DEEP LEARNING BASED FRAMEWORK FOR DYNAMIC BABY SIGN LANGUAGE RECOGNITIONSYSTEM

Sulochana Nadgeri

Research scholar, Computer Science and Engineering, Sir Padampat Singhania University,
Udaipur, 313601 Rajasthan, India
Sulchana.nadgeri@spsu.ac.in

Dr.Arun Kumar

Dean, School of Engineering, Sir Padampat Singhania University,
Udaipur, 313601,
Rajasthan, India
arun.kumar@spsu.ac.in

Abstract

In spite of Sign language being widely used by the Deaf and Hard of Hearing communities, this paper illustrates a different application of sign language, namely Baby Sign Language used to communicate with hearing infants and toddlers by hearing parents. To accomplish continuous sign recognition, we propose a model based on convolutional neural networks (CNN) and long short-term memory networks (LSTM). Using CNN, videos' information can be translated into vectors. The LSTM model is employed toconnect with the fully-connected layer of CNN, as the video can be viewed as an ordered sequence of 9 frames. The method is evaluated on a dataset that includes 34 daily vocabularies that we built ourselves. CNN-LSTM demonstrates a high recognition rate with minor changes in the optimizer hyperparameter from SGD to Adam. SGD gives a 75% accuracy, while the Adam optimizer gives 99% accuracy.

***Keywords*: Baby Sign Language; Deep Learning; CNN-LSTM.**

1. Introduction

Through communication, we can convey our thoughts and feelings at the same time, enabling us to understand what others are feeling and thinking. There are two ways in which we communicate: verbally and nonverbally. Communication through verbal means involves speaking to others; however non-verbal means expressing ourselves through facial expressions, gestures, posture, and hand movements. Those deaf or hard of hearing people are most likely to use sign language to express their emotions, thoughts, and ideas, but different countrieshave different Sign Languages similar to spoken languages. There are somewhere between 138 and 300 different types of sign language used around the globe today [https://www.ai-media.tv]. The most common sign language in the world is American Sign Language (ASL). The use of sign language helps deaf and hard of hearing people communicate and children with disorders such as autism and aphasia. Gestures, posture, eye gaze, and facial expressions all play a vital role in Sign Language. The most common way of learning sign language is by learning the A-Z alphabet in sign form. Fingerspelling is the use of hands to represent individual letters of a written alphabet. As shown in Fig. 1, finger spellings represent the alphabet A-Z in American Sign Language.

Fig.1 Fingerspelling for alphabet A-Z in American Sign Language for words (courtesy: https://www.ai-media.tv/)

Hearing parents use sign language to enhance communication with their hearing babies (aged between 6months and 6-8 years). Alternatively, it is called Baby Sign Language. Infants begin to express facial expressions to their caregivers at a young age, facilitating their interaction with them. Infants communicate primarily through their eyes and vocalizations such as cooing and crying. Caretakers must often rely on contextual cues to determine the appropriate response to crying, despite its effectiveness at eliciting various caregiving responses [Yale *et al*., (2003)]. Baby signs differ from signs used by other deaf or hard of hearing people. A family can establish a bond between themselves and their infant by using baby sign language, a form of gesture-based communication and verbal communication. This early communication helps children learn more quickly, reduces frustration, and strengthens the parent-child relationship. According to Melissa J. Cesafsky, Baby Sign Language is a way for you and your infant (or toddler) to communicate by using specific hand shapes, gestures, and motions. Parents and infants of specific hearing ability use Baby Sign Language. In contrast to the sign language used by the deaf, Baby sign includes keywords and does not have complex linguistic structures since infants will use it. A recent study conducted by Professor Gwyneth Doherty-Sneddon of Stirling University, UK, recently confirmed that baby signing enhanced a child's vocabulary and mental development, reduced tantrums and improved relationships with parents, especially as communication is the foundation of a child's development. [Donoghue *et al*.,(2003)]. Fig 2 shows a chart of Baby Sign Language.

