iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42078

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Conversion of Sign Language Video to Text and Speech

Mr. G. Sekhar Reddy[1], A. Sahithi [2], P. Harsha Vardhan[3], P. Ushasri[4]
[1]*Assistant Professor, Dept of IT,* [1, 2, 3] *Bachelor of Technology, Information Technology, Anurag Group of Institutions, Telangana, India*

Abstract: ***Sign Language recognition (SLR) is a significant and promising technique to facilitate communication for hearing-impaired people. Here, we are dedicated to finding an efficient solution to the gesture recognition problem. This work develops a sign language (SL) recognition framework with deep neural networks, which directly transcribes videos of SL sign to word. We propose a novel approach, by using Video sequences that contain both the temporal as well as the spatial features. So, we have used two different models to train both the temporal as well as spatial features. To train the model on the spatial features of the video sequences we use the (Convolutional Neural Networks) CNN model. CNN was trained on the frames obtained from the video sequences of train data. We have used RNN(recurrent neural network) to train the model on the temporal features. A trained CNN model was used to make predictions for individual frames to obtain a sequence of predictions or pool layer outputs for each video. Now this sequence of prediction or pool layer outputs was given to RNN to train on the temporal features. Thus, we perform sign language translation where input video will be given, and by using CNN and RNN, the sign shown in the video is recognized and converted to text and speech.***

Keywords: CNN (Convolutional Neural Network), RNN(Recurrent Neural Network), SLR(Sign Language Recognition), SL(Sign Language).

I. INTRODUCTION

The sign language system is a technique for deaf and dumb individuals to communicate. Those who know sign language can communicate with dumb and deaf individuals and can talk and hear appropriately. Untrained persons, on the other hand, are unable to speak with the dumb and deaf, because a person can communicate with the dumb by learning sign language. For such people, a sign language to text system will be more useful in allowing them to converse with normal people more fluently. Sign language is a physical movement that uses the hands and eyes to communicate with the deaf and dumb. Different hand shapes and movements can be used to describe their feelings. The task is to translate their sign language into text or speech. Here, we are dedicated to finding an efficient solution to the gesture recognition problem. This research uses deep neural networks to create a sign language (SL) recognition framework that directly transcribes films of SL signs to words. As a result, both the temporal and spatial features were trained using two separate models. Convolutional neural networks (CNN) are used to train the model on the spatial properties of video sequences. Recurrent neural networks are used to train the model on temporal features (RNN). As a result, we do sign language translation, in which a video is provided as input, and the sign exhibited in the video is detected and converted into text and speech using CNN and RNN.

II. LITERATURE SURVEY

There has been a lot of research into hand sign language gesture recognition in recent years. The technology used to recognize gestures is listed below

A. Vision-based

In vision-based approaches, a computer camera is used to observe information from the hands or fingers. The Vision-Based approaches just require a camera, allowing for natural human-computer contact without the usage of any additional technologies. By describing artificial vision systems that are implemented in software and/or hardware, these systems tend to complement biological vision. This is a difficult challenge to solve since, to attain real-time performance, these systems must be background insensitive, illumination insensitive, and person and camera agnostic. Furthermore, such systems must be tailored to satisfy the requirements, which include accuracy and resilience.

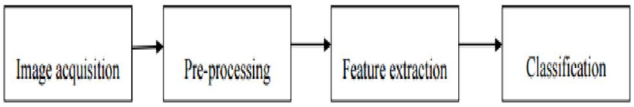


Fig1: Block Diagram of the vision-based recognition system

1) *Handshape recognition for Argentinian Sign Language using ProbSom [1]* This paper proposes a method for gesture identification of Argentinian sign language (LSA). The following are the two primary contributions of this paper: To begin, a database of hand shapes for the Argentinian Sign Language was created (LSA). Second, the problem is a supervised adaption of the self-organizing maps technique for image processing, descriptor extraction, and subsequent handshape classification. Support Vector Machines (SVM), Random Forests, and Neural Networks are all examples of state-of-the-art techniques that are compared to this one. Using the proposed descriptor, the ProbSom-based neural classifier attained an accuracy of 90%.

