Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Mudra: Convolutional Neural Network based Indian Sign Language Translator for Banks

Gautham Jayadeep, Vishnupriya N V, Vyshnavi Venugopal, Vishnu S, Geetha M

Dept of Computer Science
and Engineering
Amrita Vishwa Vidyapeetham
Amritapuri,India
gauthamjayadeep56@gmail.com, vpriyanambisan@gmail.com, vyshnavivenugopal98@gmail.com, vishnu.sasikumar581@gmail.com

Abstract-Signlanguageisamediumofexpressing thoughts and feelings by the deaf-dumb community. It could be extremely challenging for deaf-mute people to communicate efficiently in banks, where they might have to explain their needs. There are very few people who can understand sign language. The main focus of our proposed method is to design an ISL (Indian Sign Language) hand gesture motion translation tool for banks for helping the deaf-mute community to convey their ideas by converting them to text format. In the fields of ASL (American Sign Language) and other languages, ample amounts of work have been done. Apart from other algorithms, our proposed method recognizes human actions considering isolated dynamic Indian signs related to the bank as a novel approach. There are very few research works carried out in this field of ISL recognition for banks. Over and above that, an insufficient amount of dataset along with dissimilarity in gestures length was a difficulty. We used a self-recorded ISL dataset for training the model for recognizing the gestures. Unlike image data, the video domain was a new challenge. Larger lengthened video gestures were taken and actions were recognized from a series of video frames. CNN (Convolutional Neural Network) named inception V3 was used to extract the image features. LSTM (Long Short Term Memory), an architecture of RNN (Recurrent neural network) classified these gestures and are translated into text. Experimental results display that this approach towards isolated word dynamic hand gesture recognition systems provides an accurate and effective method for the interaction between non-signer and signer.

***Index Terms*—Hand gesture recognition, Convolutional neural network, Indian Sign Language, Long short term memory, and gesture.**

I. Introduction

The objective of the sign language detection system is to deliver a strategy to translate signs patterns in sign language related to the bank into text. It would be effective for the deaf-mute communities to do any of the bank procedures without anyone's help. Different regions or countries use different sign languages. Indian sign language (ISL) is chosen so that the system can be used by the Indian banks for communication between deaf and normal people.

Sign Language (SL), is one of the widely used natural and expressive ways of exchanging information visually among deaf communities[1]. The main goal of the project is to classify the bank related symbols or sign gestures from the Indian Sign Language dictionary and build an alternative communication medium for the hearing impaired community. The motivation for the development of such a helpful application came from the fact that it would help in expanding social awareness and will lower the isolation of these people from bank activities. Less research has been produced in this field of ISL due to the difficulties in complex gesture patterns.To tear away the communication barrier between the verbally challenged people and banks, our contribution considers a methodology to identify isolated words dynamic hand gesture recognition system for ISL (Fig1) using Convolution Neural Networks (CNN) based feature extraction combined with a Long short-term memory (LSTM), which is an RNN architecture. Convolution is seen as the most efficient way of feature extraction as it reduces the data dimension and redundancy of the dataset. LSTM is used in the field of deep learning for image classification from the video sequences.

It is really challenging to develop an automatic sign recognition system for bank-related ISL recognition. The major challenge was the difference in length of the sign videos. Hand shapes were complicated. It is also more complex in the case of ISL compared with other sign languages since most of the signs require both hands. For some of the signs, hand contacts the body. Change in backgrounds was also a factor. Since ISL was standardized only recently, we considered all these challenges to bring a better version of the hand gesture recognition system for banks.

II. Related Works

Earlier the datasets for images and videos were generated with basic camera technologies. In our ISL hand gesture recognition system, video datasets were required and videos in the created dataset were of different lengths, which was a new challenge. And also it was a new approach to discuss the ISL hand gesture recognition system for banks.

The paper [2] presents a sign dectecton technique having dynamic signs constructed on a three Dimentional ConvNet and LSTM networks. The methodology include of three sections. The hand objects were localized in video frames. Then the temporal features from the video sequences were automatically

extracted. Finally, the video sequences were classified and the dynamic signs is identified correctly.

