Real-Time Sign Language Interpreter using Mediapipe, Dynamic Time Warping and NLP

Yash Mankar[1], Vedanshu Thune[1], Shrikant Khawshe[1], Yogvid Wankhede[1], Madhuri Sahu[1]

[1]Department of Artificial Intelligence G.H. Raisoni College of Engineering, Nagpur, 440025, India

Abstract-

The "Sign Language Interpreter" project is dedicated to alleviating communication challenges encountered by individuals who're unable to express themselves through means. This innovative study utilizes hand gestures to generate sentences in time thereby facilitating effective communication. Inspired by the language of signs our system employs algorithms for gesture recognition. It. Translates the nuances of hand movements into sentences ensuring accuracy and contextual understanding. The Sign Language Interpreter incorporates cutting edge technologies to comprehend sign language intricacies.

Index Terms-

Indian Sign-Language, Hand Gesture Recognition, Mediapipe, Dynamic Time Warping, Landmark Extraction, NLP

I. INTRODUCTION

The proposed study urges to implement a model that detects the hand gesture and in order to interpret those conversations into written and spoken language. A real-time sign language interpreter using Mediapipe and Dynamic Time Warping is being developed. Sign language is a highly expressive and intricate form of communication that combines fluid hand movements, facial expressions, head movements, and body language. It is the seamless integration of these elements that makes it a complete and effective means of communication. However, this presents a formidable task in the field of computer science, primarily due to the complexity of data extraction and subsequent analysis. These challenges arise from several factors. To begin with, the human body is inherently non-rigid, with numerous degrees of freedom. Additionally, it has the capacity to produce an infinite array of variations for even the most basic movements. Furthermore, each individual possesses their unique body shape, volume, and distinctive gesture style, thereby complicating the recognition process. Moreover, the presence of uncertainties such as variations in viewpoint, illumination, shadows, self-occlusion, deformation, noise, and clothing further intensify the complexity of this problem.

Our primary goal is to provide an accessible solution for individuals with speech impairments enabling effortless communication with an audience. In addition to converting sign language into spoken words, the proposed sign language interpreter offers features such as an affordable and portable user interface making it suitable for sign recognition and converting those to spoken words as well as phrases. The interpreter is designed to be user friendly so that individuals with levels of proficiency can communicate seamlessly using sign language. This research paper outlines the development process, underlying technologies used in creating the Sign Language Interpreter, as its evaluation. Our goal for this project is to make a contribution to technology promoting better communication and inclusiveness, for people who have difficulty speaking by combining technical advancements such as Mediapipe, Dynamic Time Warping and Natural Language Processing.

Communication is a basic aspect of human interaction, yet individualities with speech disabilities frequently encounter significant challenges in expressing themselves conventionally. The proposed "Sign Language Interpreter" Mediapipe and Dynamic Time Warping emerges as a transformative result aimed at prostrating these obstacles and fostering inclusive communication. By employing the suggestive power of Indian subscribe Language (ISL), our design not only addresses the nuances of hand gestures but also strives to ground the communication gap for those who are dependent on the Indian Sign Language as their primary mode of expression.

The World Health Organization (WHO) states that around 1.5 billion people, which is approximately 20% of the global population, live with hearing loss, highlighting the need for accessible communication. While Natural Language Processing (NLP) has made significant advancements in simplifying language-related tasks, the progress in technology supporting sign language has been less substantial. Automatic sign language translation and generation systems offer efficient and accessible communication channels between deaf individuals and the hearing community. In recent years, there has been a growing interest in sign language technologies, with researchers exploring computer vision and deep learning approaches to address this complex task. Although many of these studies utilize text or gloss representations for sign generation, the field of speech-to-sign language generation remains relatively unexplored. Gloss, commonly used to represent sign language, has been found to lack accuracy in capturing the full linguistic and expressive aspects of sign language.

At the center of our action lies the application of a comprehensive Indian Sign Language database for the training of our model. This expansive database serves as the bedrock for developing our system. The addition of Indian subscribe Language ensures that our model is culturally sensitive, understanding the unique verbal nuances and expressions bedded in the rich shade of ISL.