Fig. 2 Sample Baby Sign Language gestures for words (source: babysignlanguage.com)

In addition to psychological benefits, babies who are taught to use gestures will experience increased certainty and confidence; for example, if they are taught to use gestures to communicate, an inability to communicate may lead to fewer feelings of frustration. In the case of a child who is excessively agitated to talk in a certain way, his or her ability to sign may be of assistance. Teaching sign language in conjunction with speech has confirmed the development of spoken communication abilities—infants who use gesture plus speech combinations to form a sign. When a baby learns sign language, it does not slow down his/her verbal development; rather, it enhances the child's desire to learn more communication techniques, including talking.

2. Related Work

Due to the lack of a universal sign language, researchers attempt to identify the sign language with either an existing dataset or build their database. The datasets they use can be static, dynamic, or both.

In [Pigou *et al*.,(2014)] identifies 20 Italian gestures using a system that uses Microsoft Kinect and convolutional neural networks (CNNs) as well as GPU acceleration, giving it an accuracy of 91.7%. They used the CLAP14 dataset that is made up of Microsoft Kinect data.

Using a convolutional neural network (CNN), [Sarkar *et al*., (2019)] found that they could recognize 26 alphabets in Indian Sign Language. They tested CNN models on two different datasets that they created and got 99.40% accuracy when the background was static and only hand gestures were inputted.

Based on Transfer Learning,[Rathi, (2018)] developed an ASL recognition system for mobile platforms, which recognized 24 alphabets from the MNIST and provided an accuracy of 95.03 %.
The authors [Morocho *et al*.,(2019)] present a real-time American Sign Language (ASL) hand gesture recognizer. It uses pre-trained Convolutional Neural Networks (CNNs) to train a dataset, which for GoogleNet gave accuracy of 95.52%, and for AlexNet, it gave an accuracy of 99.39%.

The authors [Bhagat *et al*.,(2019)] proposed a two-dimensional (3D) construction of hand gestures based on the Microsoft Kinect RGB-D camera, along with affine transformation techniques to capture the three-dimensional (3D) datasets. With 45,000 RGB images and 45,000 depth images, a Convolutional Neural Network (CNN) model successfully trained 36 static gestures and achieved an accuracy of 98.81%.

The authors [Das *et al*.,(2018)] created a method based on the vision to classify dynamic ASL performed under different lighting conditions based on an algorithm that interprets a sign from video into text using deep learning. In the 600-sample dataset, the CNN controls the spatial content, and the LSTM RNN

controlsthe temporal content. The system provides 99 % accuracy.

In [Tao *et al*., (2018)], a CNN-SqueezeNet model is proposed to recognize American Sign Language letters from RGB images. Images obtained from the dataset are resized and pre-processed before Deep Neural Network models are applied. Based on the accuracy of 87.47% in training and 83.29% in validation, the model can run on mobile devices.

In [Nadgeri *et al*., (2019)], the author presented a machine learning-based classification method that used Gray Level Co-occurrence Matrix(GLCM) to extract texture-based features and used KNN and Random Forest algorithm to classify a dataset of 60 Baby signs.On the prepared dataset, the system's accuracy is 73%.

Using a Kinect sensor, [Yeh *et al*., (2016)] detects the palm with the Otsu thresholding method and morphology operators based on skeleton data, color data, and depth data. By using SVM, the 20 alphabets and numerals in English represented in sign language are predicted with a recognition accuracy rate of 70.59%

An image intensity and depth data-driven convolutional network [Ameen *et al*.,(2016)] were developed to recognize the 24 fingers of the alphabet in American Sign Language. The first block of the proposed architecture is classified into two parts: one that extracts the edges of RGB images, and another that extracts the edges of depth images. . The next step is to combine the features. The evaluation found that the developed convolutional network performed better than previous studies with an 82% precision and 80% recall.