In recent years in the field of deep learning, Convolutional neural networks have been deeply successful. The main focus of [4] this paper is to implement a vision-based application that offers a translation from SLR to text for ASL to provide a channel for communication between the verbally disabled people. The proposed models CNN and RNN results in pulling out spatial and temporal features from the video sequence dataset. To avoid gradient vanishing problems, the deep learning model LSTM(Long Short Term Memory), an architecture of RNN was used. In [5], Vivek Bheda and N. Dianna Radpour proposed a methodology of using a stochastic gradient descent mini-batch supervised learning method for the classification of the images for every digit (0-9) and letter, in American sign language using deep convolutional neural networks. Different researchers have used different implementations techniques in this field of SLR.

III. Proposed Method

Our thesis discusses a vision based classification system having dynamic signs of Indian sign language (ISL) for bank purposes. 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. The figure below (Fig 1.) represents some of the symbols used in our dataset.

Fig. 1. Some Signs used in our Dataset

ARCHITECTURE

Sign language has well-arranged hand-code gestures, each gesture/sign has a corresponding word meaning allotted to it. Here we follow the main construction of our bank-related recognition system consisting of the following steps, Dataset Acquisition, image Pre-processing, Extraction of features, and lastly classification or detection of signs to text. The simple flow diagram of our SLR model is represented in figure 2 and the overall structural architecture for our hand gesture recognition system is mentioned in Figure 3.

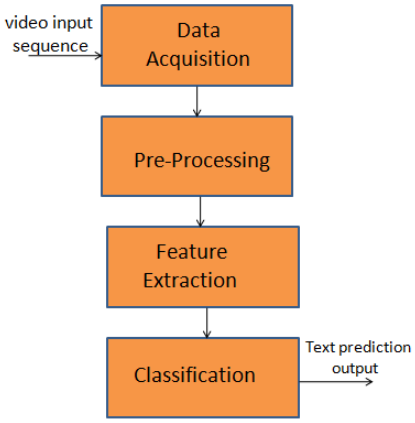

Fig. 2. Simple Flow diagram of SLR system.

In our project, we are using a CNN model called InceptionV3 for feature extraction, which is trained on more than a million images from the ImageNet database. The CNN comprise of different layers which includes;

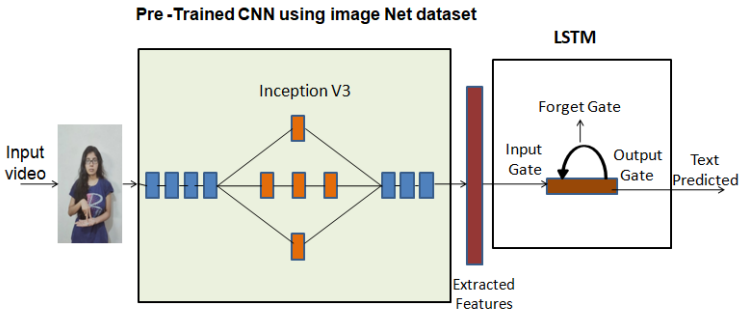

Fig. 3. Architecture for hand gesture recognition system

convolutional layers, followed by max-pooling layers, the ReLU correction layer, and 2 fully-connected layers. Convolutional layer purpose is to identify presence of set features in the images given as input. The pooling reduces the size of images and preserves their important characteristics. The ReLU correction layer converts all negative values received as input

to zeroes. The endmost fully-connected blocks classifies the image as the input for the LSTM network. But we used CNN only for feature extraction. In CNN the convolution process runs layer by layer. But only a few layers in Inception V3 are used for feature extraction. They are conv2D, Maxpooling2D, Avg pooling.

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. The parameters of LSTM used are model, kernel_filters, init, reg_lambda.