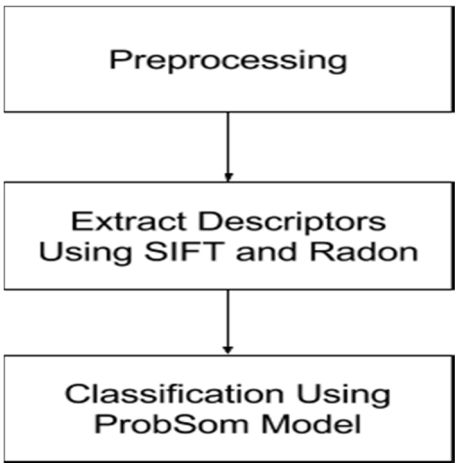


Fig 2: Block Diagram of Hand Gesture Recognition System for LSA

2) *Automatic Indian Sign Language Recognition for Continuous Video Sequence [2]* Data Acquisition, Pre-processing, Feature Extraction, and Classification are the four primary modules in the proposed system. Skin Filtering and histogram matching are applied in the pre-processing step, followed by Eigenvector-based Feature Extraction and Eigen value-weighted Euclidean distance-based Classification Technique. In this work, 24 different alphabets were considered, and a 96 percent identification rate was achieved.

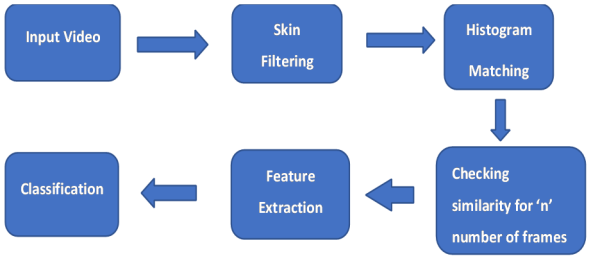


Fig 3: System Overview [2]

III. MOTIVATION

Hearing-impaired people communicate through hand signs, which makes it difficult for normal people to recognize their language. As a result, systems that recognize various signs and deliver information to ordinary people are required. The fundamental issue is that many indicators cannot be expressed in images; nevertheless, video sequences can. The key aim here is to detect the sign in the video sequences and translate it into text and speech that people can understand. Normal people have a hard time understanding hearing-impaired people's language, so a system that understands signs and gestures and relays information to normal people is needed.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

IV. EXISTING SYSTEM

A. Typically, Image classification is used for sign language recognition using machine learning algorithms like K nearest neighbour, Decision tree, and Support Vector Machines to classify the sign shown in the image.

B. Some researchers employed a Leap Motion Controller (LMC) sensor to measure the angles between the fingers' joints. Devices such as the Kinect sensor have also been used to extract the skeletal features of people.

C. Various works on gesture recognition have used finger-tip detection. • For hand gesture recognition, this system uses flex sensors, an onboard gyroscope, and an accelerometer to recognize hand gestures. Also employed for gesture recognition are continuous wave radar signals.

D. Previously, the Hidden Markov Model (HMM) was used to model sign language and other sequences, and it is still used for voice recognition systems, but it is inefficient for sign language. Hidden Markov models with limited capacity to capture temporal information have been used in previous methods dealing with continuous SL recognition.

V. PROPOSED SYSTEM

All the signs cannot be expressed in a single image, the system recognizes sign language exclusively from images to compensate for the limitations of the existing system, such as image categorization. As a result, we use CNN and RNN to classify videos. The spatial properties of the hand signs are extracted using CNN. The CNN model's output will be fed into the RNN model for sequence modelling, which will determine which sign is shown in the video. The discovered sign will be translated into text and speech.

VI. SYSTEM MODEL

The architecture provides the entire process flow of the system.