The variance in the performance of human actions over time results in unique feature representations for each action in different samples. This poses a challenge for widely-used classifiers like neural networks and support vector machines, which typically expect fixed-size input feature vectors and struggle with action recognition tasks.

To address this issue, our research introduces a classification model that leverages Dynamic Time Warping (DTW) to assess the similarity between two action samples. We also incorporate a voting algorithm to match a test action to a set of training action samples.

Upon recognition of the sign language input, natural language processing ways come into play to convert the text to speech for the asked spoken language. The affair is aloud presented through a speaker, providing effective communication. Also, the system can display the converted text on the screen for easy reference, further enhancing its effectiveness and inclusivity. The admixture of Indian subscribe Language, sophisticated tackle, and slice- edge software makes the" Sign Language Interpreter" design a groundbreaking bid in the realm of assistive technology.

II. EXISTING MODELS

These existing systems aim to enable communication between individualities who use sign languages and those who do not. One of the main challenges in developing sign language recognition systems is the lack of sufficient data for training machine literacy models. To address this issue, experimenters have proposed colorful styles for collecting and recycling sign language data, similar as using depth cameras, marker- grounded systems, and wearable detectors. Another challenge is the high variability in sign language gestures, which makes it delicate to train models to recognize specific signs with high delicacy. To overcome this, experimenters have proposed colorful ways similar to using deep literacy models, transfer literacy, and sphere adaption. In terms of speech conversion, some studies have used Text- to- Speech (TTS) technology to convert the sign language textbook into speech. TTS technology has been extensively used in sign language recognition systems to give audio feedback to the druggies. Still, TTS technology is still facing challenges similar to the lack of suggestive voice and limited support for different languages. In terms of using a camera module, screen, and speaker to make a sign language recognition and speech conversion system, there are many studies that have proposed such a system. These studies have concentrated on the perpetration of the system, including the tackle and software conditions, and have reported promising results in terms of recognition and conversion delicacy. In summary, while there's a growing body of literature on sign language recognition and speech conversion, there's still a need for further exploration in this area, particularly in terms of collecting and recycling sign language data, developing more accurate recognition models, and perfecting the quality of speech conversion.

Comparison of crucial features from former exploration:

In Sign Language Recognition, there have been numerous former exploration papers that have proposed different styles and ways to improve the delicacy of the recognition system. A comparison of crucial features from former exploration and a specific paper in Sign Language Recognition would involve looking at factors similar as the dataset used, the AI model armature, the input and affair, the evaluation criteria, and the results achieved. For illustration, a comparison of crucial features between former exploration and a specific paper in Sign Language Recognition could include:

Dataset prior exploration may have used a different dataset than the specific paper, which can affect the results achieved.

Input and output former exploration may have used different input and affair formats than the specific paper, which can affect the performance of the recognition system.

Evaluation criteria prior exploration may have used different evaluation criteria than the specific paper, which can affect the interpretation of the results.

Results achieved by former exploration may have achieved different results than the specific paper, which can affect the overall conclusion of the exploration.

In conclusion, a comparison of crucial features between former exploration and a specific paper in subscribe Language Recognition would involve looking at factors similar as the dataset used, the input and affair, the evaluation criteria, and the results achieved. Understanding these parallels and differences.

Former experimenters used different styles similar as:

Figure discovery approach Contour discovery is a fashion that can be used to identify and prize the boundaries of objects or regions of interest in an image or videotape. In the environment of sign language recognition, figure discovery can be used to describe the boundaries of the hand regions in the videotape.

RNN combined with another AI approach intermittent Neural Networks (RNNs) are a type of neural network that can be used to reuse successional data similar as videos, audio, or textbook. They're particularly well- suited to tasks similar to sign language recognition, where the input data is a sequence of hand gestures. RNNs can be used to model the temporal dependencies between the gestures, which is an important aspect of sign language recognition. An RNN can be combined with another AI approach, similar as a convolutional neural network (CNN) or a support vector machine (SVM), to ameliorate the performance of the sign language recognition system. One possible way to do this is to use a CNN to prize features from each frame of the videotape, and also use an RNN to model the temporal dependencies between the frames.

