Vision Based Indian Sign Language Recognition Model

Karthik H S
Under-Graduate Student, Dept. Of Computer Science & Engineering Jyothy Institute of Technology, Visvesvaraya Technological University Thataguni Post, Bengaluru-560082, India

Pannagadhara K S
Under-Graduate Student, Dept. Of Computer Science & Engineering Jyothy Institute of Technology, Visvesvaraya Technological University Thataguni Post, Bengaluru-560082, India

Suhas Hanjar
Under-Graduate Student, Dept. Of Computer Science & Engineering Jyothy Institute of Technology, Visvesvaraya Technological University Thataguni Post, Bengaluru-560082, India

Vinayak Rangannavar
Under-Graduate Student, Dept. Of Computer Science & Engineering Jyothy Institute of Technology, Visvesvaraya Technological University Thataguni Post, Bengaluru-560082, India

Saravana M K
Associate Professor, Dept. Of Computer Science & Engineering Jyothy Institute of Technology, Visvesvaraya Technological University Thataguni Post, Bengaluru-560082, India

Abstract: Sign language (SL) is essential for deaf and hard-of-hearing people to communicate. However, these sign languages are not known to most healthy people. There is no universal language like verbally spoken languages as every country has its native language, so every country has its way of sign language. In India, we use Indian Sign Language (ISL). This survey provides an overview of the essential Indian sign language recognition and its translation work. Much research has been conducted in American Sign Language (ASL), but unfortunately, the same cannot be in the case of Indian Sign Language. There are different types, ways between sign language recognition processes worldwide. However, a few tasks are primarily similar, such as Pre-processing, feature extraction, and classification. The main focus of our proposed method is to design an ISL (Indian Sign Language) hand gesture motion translation tool for helping the deaf-mute community to convey their ideas by converting them to text format. We used a self-recorded ISL dataset for training the model for recognizing the gestures. CNN (Convolutional Neural Network) was used to extract the image features like skeletal features. LSTM (Long Short Term Memory) model was used to classify these gestures and then are translated into text.

Keywords: Hand Gesture Recognition; Convolutional Neural Network (CNN); Indian Sign Language (ISL); Long Short Term Memory (LSTM)

I. INTRODUCTION

As important as a spoken language, sign language uses human hand actions to express the meaning. Even Sign languages are formal languages with different ways and grammar. As sign language grows, it can inherit a few elements from standard language (spoken language). In many sign languages, fingerspelling is used if we have to show the English alphabet. A gesture done using a hand is called a sign. A single hand sign can differentiate into three parts: The shape of the hand, the position, and the movement. As we know, there are N number of languages used by people to communicate across the world, and all those languages differ. In the same way, sign language has hand gestures and visual representations of many different types. The most known Sign languages are American Sign Language (ASL), French Sign Language (LSF), and Indian Sign Language (ISL). Sign language differences are signing, pronunciation, slang, and some gestures.

Sign Language Recognition (SLR) aims to develop algorithms to identify the sign's sequences and understand their meaning correctly.

Many approaches to Sign Language Recognition mainly solve the problem of Gesture Recognition (GR). Hand gestures constitute assertive inter-human communication, and they can consider it a convenient means of communication between humans and machines. The main elements of a hand Recognition System (HRS) are collecting data, locating the hand, extracting the features from the hand, and gesture recognition. Solving sign language recognition problems can be done in many ways but are mainly categorized into two types: capturing the shape of the hand and using the motion of hand gestures. The other method is a video sequence based on signs.

This survey gives a technical overview of the research going on in Indian Sign Language (ISL). We have

summarized the technical insights into the approaches in designing ISL. We believe that this survey will be helpful to researchers who are yet to start their work in this field.in detecting problems and maintain the integrity of the system.