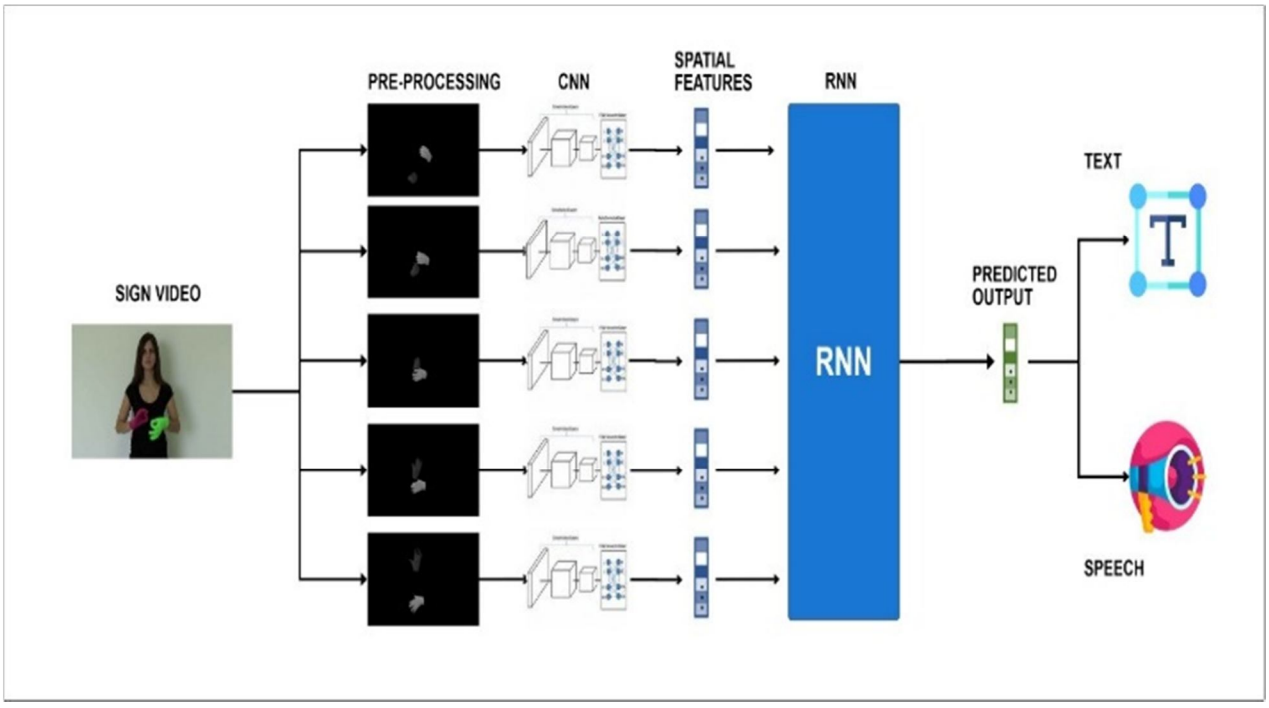


Fig-4 Architecture Diagram

The above architecture describes the entire processes involved in the system to convert signs from the video sequences to text and speech. The signed video uploaded by the user is divided into several frames with extracted hand gestures. Then the frames are given as the input to CNN(convolutional neural network) thus spatial features are extracted and an array of values are returned to RNN(recurrent neural network). It extracts the temporal features and recognizes the sign in the video. Then the sign is converted into the text and speech.

VII. IMPLEMENTATION

A. Algorithms Used

Since a video sequence comprises both temporal and spatial variables, video categorization is a difficult challenge. The spatial features are taken from the video frames, while the temporal features are extracted by linking the video frames with time. To train our model on each sort of feature, we used two different types of learning networks. We utilized a CNN to train the model on spatial features, and a recurrent neural network to train the model on temporal features.

1) *Convolutional Neural Network(CNN):* Convolutional neural networks, or ConvNets, excel in capturing data's local spatial patterns. They have a knack for spotting patterns and then classifying photos based on those patterns. The assumption in ConvNets is that the network's input will be an image. Because pooling layers are present, CNNs are unaffected by the rotation or translation of two similar images; thus, an image and its rotational image will be categorized as the same image. We used the Tensor Flow library's Inception-v3 model, which is a deep ConvNet, to extract spatial information from the frames of video sequences, due to the significant advantages of CNN in extracting the spatial features of an image. Inception is a huge image classification model with millions of parameters for images to classify.
2) *Recurrent Neural Network:* The sequence itself contains information, that recurrent neural networks (RNNs) use to perform recognition tasks. Because RNNs feature loops, their output is dependent on a combination of current input and prior output. RNNs have the disadvantage of being unable to learn long-term dependencies in practice. As a result, we used a Long Short-Term Memory (LSTM) model, which is an RNN with LSTM units. Even with noisy, incompressible input sequences, LSTMs can learn to bridge periods of over 1000 steps (Fig.5). The first layer is responsible for feeding data to the subsequent layers, the size of which is dictated by the size of the input. Our network is made up of 256 LSTM units in a single layer. A completely connected layer with SoftMax activation follows this one. Finally, a regression layer is applied to the specified input to conduct regression. To minimize the loss function, we used Adaptive Moment Estimation (ADAM), a stochastic optimizer. A deeper RNN network with three layers of 64 LSTM units each was also explored, as was a broader RNN network with 512 LSTM units. On a subset of the dataset, we discovered that the broad model with 256 LSTM units performed the best.