HMM (Hidden Markov Model) approach A retired Markov Model (HMM) is a type of probabilistic model that can be used for tasks similar as sequence bracket, sequence vaticination, and pattern recognition. In the environment of sign language recognition, HMMs can be used to model the temporal dependencies between the hand gestures in a sign language videotape.

Glove approach A glove approach in sign language recognition refers to the use of gloves equipped with detectors to describe hand gestures and movements. These gloves can be used to capture detailed information about the hand's position, exposure, and movement, which can also be used to recognize specific hand gestures in sign language.

Other approaches: There are numerous other approaches that can be used for sign language recognition, some of which include stir prisoner This approach uses cameras and detectors to track the movement of the signer's body and hands, and also uses this information to recognize specific hand gestures.

Computer vision: This approach uses ways similar as object discovery, point birth, and machine literacy to dissect videotape footage of the signer's hands and feet specific hand gestures. mongrel approaches This approach combines two or further of the below- mentioned approaches, similar as a combination of a glove with computer vision or a combination of stir prisoner with machine learning.

Deep Learning grounded approaches: This approach uses deep neural networks similar as CNN, RNN, LSTM, and other infrastructures to prize features from the videotape and recognize the signs.

Transfer Learning: This approach takes a pre-trained model on an affiliated task and fine tuning it on the sign language recognition task.

Multi-modal approaches: This approach uses multiple modalities similar to audio, video, and textbook to recognize the sign language.

It's worth noting that these are just many exemplifications and the specific approach will depend on the specific task and the data being used. Also, it's important to estimate different approaches to find the bone that performs stylish for the specific task and the data. Also, it's important to consider the real- time performance, delicacy and the computational power of the system while choosing the applicable approach.

III. PROPOSED METHODOLOGY

Here are some general steps that could be taken to develop a sign language recognition and speech conversion system using our methodology which involves implementing a real-time sign language interpreter using Dynamic Time Warping and Mediapipe.The Dynamic Time Warping Classification model is represented in Figure-1. The aim is to generate the similar word on the basis of a sign recorded in a video. Our Sign Language Recognition model uses hand landmarks (points of interest of the hand) that we track- as input. The prediction for our model is done in the following ways:

- Extract Landmarks
- Calculate the DTW distance between the recorded and reference signs.
- Predict the sign/gesture by analyzing which reference sign is likely to be recorded.

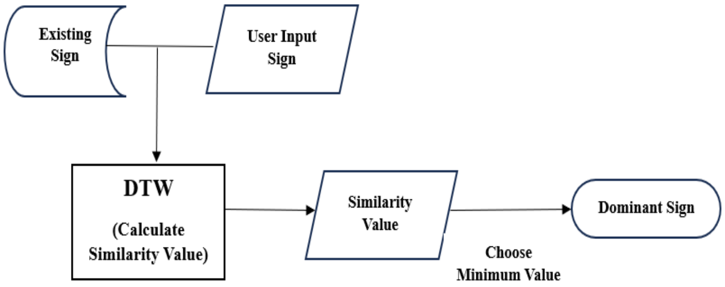


Figure-1: Dynamic Time Warping Classifier

The proposed Dynamic Time Warping (DTW) classification stands as a pivotal method within the realm of time series analysis and pattern recognition. Its purpose is to classify and contrast time series data effectively. In the realm of DTW classification, the crux lies in measuring the similarity between two time series. This is achieved by determining the optimal alignment or "warping" between the series while accommodating variations in time and shape. The versatility of DTW classification finds application in a multitude of domains, including speech recognition, gesture identification, and the categorization of temporal data. Its adaptability for accommodating time series variances positions it as an indispensable asset within the realms of machine learning and data analysis. Here, we are comparing the existing sign and the user input sign for classification and upon applying the Dynamic Time Warping Classifier, we obtain the Similarity value and based upon that value the minimum value a dominant sign would be recognized.

