raval, R. Gajjar, "Real-time Sign Language Recognition using Computer Vision," International Conference on Signal Processing and Communication (ICPSC), pp. 542-546, 2021.

[10] Ketan Gomase, Akshata Dhanawade, Prasad Gurav, Sandesh Lokare, "Sign Language Recognition using Mediapipe", in International Research Journal of Engineering and Technology (IRJET), vol. 09, 2022.

[11] Vivek Bheda, Dianna Radpour, "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language", 2017.

[12] Sakshi Goyal, Ishita Sharma, Shanu Sharma, "Sign Language Recognition System For Deaf And Dumb People", in International Journal of Engineering Research & Technology (IJERT), vol. 2, 2013.

[13] Rung-Huei Liang , Ming Ouhyoung, "A real-time continuous gesture recognition system for sign language," Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 558-567, 1998.

[14] T. Starner, J. Weaver, A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 12, pp. 1371-1375, 1998.

[15] D. Uebersax, J. Gall, M. Van den Bergh, L. Van Gool, "Real-time sign language letter and word recognition from depth data," in IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 383-390, 2011.

[16] Rajam, P. S., Balakrishnan, G., “Real time Indian Sign Language Recognition System to aid deaf-dumb people” in IEEE International Conference on Communication Technology, 2011.

[17] K. Imagawa, Shan Lu, S. Igi, "Color-based hands tracking system for sign language recognition," Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 462-467, 1998.

[18] Saquib, N., Rahman, A., " Application of machine learning techniques for real-time sign language detection using wearable sensors", Proceedings of the 11th ACM Multimedia Systems Conference, 2020.

[19] Liang, R.-H., Ouhyoung, M., "A sign language recognition system using hidden markov model and context sensitive search", Proceedings of the ACM Symposium on Virtual Reality Software and Technology, 1996.

[20] Raheja, J. L., Mishra, A., Chaudhary, A., " Indian sign language recognition using SVM", Pattern Recognition and Image Analysis, vol. 26, no. 2, pp. 434–441, 2016.

[21] B. Bauer, H. Hienz, "Relevant features for video-based continuous sign language recognition," Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 440-445, 2000.

[22] Jie Huang, Wengang Zhou, Houqiang Li, Weiping Li, "Sign Language Recognition using 3D convolutional neural networks," in IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6, 2015.