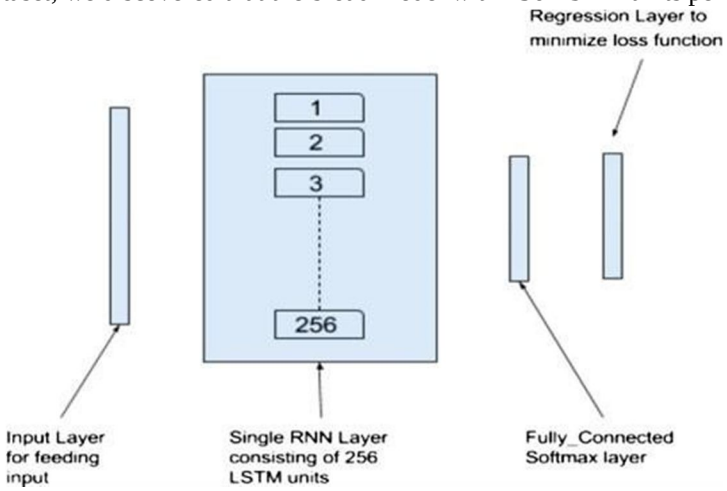


Fig-5 Architecture of the proposed RNN model.

B. Methodology

Two approaches for training the model on temporal and spatial features were used, and they differed in the way inputs were given to the RNN to train it on temporal features.

1) *Prediction Approach:* The inception model (CNN) was used to extract spatial data for individual frames, and the RNN was used to extract temporal features. After that, each video was represented by a sequence of CNN predictions for each of its constituent frames. This was fed into the RNN as input. Frames were retrieved from each video matching each gesture, and background body elements other than hands were deleted to create a grayscale representation of hands that avoided the model acquiring colour-specific information (Fig. 6). The CNN model was given frames from the training set to train on spatial features. The model was then employed to generate and store predictions for the training and test data frames. The LSTM RNN model was then trained on the temporal features using the predictions corresponding to the frames of the training data. The predictions corresponding to the frames of the test data were submitted to the RNN model for testing after it had been trained.

2) *Train CNN(Spatial Features) and Prediction:* CNN's role is based on the inception concept. All frames from each video corresponding to each gesture "X" were labelled "X" and provided to the inception model for training from the training dataset for each gesture "X."

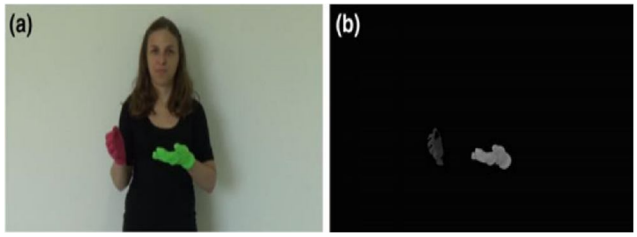


Fig-6 a sample frame from the dataset. **b** Frame after background removal

The trained model was used to make predictions for frames from both the train set and the test set of films. Each video of a gesture was broken down into a series of frames. The video is then represented as a sequence of guesses after CNN has been trained and predictions have been made.

3) *Training RNN (Temporal Features)* The videos for each gesture are fed to RNN as a sequence of predictions of its constituent frames. The RNN learns to recognize each gesture as a sequence of predictions. After the Training of RNN completes a model file is created (Fig. 7)