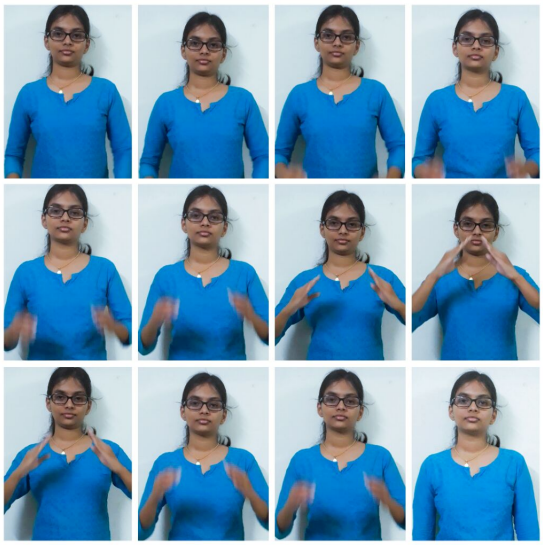

Fig. 4. Image Frames

ALGORITHM

1.Bank-related video datasets are created.
2.The videos are converted to image frames.
3.Training
(3.1)Cnn$_{f}eatureExtraction(videos, image,$ inceptionV3, imageset)
(3.2)Lstm$_{c}lassification(model, kernel_filters,$ init, reg_lambda)
4.Testing
(4.1)Signs are classified to text format.

METHODOLOGY

A. Dataset Acquisition

In this stage, [6] the images related to the bank representing dynamic signs of ISL are retrieved. Hand gestures were taken in video format and converted to a series of video frames or images. Each image has a resolution of 1080x1920. In this system, we used a self-customized dataset from the Indian sign language dictionary containing different lengthened gesture videos taken from students. The dataset comprises 1100 videos representing bank-related signs and everyday signs. 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 multiple times using a mobile phone at 40 fps. After this phase, the videos are taken for processing and are converted to image frames. 80% of these frames are taken for training and the rest 20% for testing. Fig 4 shows the image frames obtained.

B. Feature Extraction

In our proposed model, features are taken out from the video frames obtained from the created videos representing dynamic signs related to the bank using CNN (Convolutional Neural Network). The extracted features are stored in a file after extraction. Many other machine learning algorithms can be used for feature extraction. But among that CNN is the best technique in the field of deep learning. So for a wide range of video frames, CNN was used to extract potential characteristics for classification.

C. Classification or Recognition Phases

The features extracted from the previous layer is taken as an input to the system that classifies the sign. LSTM (Long Short Term Memory) network is considered as an classification tool. Using the LSTM model, the image is classified in the form of text. The benefit of using LSTM for classification is that it does not require expertise to manually engineer input features. In the training phase, the neural model network is trained to classify the hand motion. The dataset comprises of 1100 videos. In the testing phase, image frames of these videos are used for testing which includes all the signs.

IV. Test Result and Performance Evaluation

In this research work, we have discussed the implementation of an automatic sign language fingerspelling interpretation system for bank-related procedures, where algorithms based on deep neural networks were deployed. We aimed to suggest a system to aid the communication of dead and mute. people using sign language and provide a better communication bridge between signers and banks. Due to the reduced number of standardized dataset for ISL, we used a self-recorded custom dataset for our prediction. The dataset contains a total of 1100 videos of different bank related signs taken from different people. Dataset consists of different signs where each video was repeatedly recorded. Each sign gesture motion was different from one another. Video signs in our self-recorded dataset include vocabulary from the bank category and other everyday words. The new proposed improvement in this field is to test the models with larger lengthened gesture videos varying from 1 to 6 seconds at 40fps with a resolution of 1080x1920.