[Cui *et al*., (2017)] proposed an RNN-based continuous sign language recognition method using LSTMs on an RWTH-PHOENIX-Weather multi-signer benchmark dataset, using convolutional neural networks and video sequences to recognize continuous sign languages.

A framework of hierarchy-awareness networks with latent space (HAN-LS) has been proposed by [Huang *et al*., (2018)], which recognizes signs in the German sign language dataset RWTH-PHOENIX-Weather by generating global-local video features.

Using a new data-level blending method, [Köpüklü et al., (2018)] show Motion Fused Frames (MFFs) that do not use each frame in all video segments but rely on color and optical flow modalities. Based on an analysis of three available databases, the system showed improved performance in all cases. However, this method was difficult to implement globally due to the inability to turn various hand motion information into static images.

In [Liu *et al*., (2017)] presented a spotting-recognition framework for recognizing gestures at scale with an RGB-D input. The continuous gestures are initially segmented into isolated gestures by getting the exact hand position through Faster R-CNN A hand-specific spatiotemporal (ST) feature is extracted by 3D convolutional networks (C3D), where only the hand regions and the face location are examined, thus blocking the adverse effects of visual stimuli such as the background, cloth, and body. The system is 61.03 % accurate.

An innovative approach to deep learning with SubUNets, a novel architecture that learns intermediary representations to guide the learning operation, has been proposed by [Camgoz *et al*., (2017)]. Based on the One Million Hands dataset, the proposed method trained and evaluated SubUNet to recognize hand shapes and reported an improvement of 30% over earlier methods.

In continuous Indian sign language, [Kumar *et al*., (2016)] captured 18 signs using selfie sticks with 10 different signers. Based on discrete cosine transforms, the contour features of the hands and heads were reduced with principal component analysis [G. Rao *et al*.,(2018)]. Then, the system classifies the signs based on Euclidian normalized Euclidean and Mahalanobis distance metrics. The Mahalanobis distance consistently produces an average word matching score of 90.58 % for the identical train and test sets compared to the two other distance measures, although the dataset size causes concern. The use of the Fuzzy Inference Engine further improved the performance, with an average word matching of 92.5% [PVV Kishore *et al*., (2016)]

[Rao *et al*.,(2017)] recognized continuous sign language using a neural network classifier. Dynamic sign language recognition was performed using a neural network classifier based on deep learning approaches, such as

convolutional neural networks (CNNs) in [A. Krizhevsky *et al*., (2017)]. A combination of CNN- and RNN-based approaches to Spatio-temporal learning has been proposed by [N.Srivastava *et al*.,(2015)] and [M.Baccouche *et.al*., (2018)] for the learning of video sequences.

For skeleton-based action recognition, [Du *et al*.,(2015)] proposed a hierarchical RNN. They divided the entire skeleton into five subnets to achieve state-of-the-art performance and then fed them into the simulation. The resulting model demonstrates high computational efficiency, using 5 subnets to process the whole skeleton.

A study by [Norah *et al*.,(2019)] examined seven commonly used dynamic hand gestures captured by mobile cameras. The videos were captured 20 times per gesture, resulting in 24,698 frames per video. Features were extracted from the images, and the deep convolutional neural network (ADCNN) was used to classify the hands. Based on training data and testing data, the system was evaluated at 100% and 99.73%, respectively, with a duration of 15,598 seconds. The ADCNN model outperformed Base CNN by 4%. There are limited gestures in the system; instead, signs are based on single-handed signs.

[Hung *et al*.,(2019)] proposed a system that uses webcams to track a user's hand and control appliances. After doing so, they then used YCbCr color space techniques for morphology and skin color to eliminate the background. The ROI was also tracked using kernel correlation filters (KCF). After processing the image, it was run through Alex Net and VGG Net, a variant of deep convolutional neural networks (CNN). In this study, the performance of both Alex Net and VGGNet models were compared. The VGGNet model gives recognition rates of 99.90% and 95.61% for training data and testing data, respectively, whereas Alex Net gives 99.68% and 84.99% respectively for training and testing data.