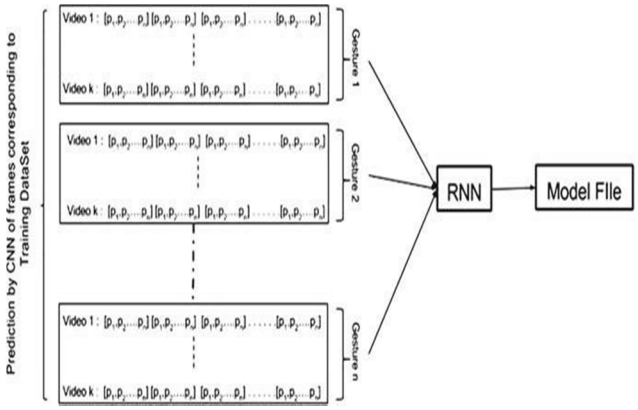


Fig-7 Training RNN after getting the results from CNN

VIII. RESULTS

The proposed system was successfully tested to denote its effectiveness and achievability. Thus sign in the video sequence is converted into the text and speech. It is the command-line interface(CLI),fig.8 where the user can upload the sign video. The result will be displayed as follows:

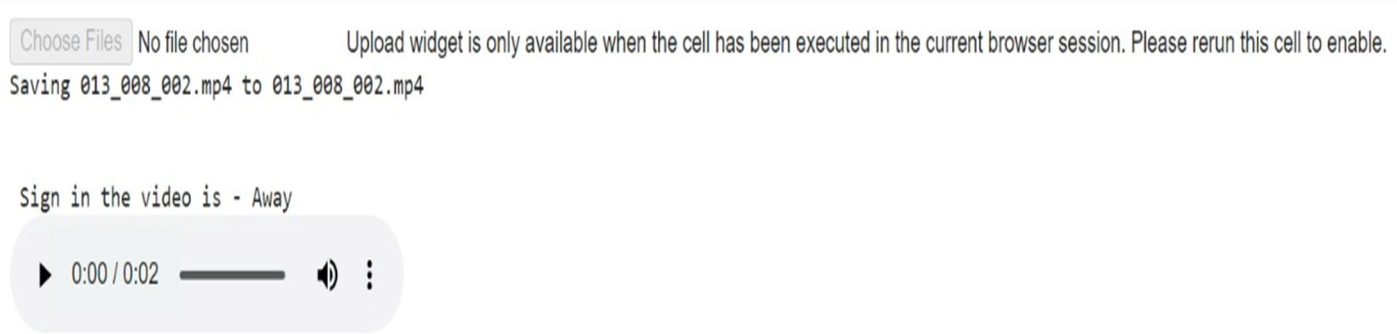


Fig 8 Interface to upload sign video.

After a user uploads the video, the sign in the video is displayed as text, and the sign spells out the speech as in fig 9.

Fig 9 Text and speech output of sign

IX. CONCLUSION

We presented a vision-based method for interpreting single hand motions from Sign Language in this paper. To classify the spatial and temporal features, this study used a prediction approach. The geographical features were classified using CNN, whereas the temporal features were classified using RNN. This demonstrates how CNN and RNN may be used to learn spatial and temporal information and interpret sign language gesture videos as text or speech.