TABLE I represent all the symbols and the count of each sign video recorded. As per the implementation architecture of our proposed SLR(Sign language recognition) system is addressed in Fig 3, the video sequence data is gathered, filtered, and cleaned. Later before entering the model, the video sequence was broken down into image frames as shown in Fig 4. Then with the help of CNN (Convolutional neural network) inception V3 architecture, the features were

TABLE I

Sl. No	**Dataset**		
	Signs	***No.of videos***	***Total No of videos***
1	Come	70	
2	Sleep	70	
3	Food	70	
4	Home	70	
5	Red	70	
6	Orange	70	
7	Hi	70	
8	Tea	60	
9	Love	60	
10	Stand	60	1100
11	Debit Card	45	
12	Balance sheet	45	
13	Working Hours	45	
14	Loan	45	
15	Deposit	45	
16	Withdrawal slip	45	
17	Expiry	40	
18	Teller	40	
19	Provident Fund	40	
20	Transaction	40	

extracted. Feature Extraction is done to identify the features of interest and reduce computational time without sacrificing accuracy. Here the area of interest is the image frames as a whole and those features are saved into a NumPy array.

TABLE II

Signs	Debit Card	Balance sheet	Working Hours	Loan	Deposit	Expiry	Teller
Debit Card	1	0	0	0	0	0	0
Balance sheet	0.17	0.83	0	0	0	0	0
Working Hours	0	0	0.8	0	0.2	0	0
Loan	0	0	0	0	1	0	0
Deposit	0	0	0	0	0	1	0
Expiry	0	0	0	0	0	0	1
Teller	0.25	0	0	0	0	0	0.75

During training, we obtained a confusion matrix (TABLE II and III) for the datasets. Later the extracted features were classified or interpreted into text format using the LSTM model.

Fig 5 shows the number of times the recognition system rightly classifies the bank related signs. The testing was repeated many times and we obtained

TABLE III

Signs	Come	Sleep	Food	Home	Red	Orange	Hi	Tea	Love
Come	1	0	0	0	0	0	0	0	0
Sleep	0.43	0.57	0	0	0	0	0	0	0
Food	0	0	1	0	0	0	0	0	0
Home	0	0	0	1	0	0	0	0	0
Red	0	0	0	0	1	0	0	0	0
Orange	0	0	0	0	0	1	0	0	0
Hi	0	0	0	0	0	0	1	0	0
Tea	0	0	0	0	0	0	0	1	0
Love	0	0	1	0	0	0	0	0	1

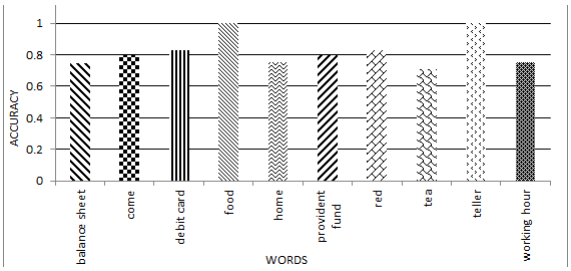

Fig. 5. Bar graph

the following information from the experiments. 1.The properly classified bank-related signs were debit card, loan, deposit, and expiry. In the case of everyday signs, the system correctly predicted the words: come, food, home, red, orange, hi, tea, and love. 2.The system also confused certain words with the other. They were balance sheet, working hours and teller etc.

TABLE IV

S2. No	**Accuracy Comparison Table**	
	Dataset	***Accuracy***
1	Bank Category	85%
2	Everyday	92%
3	Entire dataset	81%

While comparing the bank vocabulary gestures and every day sign symbols, an accuracy variation was measured and it is illustrated in Table IV. It was recognized that the everyday dataset resulted in higher accuracy compared to the other category. Considering the experimental result obtained by using the entire dataset, it was observed that at the training phases an accuracy of 100% was recognized. Classification is done where the hand gestures are captured and interpreted as commands. The models were tested with the larger lengthened gestures and the corresponding sign text word allotted to each gesture was also predicted. Thereby testing the entire dataset, an accuracy of 81% percentage was achieved and the correct text representation for the input gestures was predicted. Here, these signs patterns are focused and the system effectively converts sign language video sequence gestures into text form so the visual interaction in between deaf people and the bank is more convenient.