[Zhan,(2019)] presented a system that recognizes 9 hand gestures dynamically. Because the dataset size is relatively small, Spatio-temporal data augmentation is done by horizontally mirroring the images to avoid overfitting. A total of 4500 images were added to the dataset after augmentation.SGD optimizer was used to train the CNN using an incremental learning rate. In terms of classification accuracy, CNN classifiers achieve an average of 98.76%.

[Chou *et al*.,(2020)] have developed an American Sign Language wearable decoder that identifies 156 characteristics from the acquired sign language data to classify the words. Using the long short-term memory (LSTM) method, this system achieved up to 99.89 % accuracy.

Using the Microsoft Kinect sensor, [Gangarde *et al*.,(2020)] detected and separated a hand region from a depth picture. The proposed approach works well in cluttered environments, such as those with dark backgrounds and an overlapping hand. Using convolutional neural networks (CNNs), features from Indian signs are automatically generated. These characteristics remain unchanged regardless of rotation or scaling. With the method proposed, gestures can be correctly classified with an accuracy of 99.3%.

We propose a model for dynamic Baby Sign Recognition, a process that takes into account more than hand gesture alone, and also comprises hand occlusion with the hand, hand occlusion with the face, in addition to the non-constant signer's orientation as well as occlusion of the hand with upper body parts.

3. Methodology:

Because the standard dataset is not accessible for research, it is necessary to create your dataset. There are several key terms in baby sign language that babies and infants will use to express their feelings and needs. Therefore, it does not contain signs for the alphabet or numbers like other sign languages. In addition, there are fewer complexes in baby signs because it does not follow linguistic structure, and signs are made up of one or more than one gesture. So while creating the dataset for research, static gestures are acquired independently thenmerged in order to represent a word from Baby Sign Language.

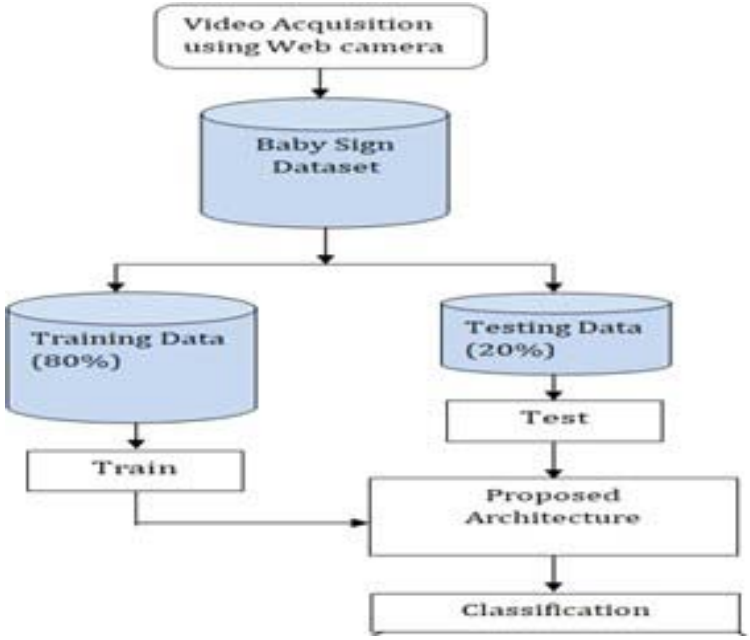

Fig. 3 Proposed Architecture

3.1 Dataset preparation/Acquisition:

The dataset is premised on the following assumptions:

(1) The web camera has a resolution of 720p.
(2) A Web camera is used to capture the three-channel RGB image
(3) There is no need for gloves or Kinect sensors
(4) There were no other people in view during the signing process.
(5) The signer stood approximately 0.5 meters from the Web camera.
(6) The hand gestures were captured under indoor lighting conditions.
(7) Signers wore black t-shirts during the signing process.
(8) The signer's upper body was visible in the video.
(9) The video was captured for 4 seconds.