For this study, we are considering our own custom dataset for training our model. The contents of our model are given in Table-1. These are some of the basic sign language gestures which are widely used for communication. It also mentions the number of training samples for them respectively.

Index	Action Type	Training Sample set
1	How	5
2	You	5
3	I	5
4	Good	5
5	What	5
6	Time	5
7	Where	5
8	Washroom	5
9	Please	5
10	Help	5
	Total	**50**

Table-1: Training Dataset

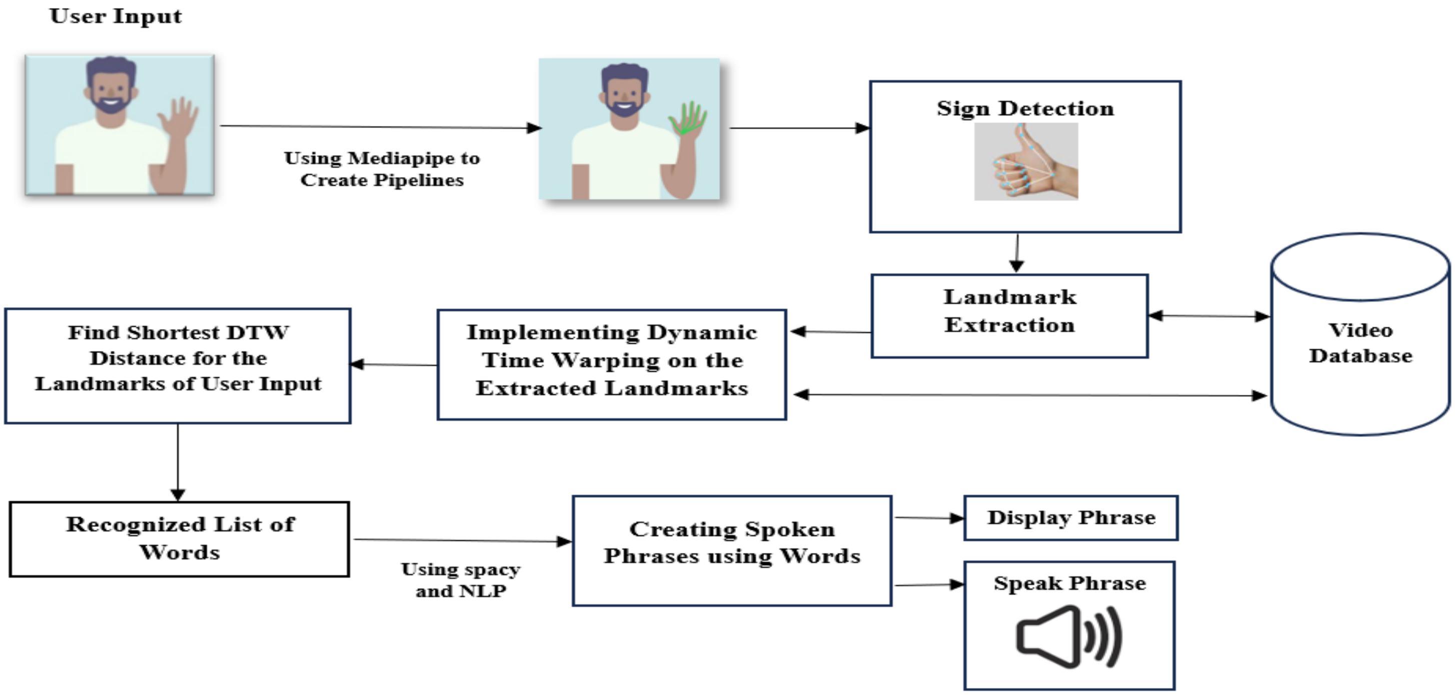

User Input
Using Mediapipe to Create Pipelines
Sign Detection
Landmark Extraction
Video Database
Implementing Dynamic Time Warping on the Extracted Landmarks
Find Shortest DTW Distance for the Landmarks of User Input
Recognized List of Words
Using spacy and NLP
Creating Spoken Phrases using Words
Display Phrase
Speak Phrase

Figure-2: Model Architecture

The flow of our suggested model using Dynamic Time Warping is represented in Figure-2. The implementation of our model is given by:

1. MediaPipe Detection/Holistic Model

MediaPipe is an open- source framework for computer vision results released by Google a couple of years ago. Among various models provided by MediaPipe, the Holistic Model can track in real- time the position of the Hands, the Pose and the Face landmarks. For now, our model uses Hands positions as well as the Pose landmarks (position of shoulder, elbow and wrist) to make the prediction.