X. FUTURE SCOPE

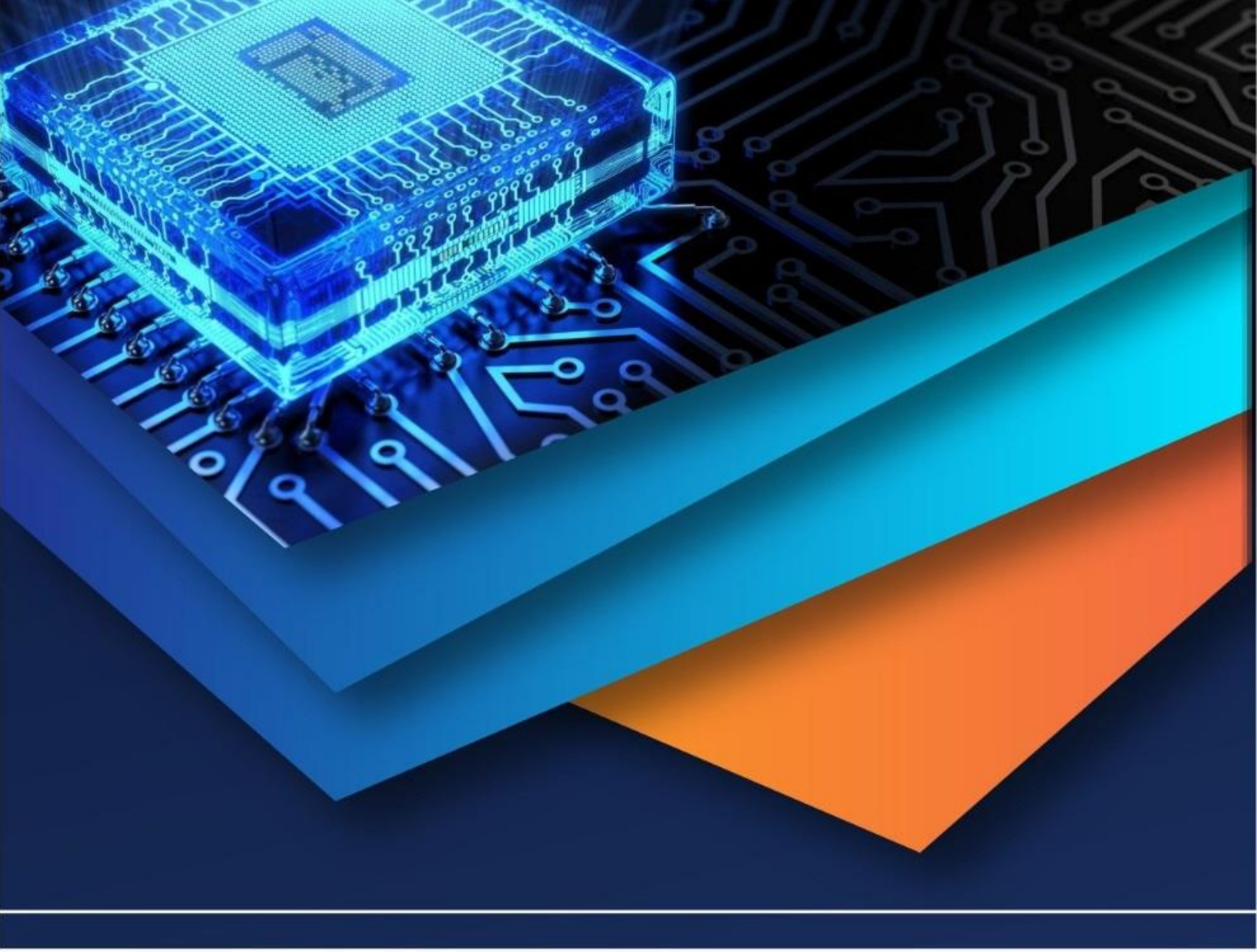
This research can be expanded in the future to recognize continuous sign language motions with greater accuracy. This strategy for individual gestures can likewise be applied to sign language at the sentence level. In addition, the current procedure employs two distinct models: training inception (CNN) and training RNN. Future work could concentrate on fusing the two models into a single model. The proposed system can be developed and deployed utilizing Raspberry Pi in the future. Image processing should be upgraded so that the system can communicate in both directions, i.e., it should be capable of converting conventional language to sign language and vice versa, and it should be able to concentrate on transforming the sequence of gestures into sentences and subsequently text and voice.

XI. ACKNOWLEDGMENT

We express our sincere gratitude to our guide, Assistant Professor Mr. G. Sekhar Reddy for suggestions and support during every stage of this work. We also convey our deep sense of gratitude to Professor Dr. K. S. Reddy, Head of the Information Technology department.

REFERENCES

[1] “Handshape recognition for Argentinian sign language using problem”. Journal of Computer Science and Technology 16(2016). Ronchetti, Franco, Facundo Quiroga, Cesar Armando Estrebou and Laura Cristina Lanzarini.
[2] “Automation Indian Sign Language Recognition for Continous Video Sequence.” ADBU Journal of Engineering Technology 2, no. (2015). Singha, Joyeeta, and Karen Das.
[3] “Continuous Indian Sign Language Gesture Recognition and Sentence Formation.” Procedia Computer Science 54(2015): 523-531. Tripathi, Kumud, and Neha Baranwal GC Nandi.
[4] “Si language recognition using subunits.” Journal of Machine Learning Research 13 no. Jul(2012): 2205-2231. Cooper, Helen, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden.
[5] “Sign language recognition.” In visual Analysis of Humans, pp. 539- 562. Springer London,2011. Cooper, Helen, Brian Holt, and Richard Bowden.
[6] “Learning long-term dependencies with gradient descent is difficult.” IEEE transactions on neural networks 5, no.2(1994): 157-166. Bengio, Yoshua , Patrice Simard, and Paolo Frasconi.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY