International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Real-Time Translation of Dynamic Hand Gestures Sequences into Synthesized Speech Using Spatiotemporal Deep Neural Networks

Sandeep Kumar Arudra[1], P.NagaVenkata Dileep[2], K.Rakesh[3], R.Venkata Sai Lokesh[4], P.Irfan Khan[5]

[1]Associate Professor, Department of CSE – Data Science, KKR & KSR Institute of Technology and Sciences, Guntur.
sandeepkumararudra@gmail.com[1]

[2,3,4,5] B-tech Student, Department Of CSE – Data Science, KKR & KSR Institute of Technology and Sciences, Guntur,
22jr1a44a6@gmail.com[2], 22jr1a44b5genai@gmail.com[3], sailokeshrachuri@gmail.com[4], irfankhanx1126@gmail.com[5].

ABSTRACT

People who communicate using hand gestures, especially those who are deaf or unable to speak, face difficulties in talking with others. Real-time translation of hand gestures, which can then be converted to speeches, may overcome communication difficulties between people using hand gestures and normal people. This project presents a Real-time translation of Dynamic hand gestures into synthesized speech using the spatiotemporal deep neural networks, The system that captures the video/Photo of continuous hand gestures that convert into the frames. The CNN to extract the visuals from frames and the LSTM to understand the movement of the gestures. Then the gestures are converted into Text and immediately transformed into real time Speech. The result shows that produces the clear gestures of spatiotemporal deep neural networks converting the Text into the real time speech, The project that can be useful for the individuals with speech and hearing disabilities and human–computer interaction and the real time communication systems using the spatiotemporal deep neural networks

Keywords: Gesture-to-Speech Translation, Computer Vision, CNN - LSTM, Deep Neural Networks, Sign language Translation, Text to Speech, Real-Time Processing

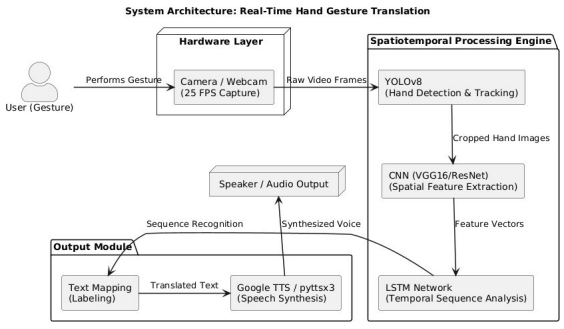
INTRODUCTION

Hand gestures are widely used for the people who are unable to speak like a normal people, human-computer interaction plays an important role in the hand gestures in the real time communication by using the devices like cameras and sensors to detect the faces and movement of the live hand gestures and gives the output as translation of text to speech by using the spatiotemporal deep neural networks. The result which shows the real time hand gestures of their hands and position of the gestures make them more accurate and train the models of the spatiotemporal features to make it more effective.

This project focuses on real time translation of dynamic hand gestures into synthesized speech using spatiotemporal neural networks. The system that captures the hand gestures in real time as an input, and processes the gestures in deep neural networks models and designs the spatiotemporal features,and converts it into corresponding text to speech. The improvement in deep learning to understand the spatiotemporal neural networks, it helps to understand the shape and position of the hand gestures and how it moves over time. By this motion and position of the hand gestures in real time and making the dynamic hand gestures more accurately from the video that was captured.

It is fail to recognize the hand gestures correctly because of the speed, camera angle, lighting and signing styles by improving the deep neural network models like Convolution neural networks (CNN) where it extracts the information and converts the video into frames, such as hand shapes and the position in real time. Where as the Long short term memory (LSTM) it understand the movement of the hand gestures and learn the patterns from the features and also using the deep neural networks converts into speech and becomes the gestures more accurate and reliable

With this system make easily to convert the sign languages into the spoken language and make the communication more easy and accurate to understand the gestures who are not able to speak or hearing disabilities by using the networks like CNN, LSTM and deep neural networks works in the real time that captures the video of the hand gestures and translate the speech and helps to convert the sign language into the spoken language.



LITERATURE REVIEW

Earlier, many researchers have worked on static hand gestures by using image processing techniques and it only identifies hand shapes but did not identify hand movements/motions these are very important for the dynamic hand gesture language. Years later, systems use wearable devices like data gloves and sensors to make it

easier to understand the recognition of the gestures but they are uncomfortable for daily use. After that they introduce the camera-based systems that capture the hand gesture features like hand shape, finger position, hand direction. These features are given to machine learning like support vector machines or decision trees but these systems need improvement in lighting condition, backgrounds with different users.