2. Extract landmarks

First, let's draw landmarks from the video with the help of the Holistic model's process method. Colour conversion has to be done because OpenCV uses BGR colours and MediaPipe RGB ones.

3. Draw landmarks

The drawing_utils sub-package of MediaPipe conveniently provides all the essential tools required for rendering landmarks on a single image.

4. Models

The project's current interpretation involves the execution of two distinct models.

The "HandModel" class encapsulates detailed information about hand gestures within a single image. Meanwhile, the "SignModel" class encompasses and extends the "HandModel" information to cover all frames within a video.

5. Hand Model

I. The main challenge when using landmarks positions as input data, is that the prediction is acute (sensitive to the size and the fixed position of the hands).

II. A logical way to extract the information about the hand gesture is to use the angles between all the parts of the hand, called joints. We'll use all 21 connections of MediaPipe's Hand Model in this program.

III. The HandModel class applied in this program is defined by its feature_vector that gives a representation of the hand gesture.

IV. The HandModel class has two arguments:

- connections: List of tuples containing the ids of the two landmarks representing a connection
- feature_vector: List of the 21 * 21 = 441 angles between all the joints of the hand

6. Sign Model

I. Now that we can extract the information of the hand gesture, we've to construct an object containing both spatial and temporal information of a sign.

II. To accomplish this task, it's simply a matter of retaining the feature vectors for both hands across every frame in the video.

III. The SignModel class applied in this program has four attributes:

- The variables has_left_hand and has_right_hand are set to "True" if the video includes the presence of the left and right hands, respectively.
- The variables lh_embedding and rh_embedding store lists of feature vectors corresponding to the left and right hands for every individual frame.

7. Sign Prediction

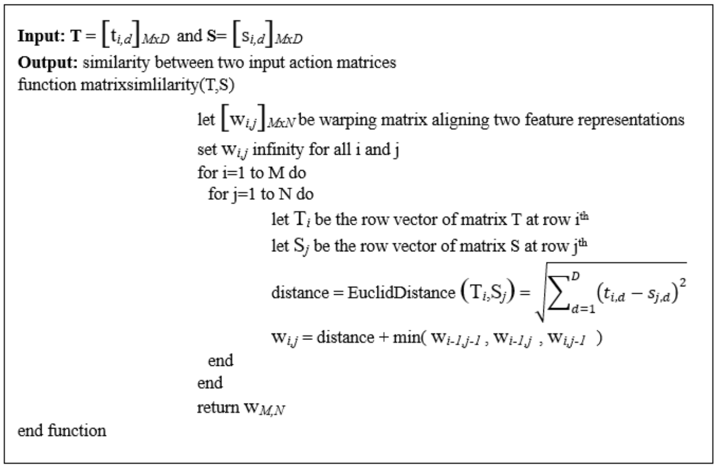
So, we've now created sign embeddings that contain the temporal and spatial information of a sign. The coming step is to distinguish the signs precisely. At present, there exist multiple methods that can be employed to address the same issue. But the approach used by us requires far less training data and is suitable to calculate a similarity between two signs. With Dynamic Time Warping, we can calculate the distance between embeddings as they're time series of feature_vectors.

8. Dynamic Time Warping (DTW)

Dynamic Time Warping is an algorithm broadly utilized for time series comparison.

It identifies the most optimal alignments between two time series by applying a warping technique. This allows us to analyze patterns rather than sequences. In our case, DTW will find resemblances between embeddings of the same signs even if they're done at different speeds. To calculate the similarity between two signs we compare their embeddings. We calculate the DTW distance between the embeddings of the sign recorded and the embeddings of all the reference videos (each sign has multiple reference videos).