II. LITERATURE SURVEY

To enhance the recognition accuracy and efficiency among different action videos, there has been some investigation on how to utilize structural manifold information; one which they used in the paper is DML to incorporate the manifold of training samples into deep learning. The discriminative capacity of the next layer can be promoted by applying it to a CNN. The over-fitting problem can be alleviated by applying DML on a restricted Boltzmann machine. [4]

Describes a basketball shooting gesture recognition method based on extraction of image feature and ML. Action posture data of players is collected by the extraction of image feature method, and by using time and frequency domains, multi-dimensional motion posters are extracted. Then, by using Gaussian Hidden and feature selection, accurate classification and recognition are obtained. This method gains high efficiency and low error rate compared to other old basketball action recognition technology. [6]

Unsupervised representation learning has no necessity of label info with observed data. Deep learning models consume a high amount of time, and because of this entity, there are many ML models which are adapting well-trained models obtained in the supervised and end-to-end manner as extractors of features. In a video sequence of human actions are 3D signals which contains visual appearance and motion dynamics of objects and humans. Hence, capturing both spatial and temporal correlation in videos along with the data representation approach is meaningful. The recently researched human motion recognition models build classifiers based on convolutional networks, which is a deep learning structure. Requirement of these models are a huge amount of training data. But without retraining, and models cannot be recognized by samples from the distinct dataset. A new 3D Deconvolutional network for representation of learning of high dimensional data in which, through optimization approach, high features are obtained. This proposed 3DDN decomposes video frames into spatiotemporal features in an unsupervised way. [15]

Recognition of sign language is accomplished by deep learning techniques consisting of hand semantic segmentation, feature representation of hand shape, and deep recurrent neural network. A semantic segmentation method named as DeepLabv3+ is trained with the help of hand images which are pixel-labeled. This is done for extraction of hand regions in the input video for every frame. After the extraction process, hand regions are cropped and scaled to 64 ×64 pixels to alleviate hand scale variations. All cropped images are converted into grayscale images. This process is achieved with the help of a single layer Convolutional Self-Organizing Map. Deep Bi-directional LSTM RNNs are used to recognize the sequence of extracted feature vectors. [1]

The proposed methodology uses the MediaPipe framework to extract the hand landmark from input video for the recognition of Thai sign language. After processing, with the help of Recurrent neural networks, that landmark is used to build the model for recognition of hand gestures. The model is built with LSTM, BLSTM, and Gated Recurrent Unit. Recorded five videos of 5 gestures,500 videos in total, for testing. The videos are 50 FPS with format H.264 and don't require any equipment other than a mobile camera. Hand key point is extracted and written to a text file (CSV).42 key points will be recorded into the CSV file. With the hand's key point extracted, it is suitable for model training using RNN. [3]

Signals are obtained for each input video by extracting skeleton joints of the person doing the action. After the signals are obtained, Common Spatial Patterns (CSP) algorithm is applied to these signals. It is commonly used in the field of electroencephalography. A summary image is obtained from the signals for every video. Positions of the joints in the skeletal extraction process are then used to create signals. These images are classified into two types of approaches; global visual descriptors and CNN. Visual descriptors explain essential characteristics such as shape, color, texture, or motion. OpenPose is used for the extraction of skeletal joints. Two datasets are used for this experiment. [13]

The proposed solution for hand detection and gesture recognition is that raw data coming from Kinect can be used to recover depth information on all the pixels of an image. Then a novel faster algorithm is used to identify each point within a given depth interval. Fingertips are localized by using the k-curvature algorithm. For dynamic gesture recognition, the DTW algorithm is used. The approach allows the user to record a sequence of reference gestures saved in the FIFO list. In the end, 55 static and dynamic gestures were properly recognized while testing the model. We can add another module to accurately get the user's hand's size. [11]

The proposed method for feature extraction is the training phase, which is performed using DLSTM with LSTM and RNNs. According to the research, RNN is more suitable than CNN to support the data. Here the implementation is done on ASL (American sign language). To overcome the failures, we can merge the features extracted from hand with stream information and integrate the network with the 3D CNN. [2]

The approach is to solve the problems like sequence learning. A sequence of specialized expert systems is referred to as SubUNets (novel deep learning architecture). Spatial feature representations can be learned using 2D CNNs, and for temporal aspects, RNNs are used. The problem faced was the vanishing gradient problem which was then solved by using LSTM. The experiments showed that using SubUnets, the network

generalizes better. For future work, investigating more about hierarchical SubUNets will be helpful. [5]

The research aims to solve the problem of fewer models which can recognize the hand sign language. 3D convnet and Bi-directional LSTM are used for multi-model recognition. BLSTM-3D uses deep learning technology to achieve hand sign recognition, CNN for extracting the features of hand gestures, RNN to learn the video sequences, and a combination of CNN and RNN to learn the spatiotemporal features. [10]

The methodology is based on processing video sequences and extraction and fusion of discriminate features for the classification of video sequences to isolated signs. With the help of the VGG-16 network, which is trained with both raw video sequences and optimal flow images on ImageNet, also they used FlowNet2 (which is an accurate optimal flow deep network) to get optimal flow images. For future work, creating a new dataset for sign language which will be suitable for a deep learning framework is very much needed. [8]

The aim of this research is hand gesture recognition using CNN algorithm, by identifying gestures in the image. CNN is used here because of the layers it has. The primary layer is the feature extraction layer and another layer is the characteristic map layer. The system uses the CNN algorithm as an interpreter which interprets and builds statements, which means it converts the gestures into a statement or text. Keyframe extraction from video, Keyframe extraction from a set of frames, Applying CNN Algorithm, and Mapping Gestures into the text are the process of the method. 6 types of gestures were recognized by the model. [7]

Based on continuous sign language recognition methods, current methods of SLR use the static module as the base and sentence segmentation to be continuous. To solve the issues of current methods like sentence segmentation, manual labeling, and word alignment bidirectional spatial-temporal LSTM fusion network is proposed. RGB videos are taken as input then detect hand and face for each frame using R-CNN. The frame information is taken in two parts 1)Spatial-temporal 2) Local hand information. Both are fed to a bi-directional ST-LSTM encode-decode framework. The SLR accuracy achieved 81.22% on the 500 CSL dataset and 76.12% on the RWTH-PHOENIX-Weather dataset. [14]

Previous records and studies show that they have issues in real-time conditions. A novel facial expression database contains 30,000 facial pictures with real-time conditions and multiple people of different ages and races. Each picture is labeled by approximately 40 annotators. Lab-controlled expressions differ a lot compared to real-time expressions. A new deep locality-preserving convolutional neural network (DLP-CNN) method is introduced to overcome this problem. The method improves the classification power of deep features by storing the locality closeness, maximizing inter-class scatters. DLP-CNN performs better than traditionally used methods for expression recognition in real-time. [9]

Human vision can display a natural mechanism known as human attention, where the system tends to focus on only a part of the scene for quick feature extraction. Recent deep learning models try to imitate this mechanism. Such models are known as Machine/Neural/Artificial attention. The human attention mechanism is the key to achieving the best deep networks. [12]

III. PROPOSED METHODOLOGY

Our report discusses a vision based classification system having dynamic signs of Indian sign language (ISL) to recognize hand signs. A dynamic sign motions contains complex motion of gestures with more movements, unlike a static sign. Special arrangement of the hand determines a static sign whereas a sequence of the hand movements and configurations determine a dynamic sign. We used a self-customized dataset from the Indian sign language dictionary. CNN algorithm is used for feature extraction. The hand region is considered as the region of interest by the algorithm. Then the features extracted are given as an input to the LSTM model and then signs are converted to text form.

A. Dataset Acquisition:

In this stage, the images representing dynamic signs of ISL are retrieved. Hand gestures were taken in video format and converted to a series of video frames or images. This dataset is created by taking some of the significant words from the Indian Sign Language dictionary which consists of more than 6000 sign words. Each sign was individually recorded 30 times using a web-cam at 30 fps. After this phase, the videos are taken for processing and are converted to image frames.

B. Feature Extraction:

In our suggested model, CNN (Convolutional Neural Network) is used to extract features from the video frames of created video recordings . After extraction, the characteristics are kept in a file. For feature extraction, a variety of alternative machine learning techniques can be applied. But among those, CNN is the most effective deep learning method. CNN was therefore utilised to extract probable categorization parameters for a variety of video frames.

C. Classification:

The features extracted using CNN are fed into the system that classifies the sign as input. The LSTM (Long Short Term Memory) network is a classification tool. The image is classified as text using the LSTM model. The advantage of using LSTM for classification is that no expertise is required to manually engineer input features. The neural model network is trained to classify hand motions during the training phase. During the testing phase, image frames from these videos are used to test all of the signs. The LSTM model is used for the classification of the extracted features.

An LSTM unit has a cell, an input gate, an output gate, and a forget gate for feed-back. The flow of information into and out of the cell is controlled by these 3 gates. The cell keeps track of the values over arbitrary time intervals. The input gate checks the flow of extented latest value into the block of cell. The forget gate regulates the extent to which a value remains in the cell. Whereas the output block manage the value used in computing activation of the LSTM unit.

D. Testing:

We used a self-recorded custom dataset for our prediction due to the limited number of standardised datasets for ISL. The dataset is made up of various signs, each of which was recorded 30 times. Each sign gesture motion was distinct from the next. Everyday words can be found in our self-recorded dataset of video signs.

IV. IMPLEMENTATION

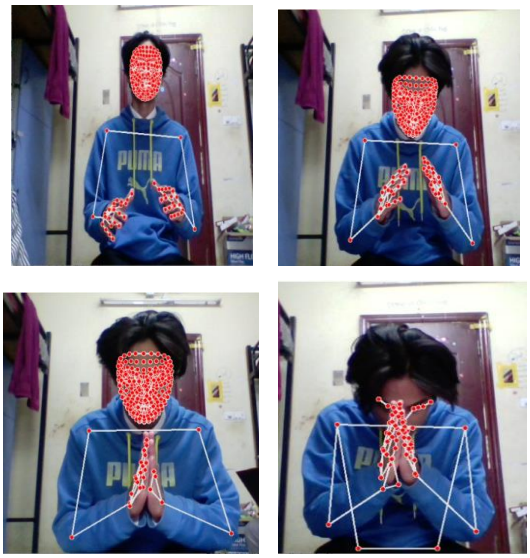

Fig 1. Hand Gestures frames for the word “Namaste”

Some of the signs used in our dataset:

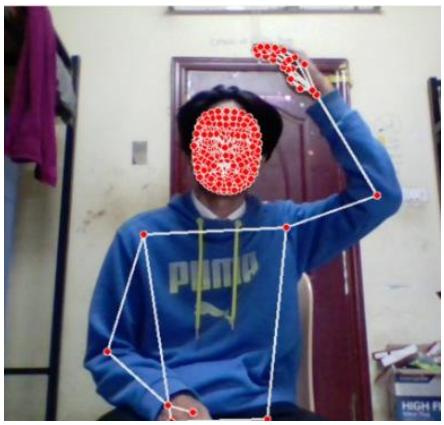

Fig 2. Gesture for word “Tall"

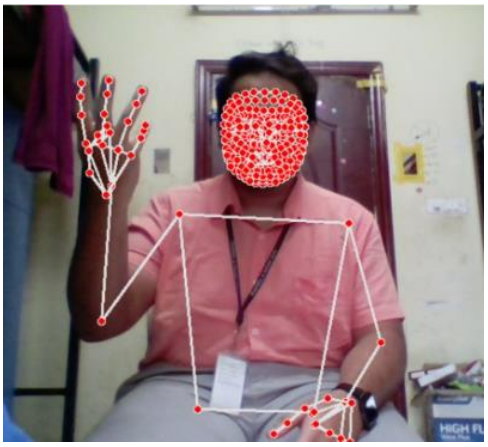

Fig 3. Gesture for the word “What?”