A dataset of 34 Baby sign language gestures has been created for this experiment. Each sign is repeated on 11 times with different occasions, and just one signer is considered. Videos have been stored in respective folders and named with each sign's name.

3.2 Training and Testing:

There are 80:20 proportions for the training and test (i.e. validation) sets of dataset. Lastly, the training set is divided into batches, and the images are shuffled randomly for rearrangement.

3.3 Network Architecture:

Fig 4. illustrates a proposed network architecture for deep learning combining input layer, LSTMs, time-distributed layer, convolution layer, flatten layer, dense layer, and dropout layer.

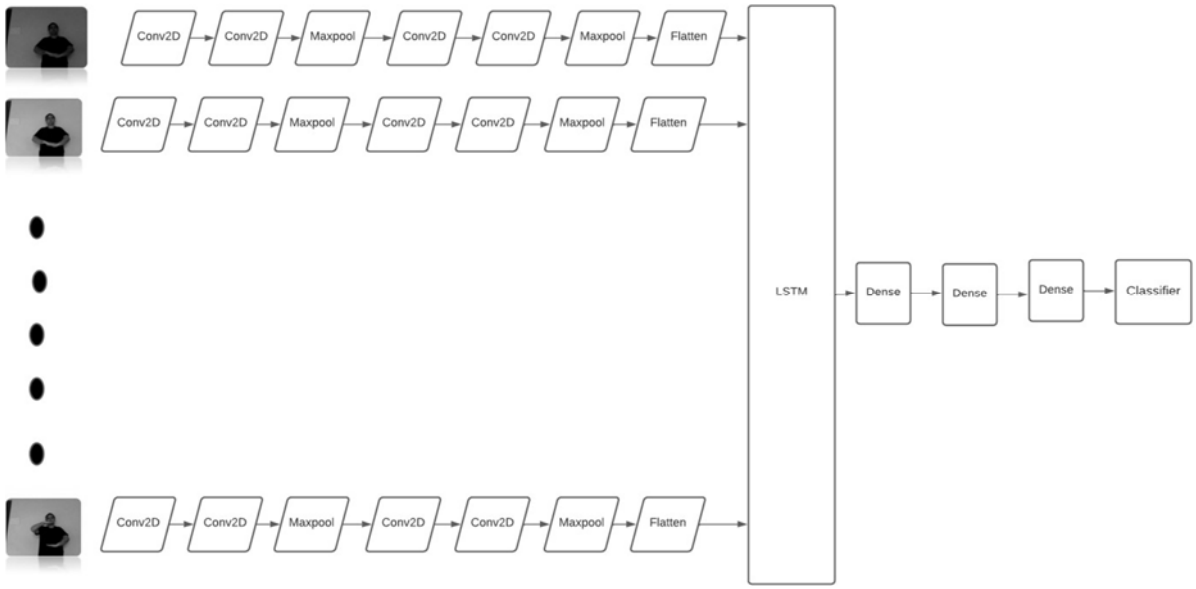

Fig 4. Customized Network architecture

As shown in Model architecture, the Time distributed layer plays a key role as we desire to apply the identical operation to a sequence of frames.

3.3.1 Input Layer: Every video is converted into 9 RGB frames of 64x64 dimensions. Each layerwill therefore receive 9 input images. We need the same filtering for each image, and we want to continue with common layers.

3.3.2 Time Distributed Layer: Using this wrapper, we can apply a layer to every temporal slice of an input. Instead of managing many inputs, we can manage one model for every input. It implements the same transformation for a list of inputs. Input data can be transformed using a combination of convolutions, pooling, densification, batch normalization, etc. All 9 images in the input will be subjected to the stated transformations.