The following approach returns all the signs in the list, sorted by their distance to the recorded sign.The calculation part of Dynamic Time Warping is explained in Figure-3.



```
Input: T = $[t_{i,d}]_{MxD}$ and S= $[s_{i,d}]_{MxD}$
Output: similarity between two input action matrices
function matrixsimlilarity(T,S)
    let $[w_{i,j}]_{MxN}$ be warping matrix aligning two feature representations
    set $w_{i,j}$ infinity for all i and j
    for i=1 to M do
      for j=1 to N do
            let $T_i$ be the row vector of matrix T at row i$^{th}$
            let $S_j$ be the row vector of matrix S at row j$^{th}$
            distance = EuclidDistance $(T_i,S_j) = \sqrt{\sum_{d=1}^{D}(t_{i,d} - s_{j,d})^2}$
            $w_{i,j}$ = distance + min( $w_{i-1,j-1}$ , $w_{i-1,j}$ , $w_{i,j-1}$ )
      end
    end
    return $w_{M,N}$
end function
```

Figure-3: Using Dynamic Time Warping to Calculate Similarity

9. Sign prediction

We have calculated the distances between the recorded sign and all the reference ones. So, by sorting them we can take a batch of the most resembling signs to our record. Using this batch, we can assess whether a sign repetition occurs frequently enough to instill confidence in our prediction.

In our study, we choose batch_size = 2 and threshold = 0.8, meaning that if the same sign appears at least 2 times in the batch then we generate output as a label of sign. Else, we generate output as "Unknown sign". The choice of batch_size and the setting of threshold values are contingent on the quantity of videos available for each sign within the dataset.

10. Phrase Creation

Based upon the keywords obtained from our sign language recognition model, we would pass a list of these words to form meaningful sentences by using the spacy library.

IV. IMPLEMENTATION

We have used the combination of Mediapipe, Dynamic Time Warping (DTW) and Natural Language Processing (NLP). The implementation of our model is described in Figure-4 for certain sentences. Also, we have described several words and greetings in Figure-5.