In the rise of deep learning, using CNN models has improved the performance by learning features from images. It is better than the traditional method and can handle the backgrounds and the lighting. CNN mainly focused on single frames and did not focus on gesture movements in the real time. The gesture based communication contains both the hand shapes, and hand movements. By using these two methods it developed the spatiotemporal neural networks.

The models which include the 3D CNNs and combination of the CNN and LSTM to understand the dynamic hand gestures and have focused on real-time hand gesture recognition. This system is difficult to develop, because it should process the video frames very fast and give the result very late. To solve this system researchers use the neural networks design to improve speed.

PROBLEM STATEMENT

People who rely on hand gestures for communication often face difficulties in interacting with others who do not understand sign language leading to communication barriers in daily life. Existing gesture recognition systems struggle to accurately interpret dynamic and continuous hand movements in real time and often suffer from limited accuracy, high latency, and dependence on specialized hardware. The challenge lies in effectively capturing both the special features of hand gestures and their temporal motion patterns to produce meaningful interpretations. This project aims to address these issues by developing the real time systems that translate dynamic hand gestures sequences into synthesized speech using spatiotemporal deep neural networks, thereby enabling effective and accessible communication between gesture-based users, and the general population.

METHODOLOGY

The Real-Time Translation of Dynamic Hand Gesture Sequences into Synthesized Speech systems work together in multiple steps. First it records the hand gestures using camera, video capture, hand detection, gesture recognition, text conversion, and finally speech recognition, these stages are very important to develop the real time dynamic hand gestures into synthesized speech

1. System Architecture

The system architecture which includes several parts. The first camera captures the hand gestures. The frames of the hand gestures check the hand detection modules, next the gesture recognition that mainly focus on hand movements and hand positions by using the neural network techniques. After that gesture recognition is converted into text andfinally it converted into text to speech with the useful for the individual of hearing disabilities and could not unable to speak which included in system architecture and these modules are easy to understand

2. Video Captures and Pre-Processing

The system which used to capture the videos by using the Webcam or Mobile camera. By using the camera it captures the video of hand gestures and detect the hand gesture and camera should maintain the Quality and clear hand visibility, the video that captures the frames and and maintain the high quality of the frames in camera, It captures the frames of the hand gestures and the frames which have the standard quality and that convert it into the pixels, and frames that are divided into pixels with the help of neural networks. The pre-processing techniques which use the deep neural network models that can reduce the background noise and focus on the hands and it can also handle the more complex background automatically.

3. Hand Detection and Tracking

The video that captures the hand gestures and detect the hand gestures by using the simple models to detect the hand gestures, the models like YOLO based hand detectors these models that detect the hands in real time and make more accurate After hand detection, the system introduce the hand tracking, the hand tracking which tracks the same hands in the hand-gestures it also helps to handle the situation when hand temporarily goes out of the frame and then the detect hand is made in frames and further process it includes the collection of data frames and make the frames into process the data which in a automatically done while processing the systems.

4. Spatiotemporal Feature Extraction

The hand gestures which have two main features are Spatial features and Temporal features. The spatial features which include the hand location, hand orientation, finger position, hand shapes and the Temporal features which include the hand movement, speed movement in real time and it captures both the features of spatial and temporal features by using the spatiotemporal neural networks. The system models which include the 3D CNNs, and the combination of the CNN and LSTM the CNN mainly focus on single frames and did not work on the gesture movements in over time and the LSTM that works on the gesture movements by using the spatiotemporal neural networks and understand the gesture movements

5. Classification of the Gestures

After the hand gestures the system is trained and sends the information to the classification layers. The classification layer which uses the neural networks that figure out the gestures hand, and make the hand gestures into possible gestures and uses something called softmax activation.and makes the highest priority and makes the correct gestures recognition and convert into text to speech. The system which has the datasets that are to be trained of the video hand gestures, during the training the neural networks of the datasets of the video hand gestures into the corresponding labels, and the system is trained and helps to work well with different camera angles, lighting conditions, clear visibility of the camera, background.

6. Gesture to Text Conversion

After the classification of the gestures is converted into the text for example it recognizes the gestures as “hello” and the system makes the output text as the “hello”. The system recognizes the multiple hand gestures of one another, after the hand gestures it converts into text and gives as the simple text and easily understand and make the sentence more natural. And the system directs the word to word translation and may not get perfect sentences. The system is trained and reduces the errors and gives the simple output of the word to sentence.