3.3.3.Convolution Layer: The proposed architecture uses 4 convolutional layers. There are different numbers and sizes of the convolution kernels for each layer. Therefore, the system uses more kernels for deeper layers, thereby capturing more features. Specifically, we used 128-64-64-32 convolution kernels sequentially, and the 3x3 kernel size was utilized for each layer. As a result, the output feature map size is maintained by using the zero-padding method, i.e., by padding the original image with zeros in order to reduce the impact of the image edge and maintain the original image size. The rectified linear unit (ReLU) activation function is subsequently applied to each layer of the convolutional layer [H.-Y. Chung *et al.,(* 2019)].

3.3.4. Dropout Layer: Regularization by dropout lets us approximate parallel training of many neural networks with various architectures. Specific layer outputs are dropped out of the training at random. We used a dropout value 0.2 for the proposed model.

3.3.5 .Max pooling Layer: Pooling layers reduce the size of the feature maps. Reducing the number of parameters to learn and computation done in the network reduces the amount of learning. Convolutional layers produce feature maps whose areas are summarized by the pooling layer. Therefore, subsequent operations on theresulting features are based on summarized information rather than precisely positioned information generated by the convolution layer. Therefore, the model can better cope with changes in the location of features in the input image. The largest element from the region of the feature map covered by the filter is picked in a max pooling operation. As a result, the output of the max-pooling layer would be a feature map that contained the most prominent features of the prior feature map.

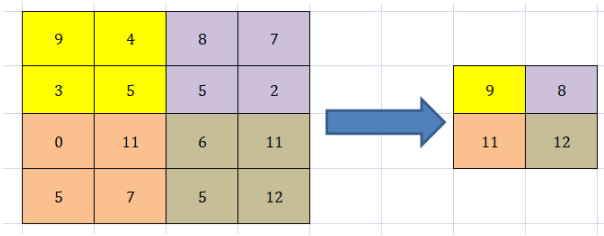

Fig. 5. Max pooling of a single depth slice with a 2x2 filter and two stride.

3.3.6. Dense Layer: In the network, we employed three dense layers. The dense layer is a strongly connected neural network layer, which means that each neuron in the dense layer receives input from all neurons in the layer preceding it. The dense layer was the most often used in the models. In the background, the dense layer performs matrix-vector multiplication. The values in the matrix are actual parameters that may be learned and altered via backpropagation. As an output, the dense layer generates the 'm' dimensional vector. As a result, the dense layer is typically used to change the dimensions of the vector. Vector operations such as rotation, scaling, and translation are also performed by dense layers.

3.3.7. Classifier: The final layer is a sigmoid function-based 34-node layer, with each node representing a class in the input dataset. The sigmoid utilizes a probabilistic strategy for decision making that ranges from 0 to 1, thus when we need to make a choice or foresee an event, we choose this activation function since the range is the shortest, resulting in more accurate prediction.

4. Results & Discussion

4.1 Experimental Setup

(1) The dataset was split into 80% and 20% for training and testing in the experiment, respectively. The proposed network consists of 14 layers, and details are described in Fig. 4. The CNN and CNN-LSTM networks were implemented using Python and the Keras package with tensorflow2 on 90GHZ CPU frequency, 8 GB RAM, and GPU with NVIDIA geforce®GTX730M graphics card for speedy operations and run several times to finalize the proper hyper-parameters viz. Batch size, number of epochs, number of iterations, validation frequency.

(2) The model is trained with SGD optimizer, categorical-cross entropy loss function, L2 Regularizer to reduce overfitting and initial learning rate = 0.00001, but it decreases during training when Validation_loss becomes stagnant by a factor of 0.02.

(3) Fit and train the data for mini batch_size = 9 and epochs = 50

(4) Verify the system on validation data and calculate the training and test accuracy,

(5) Training and test loss, and confusion matrix with performance parameters like precision, recall, F1-score, and support

4.2. Performance Evaluation

The experimental results of Baby Sign Language Recognition are presented here in this section. Fig.6 and Fig. 7 show model loss and accuracy graphs. As we can also see from the plot of accuracy, the model has not yet over-learned the training data and shows similar performance on the validation data.