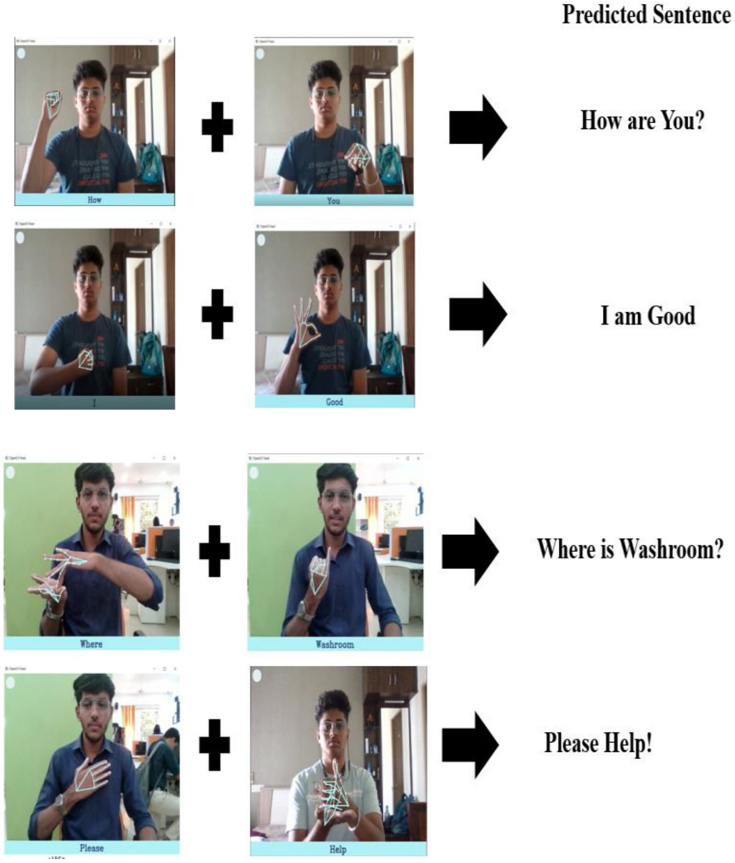


Figure-4: Sign to Sentence Interpretation using Dynamic Time Warping

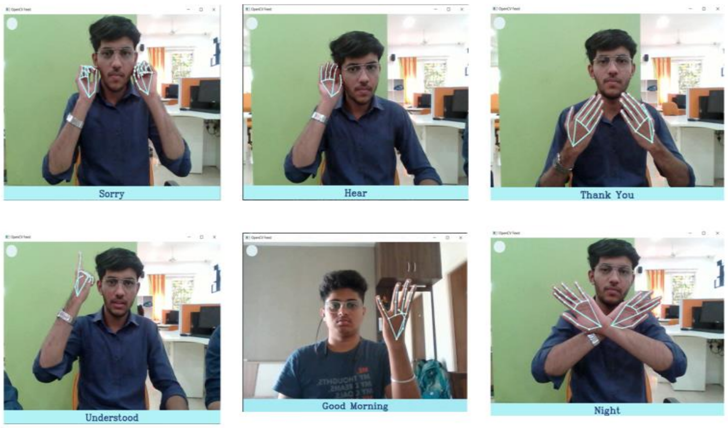


Figure-5: Greetings and Common words Recognition

V. RESULT

In this section, we evaluate the recognition performance of our method in terms of both accuracy and complexity. The accuracy and complexity of both the approaches used for recognition were compared. The first method was LSTM for recognition and the other one was DTW.Model in which recognition was done through LSTM requiring a huge amount of data for each sign to work efficiently. To train a LSTM model, we have to run iterative epochs, and the time required to complete these epochs depends on the size of the dataset and also on the computational power of the machine. Thus, LSTM models have worse time and space complexity. Whereas, the model in which DTW was used for recognition, required less data as it compared similarity between two signs. Due to which model was more accurate and had less time and space complexity. Thus, based on our study and experimentation DTW is recommended for sign recognition. On the basis of 3 Test Cases Conducted, accuracy was obtained as mentioned in Table-2.

Test Case	LSTM Accuracy	DTW Accuracy
Test 1	80%	92%
Test 2	83%	95%
Test 3	82%	93%
Average	81.6%	93.3%

Table-2: Model Accuracy Comparison

VI. CONCLUSION

This research paper outlines a successful and comprehensive approach to sign language recognition using the MediaPipe and Dynamic Time Warping. By leveraging the model's capabilities to track hand positions, pose landmarks, and facial landmarks in real-time, the study has enabled the prediction of sign language gestures. It also provides an effective and efficient approach to sign language recognition, making it a valuable contribution to the field of computer vision and gesture recognition. The use of Dynamic Time Warping for sign comparison allows for robust recognition even under variations in sign execution speed. Additionally, the approach's reliance on feature vectors and time series analysis minimizes the need for extensive training data, making it a practical solution for real-world applications in sign language recognition.The proposed model could successfully recognize and convert the input sign language to a spoken language enabling effective communication.

VII. REFERENCES

1. Chinese Sign Language Recognition Based on DTW-Distance-Mapping Features Juan Cheng, 1Fulin Wei, 2Yu Liu, 1Chang Li, 1Qiang Chen, 1and Xun ChenAcademic Editor: Wanquan LiuCopyright © 2020 Juan Cheng et al.

2. Sign Language Recognition Using Keypoints Through DTW, Published in: 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP), DOI: 10.1109/CCIP57447.2022.10058664

3. Robust Identification System for Spanish Sign Language Based on Three-Dimensional Frame Information Jesús Galván-Ruiz 1, Carlos M. Travieso-González 1,2, Alejandro Pinan-Roescher 1 and Jesús B. Alonso-Hernández 1,

4. MultiFacet: A Multi-Tasking Framework for Speech-to-Sign Language Generation Mounika Kanakanti, Shantanu Singh, Manish Shrivastava, International Institute of Information Technology Technology Hyderabad, India

5. Ham2Pose: Animating Sign Language Notation into Pose Sequences Rotem Shalev Arkushin - Reichman University, Amit Moryossef - Bar-Ilan University, Ohad Fried - Reichman University