7. Text to Speech Synthesis

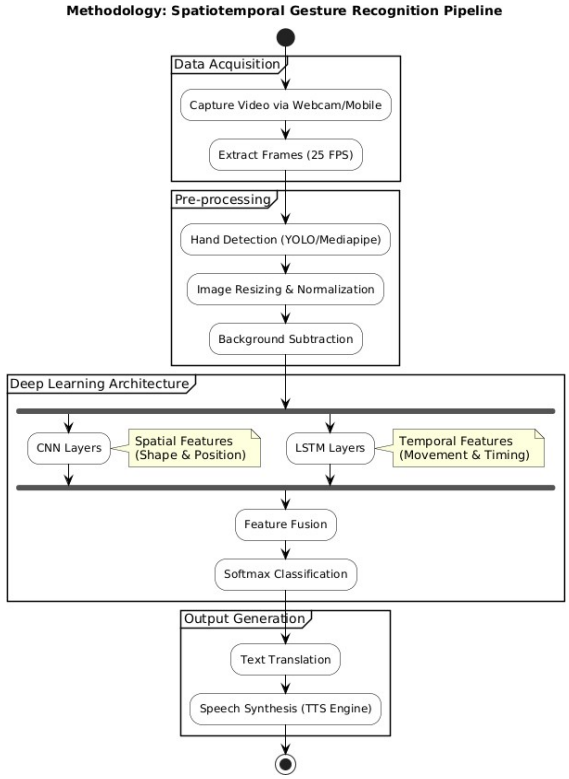
The final step that converting the text to speech, the system which uses the text to speech technologies like Google text to speech, pyttsx3 these techniques that are used to convert the text to speech synthesis, First it sends the text to the google text-speech converter and it makes the simple text to speech and plays through speakers, the speech should be clear and sound naturally, while using different voice options like male voice, female voice and select the different languages with their preferences and also the sound should also be adjusted by lower or faster or medium with makes the comfortable to the users.

8. Real time Processing Optimization

The system needs to process the frames in real time performances and also use the method of the several techniques to optimize the process in the real time frames. The deep neural networks which are to be processed and makes the techniques are optimized in real time processing data sets of the different frames, the neural network models that work in real time gestures into the speech by optimization techniques. With the GPUs it makes the performances and calculations in simultaneously, to achieve the real time performance it reduces the processing delay by the frame skipping and also uses a lightweight neural networks models in the real time processing optimization techniques, these techniques which gives the faster responses and accuracy in the real time hand gestures into speech synthesized by using the spatiotemporal neural networks.

9. Testing and Performances

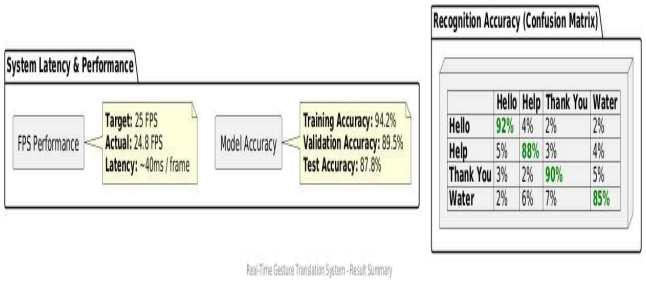
The different systems are to be tested and the performance of different hand gestures recognition in different conditions. The testing of the hand gestures in real time and checks the accuracy, speed quality, processing speed that should be maintained while checking the hand gestures in real time performances, it should maintain the 80% 90% accuracy for reliable use. To check the system works well the measurements are to used that are Recognition accuracy it shows the gestures are identified correctly or not, next the False positive rate which tells the system in wrong guess and next Processing latency means the system that takes to respond the hand gestures and finally, FPS Frames per second which makes the smoothly processed the video and the frames and helps to understand the systems in fast and accurate way.



RESULTS

The system processes the video in real time and smoothly gives the 25 frames per second. It correctly recognizes 85% of the hand gestures and accuracy gives the output as 90% . the difficult gestures that looks similar to each other, the results that changes depending on lighting, camera position, backgrounds by using the network models that are used to gesture recognition, the results that are showed the system can be useful in real world applications and further improvements and further uses.

The system that are able to recognize the hand gestures into the speech recognition and get the output as the simple languages, the gesture detection are worked well and makes the 95% of video frames at some times it is difficulty like lighting conditions or when hands move very fast detections are failed for a minute and quickly start working again. Finally the output is speech that are clear and easy to understand and maintain the voice sounds in lower, faster and medium and also uses the different voice options like male voice, female voice, and different language voices that are done by using the Google text-speech and make the performances good and accuracy of the dynamic hand gestures into synthesized speech by using the spatiotemporal neural networks.



DISCUSSION

Real time Translation of Dynamic Hand gesture Sequences into Synthesized Speech system using Spatiotemporal neural networks using deep neural network technology. It captures both the shape of the hands and hand movements over time and recognizes dynamic sign language gestures. The CNN and LSTM work together and make balanced accuracy and processing speed and the 3D CNN which gives the less accuracy and work well, and with handling of gestures variations with different users and different performances of the gestures like hand size, speed, movements these are to be handled and can be trained by collecting the datasets and by using better cameras to improve the better performances of gestures to speech

By using the simple models we detect the hand gestures and tracking the gestures in real time and tracking the frames of the videos and makes the frames into pixels, After the hand gestures it is trained and sends the information to the classification layers, it figure out the gestures hand and make the hand gestures into possible gestures that are called as softmax. It recognizes multiple hand gestures and convert it into simple text and easily understand and make to understand,with the use of Google text to speech converter it makes simple text to speech and plays through speakers and sounds it naturally and make the changes of different languages and also changes the sound to increase and decrease and better to understand and make it comfortable to use

CONCLUSION

The project is successfully developed the Real Time Translation of Dynamic Hand Gestures Sequences into Synthesized Speech System using Spatiotemporal neural networks, it converts the sign languages into synthesized speech and these are helps for the deaf and unable to speak persons with out need of human intervention. The system which captures the video of the hand gestures and makes use of spatiotemporal neural networks models that are trained the spatial and temporal network patterns and make use of CNN and LSTM work together and gives the output as the speech recognition accuracy and maintain the real-time processing speed.

By combining the computer vision and spatiotemporal neural networks and mainly focus on multi-language sign recognition, mobile and device developments and finally it support the grammar and sentence corrections, and which is trained and integrate the facial expression recognition, the artificial intelligences and computer vision focused and has the challenges in real world communication, The system that provides and can improve the quality of life for millions of deaf and unable to speak persons worldwide, by using these gestures it can easily done of translation of dynamic hand gestures into speech recognition

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] S. Sharma and C. Singh, "Sign language recognition using deep learning: A survey," in Proc. IEEE Int. Conf. Computing, Communication and Automation, 2017, pp. 1189–1194.

[3] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra, "A multimodal framework for sensor based sign language recognition," Neurocomputing, vol. 259, pp. 21–38, Oct. 2017.

[4] T. Starner, J. Weaver, and A. Pentland, “Real-time American sign language recognition using desk and wearable computer based video,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 12, pp. 1371–1375, Dec. 1998.

[5] H. Cooper, B. Holt, and R. Bowden, “Sign language recognition,” in Visual Analysis of Humans: Looking at People, T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, Eds. London, UK: Springer, 2011, pp. 539–562.

[6] N. Musthafa and C. G. Raji, ‘‘Real time Indian sign language recognition system,’’ Mater. Today, Proc., vol. 58,
pp. 504–508, Jan. 2022, doi: 10.1016/j.matpr.2022.03.011.

[7] T. Kapuscinski and K. Inglot, ‘‘Vision-based gesture modeling for signed expressions recognition,’’ Proc. Comput. Sci., vol. 207, pp. 1007–1016, Jan. 2022, doi: 10.1016/j.procs.2022.09.156.

[8] C. Y. Li, X. Zhang, and L. W. Jin, “LPSNet: A novel log path signature feature based hand gesture recognition framework,” in Proc. IEEE Int. Conf. Computer Vision Workshops, Venice, Italy, 2017, pp. 631–639.

[9] D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 3642–3649.

[10]C. J. Tsai, Y. W. Tsai, S. L. Hsu, and Y. C. Wu, “Synthetic training of deep CNN for 3D hand gesture identification,” in Proc. Int. Conf. Control, Artificial Intelligence, Robotics & Optimization, Prague, Czech Republic, 2017, pp. 165–170.

[11]W. J. Zhang and J. C. Wang, “Dynamic hand gesture recognition based on 3D convolutional neural network models,” in Proc. IEEE 16th Int. Conf. Networking, Sensing and Control, Banff, Canada, 2019, pp. 224–229.

[12]J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Boston, USA, 2015, pp. 4694–470