The performance of any detection system is analysed regarding its precision, recall, and F1- score, which are defined as Equations(1),(2) (3) below respectively.

$$Precision = TP/(TP + FP) \quad (1)$$

$$Recall = TP/(TP + FN) \quad (2)$$

$$F1\_score = (2 * Precision * Recall) \quad (3)$$

These measures are characterized by TP(True Positive),FP(False positive), followed by TN(True negative) and FN(False negative). Using these mentioned equations the classification reports are created and the performance of the system is determined. TABLE V shows the classification report of bank-related signs.

TABLE V

Signs	Precision	Recall	F1-score	Support
Debit Card	1	0.83	0.91	6
Balance sheet	0.71	0.1	0.83	5
Working Hours	1	1	1	5
Loan	0.75	1	0.86	5
Deposit	1	0.8	0.89	5
Expiry	1	1	1	5
Teller	1	1	1	5

V. Conclusion

With the recent advances in deep learning, there has been tremendous advancement in the domain of gesture recognition. Our study proves the effectiveness of using the deep neural network models CNN and LSTM for sign recognition making it easier for the deaf-mute people to communicate with banks. The main objective of our project was to build a human-computer interface for Indian sign language by using our self-customized dataset to tear away the communication obstacle between the deaf-dumb community and banks. The scope of sign language detection for banks is huge and there are multiple developments that can be done to our project in the future. Datasets can be increased and word sentences can be included for improving the efficiency of the system. The project can be completely automated and extended for text to speech conversion. Also, different models can be tried for the purpose of classification.

VI. Acknowledgment

We are grateful to our institution's Chancellor, Shri Dr. Mata Amritanandamayi Devi for the generous support and guidance extended towards us to achieve the successful outcome.

References

[1] Mrs.Dipali Rojasara and Dr.Nehal G Chitaliya;"Indian Sign Language Recognition –A Survey"; International Journal of Engineering Research and Technology (IJERT) Vol. 2 Issue 10, October - 2013 IJERT ISSN: 2278-0181.

[2] 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.

[3] Thomas Coogan , George Awad, Junwei Han and Alistair Sutherland, "Real time hand gesture recognition including hand segmentation and tracking", ISVC 2006 - 2nd International Symposium on Visual Computing, 6-8 November 2006, Lake Tahoe, NV, USA. ISBN 978- 3-540-48628-2.

[4] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4896-4899.

[5] Vivek Bheda and N. Dianna Radpour;"Using Deep Convolutional Networks for Gesture Recognition in American Sign Language" Department of Computer Science, Department of Linguistics State University of New York at Buffalo.

[6] Mrs.Dipali Rojasara and Dr.Nehal G Chitaliya;"Indian Sign Language Recognition –A Survey"; International Journal of Engineering Research and Technology (IJERT) Vol. 2 Issue 10, October - 2013 IJERT ISSN: 2278-0181

[7] Dr. M Geetha, PV Aswath; "Dynamic gesture recognition of Indian sign language considering local motion of hand using spatial location of Key Maximum Curvature Points",2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS).

[8] Dr. M Geetha, R Menon, S Jayan, R James, GVV Janardhan,"Gesture recognition for american sign language with polygon approximation",2011 IEEE International Conference on Technology for Education.

[9] Dr. M Geetha, C Manjusha, P Unnikrishnan, R Harikrishnan;"A vision based dynamic gesture recognition of indian sign language on kinect based depth images",2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA).

[10] Dr M Geetha, UC Manjusha;'A vision based recognition of indian sign language alphabets and numerals using b-spline approximation',International Journal on Computer Science and Engineering 4(3), 406

[11] Bhuyan, M. K., Ghoah, D. and Bora, P. K. "A Framework for Hand Gesture Recognition with Applications to Sign Language", IEEE Annual India Conference, Sept 2006

[12] Nandy, A., Mondal, S., Prasad, J. S., Chakraborty, P., and Nandi, G. C. "Recognizing Interpreting Indian Sign Language Gesture for Human Robot Interaction", In the proceedings of Int'l Conf. on Computer Communication Technology(ICCCT'10)

.