6. Adaptive hough transform with optimized deep learning followed by dynamic time warping for hand gesture recognition Manisha Kowdiki1 · Arti Khaparde1 Received: 20 February 2021 / Revised: 21 July 2021 / Accepted: 19 August 2021 © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

7. Data Glove with Bending Sensor and Inertial Sensor Based on Weighted DTW Fusion for Sign Language Recognition Chenghong Lu, Shingo Amino and Lei Jing * Graduate School of Computer Science and Engineering, University of Aizu, Tsuruga, Ikki-machi, Aizuwakamatsu City 965-8580, Japan * Correspondence: leijing@u-aizu.ac.jp

8. Intelligent Sign Language Recognition System for E-Learning Context Muhammad Jamil Hussain1, Ahmad Shaoor1, Suliman A. Alsuhibany2, Yazeed Yasin Ghadi3, Tamara al Shloul4, Ahmad Jalal1 and Jeongmin Park5, *Computers, Materials & Continua Tech Science Press DOI: 10.32604/cmc.2022.025953

9. Human Action Recognition Using Dynamic Time Warping and Voting Algorithm January 2014 © 2014 Published by VNU Journal of Science. Manuscript communication: received 10 December 2013, revised 04 March 2014, accepted 26 March 2014 Corresponding author: Pham Chinh Huu

10. J. Wang, H. Zheng, “View-robust action recognition based on temporal self-similarities and dynamic time warping,” IEEE Int. Conf. on Computer Science and Automation Engineering (CSAE), 2012, pp. 498-502.

11. W. Li, Z. Zhang and Z. Liu, “Action recognition based on a bag of 3D points,” IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, 2010, pp.9 -14.

12. . Zhang, L. Long, D. Shi, H. He, and X. Liu, "Research and Improvement of Chinese Sign Language Detection Algorithm Based on YOLOv5s” in Proceedings of the 2022

2nd International Conference on Networking, Communications and Information Technology (NetCIT)

13. A. Deep, A. Litoriya, A. Ingole, V. Asare, S. M. Bhole, and S. Pathak, "Realtime Sign Language Detection and Recognition” in Proceedings of the 2022 2nd Asian Conference on Innovation in Technology (ASIANCON)

14. D. Gandhi, K. Shah, and M. Chandane, "Dynamic Sign Language Recognition and Emotion Detection using MediaPipe and Deep Learning" in Proceedings of the 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)

15. G. Anilkumar, M. S. Fouzia, and G. S. Anisha, "Imperative Methodology to Detect the Palm signs (American Sign Language) using YOLOv5 and MediaPipe" in Proceedings of the 2022 2nd International Conference on Intelligent Technologies (CONIT)

16. A. D. Shetty, J. Shetty, K. K, R. Rakshitha, and S. S. B, "Real-Time Translation of Sign Language for Speech Impaired" in Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication (ICCMC)

17. Sakshi Sharma and Sukhvinder Singh “Recognition of Indian Sign Language (ISL) Using Deep Learning Model” Wireless Personal Communications, An International Journal -Springer

18. N. Rajasekhar, M. Geetha Yadav, Charitha Vedantam, Karthik Pellakuru, Chaitanya Navapete, "Sign Language Recognition using Machine Learning Algorithm” in International Conference on Sustainable Computing and Smart Systems (ICSCSS)

19. B. V. Chowdary, Ajay Purshotam Thota, Alwa Sreeja, Kotla Nithin Reddy, and Karanam Sai Chandana, "Sign Language Detection and Recognition using CNN" in International Conference on Sustainable Computing and Smart Systems (ICSCSS)

20. Shagun Katoch, Varsha Singh, and Uma Shanker Tiwary, "Indian Sign Language recognition system using SURF with SVM and CNN" in International Journal of Trend in Scientific Research and Development (IJTSRD)

21. Atharv Ganpatye and Sunil Mane, "Motion Based Indian Sign Language Recognition using Deep Learning" in 2nd International Conference on Intelligent Technologies (CONIT)