Fig 4. Gesture for the word “On”

Fig 5. Gesture for the word “Maximum”

V. CONCLUSION AND FUTURE WORK

In our final consideration in Sign Language Recognition (SLR) research, we concluded that it is open

to a few problems: Setting up valid data set for the project where the users can use the data set for their research which can also be improved and compared to other data sets, incorporating hand gestures for a larger data set and very important in Sign Language (SL).With recent deep learning advances, there has been tremendous progress in the domain of gesture recognition. The detection of sign language has a wide range of applications. To improve efficiency of the proposed system, datasets can be expanded and word sentences can be added. The project can be fully automated and expanded to include text-to-speech conversion. Different models can also be tried for classification purposes.

VI. CONFLICT OF INTERESTS

The authors declared no conflict of interests with respect to authorship, work and publication of this paper.

ACKNOWLEDGMENT

We authors like to acknowledge to our esteemed institution “Jyothy Institute of Technology” for providing us an opportunity and to our principal Dr. K Gopalkrishnan for providing us adequate facilities to undertake this project work. We would also like to thank Head of Department, C.S.E Dr Prabhanjan S, our family and friends who have guided and helped us through this work and preparation of this manuscript.

REFERENCES

[1] Saleh Aly and Walaa Aly. Deeparslr: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition. IEEE Access, 8:83199–83212, 2020.

[2] Danilo Avola, Marco Bernardi, Luigi Cinque, Gian Luca Foresti, and Cristiano Massaroni. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. IEEE Transactions on Multimedia, 21 (1):234–245, 2018.

[3] Anusorn Chaikaew, Kritsana Somkuan, and Thidalak Yuyen. Thai sign language recognition: an application of deep neural network. In 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, pages 128–131. IEEE, 2021.

[4] Xin Chen, Jian Weng, Wei Lu, Jiaming Xu, and Jiasi Weng. Deep manifold learning combined with convolutional neural networks for action recognition. IEEE transactions on neural networks and learning systems, 29 (9):3938–3952, 2017.

[5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In Proceedings of the IEEE International Conference on Computer Vision, pages 3056–3065, 2017.

[6] Rong Ji. Research on basketball shooting action based on image feature extraction and machine learning.IEEE Access, 8:138743–138751, 2020.

[7] Rashmi R Koli and Tanveer I Bagban. Human action recognition using deep neural networks. In 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pages 376–380. IEEE, 2020.

[8] Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. A deep learning approach for analyzing video and skeletal features in sign language recognition. In 2018 IEEE International Conference on Imaging Systems and Techniques (IST), pages 1–6. IEEE, 2018.

[9] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing, 28 (1):356–370, 2018.

[10] Yanqiu Liao, Pengwen Xiong, Weidong Min, Weiqiong Min, and Jiahao Lu. Dynamic sign language recognition based on video sequence with blstm-3d residual networks. IEEE Access, 7:38044–38054, 2019.

[11] Guillaume Plouffe and Ana-Maria Cretu. Static and dynamic hand gesture recognition in depth data using dynamic time warping. IEEE transactions on instrumentation and measurement, 65 (2):305–316, 2015.

[12] LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. IEEE Transactions on Multimedia, 2020.

[13] Chengcheng Wei, Jian Zhao, Wengang Zhou, and Houqiang Li. Semantic boundary detection with reinforcement learning for continuous sign language recognition. IEEE Transactions on Circuits and Systems for Video Technology, 31 (3):1138–1149, 2020.

[14] Qinkun Xiao, Xin Chang, Xue Zhang, and Xing Liu. Multi-information spatial–temporal lstm fusion continuous sign language neural machine translation. IEEE Access, 8:216718–216728, 2020.

[15] Chun-Yang Zhang, Yong-Yi Xiao, Jin-Cheng Lin, CL Philip Chen, Wenxi Liu, and Yu-Hong Tong. 3-d de-convolutional networks for the unsupervised representation learning of human motions. IEEE transactions on cybernetics, 2020.