From the plot of loss, we can see that the model has comparable performance on both train and validation datasets (labeled test). If these parallel plots start to depart consistently, it might be a sign to stop training at an earlier epoch. The plot shows a good fit , and is identified by a training and validation loss that decreases to a point of stability with a minimal gap between the two final loss values. After almost 30 epochs, training and validation losses decrease gradually.

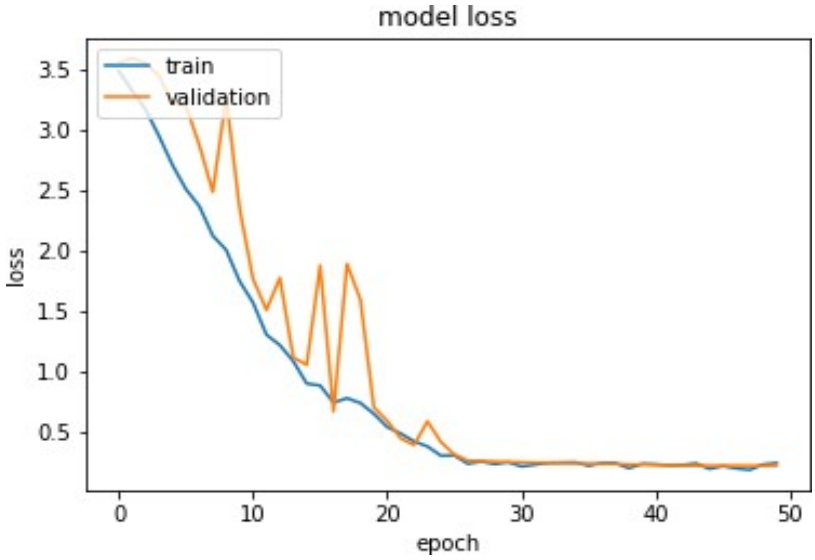

Fig 6 Model Loss during Training and Validation

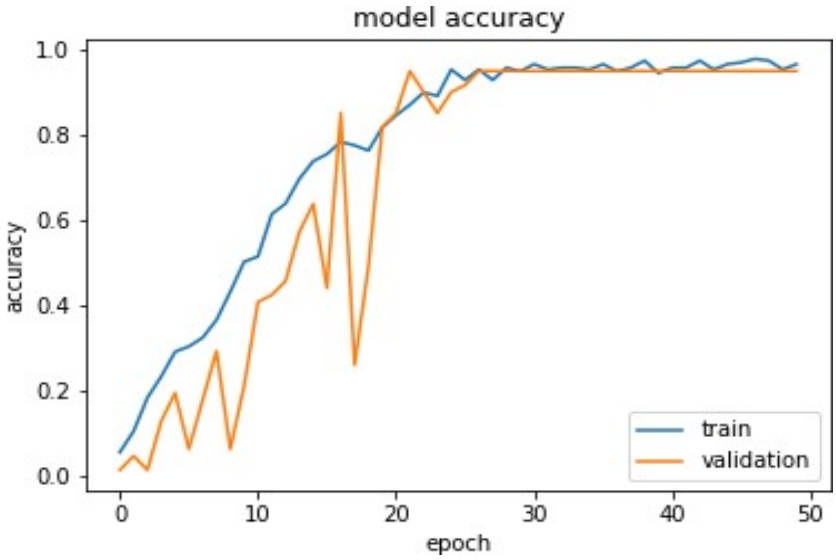

Fig 7 Model Accuracy during Training and Validation

For measuring the performance level, we use some sort of statistical tool. Firstly we used accuracy and loss So, we evaluate some performance measures like F1-score, precision, recall, FPR, and FDR based on confusion matrix illustrated in Table 1 and Fig 7, the confusion matrix that determines the efficacy of the proposed network across the test dataset where diagonal epitomizes correct results and from the figures, it canbe observe that most of results predicted correctly.

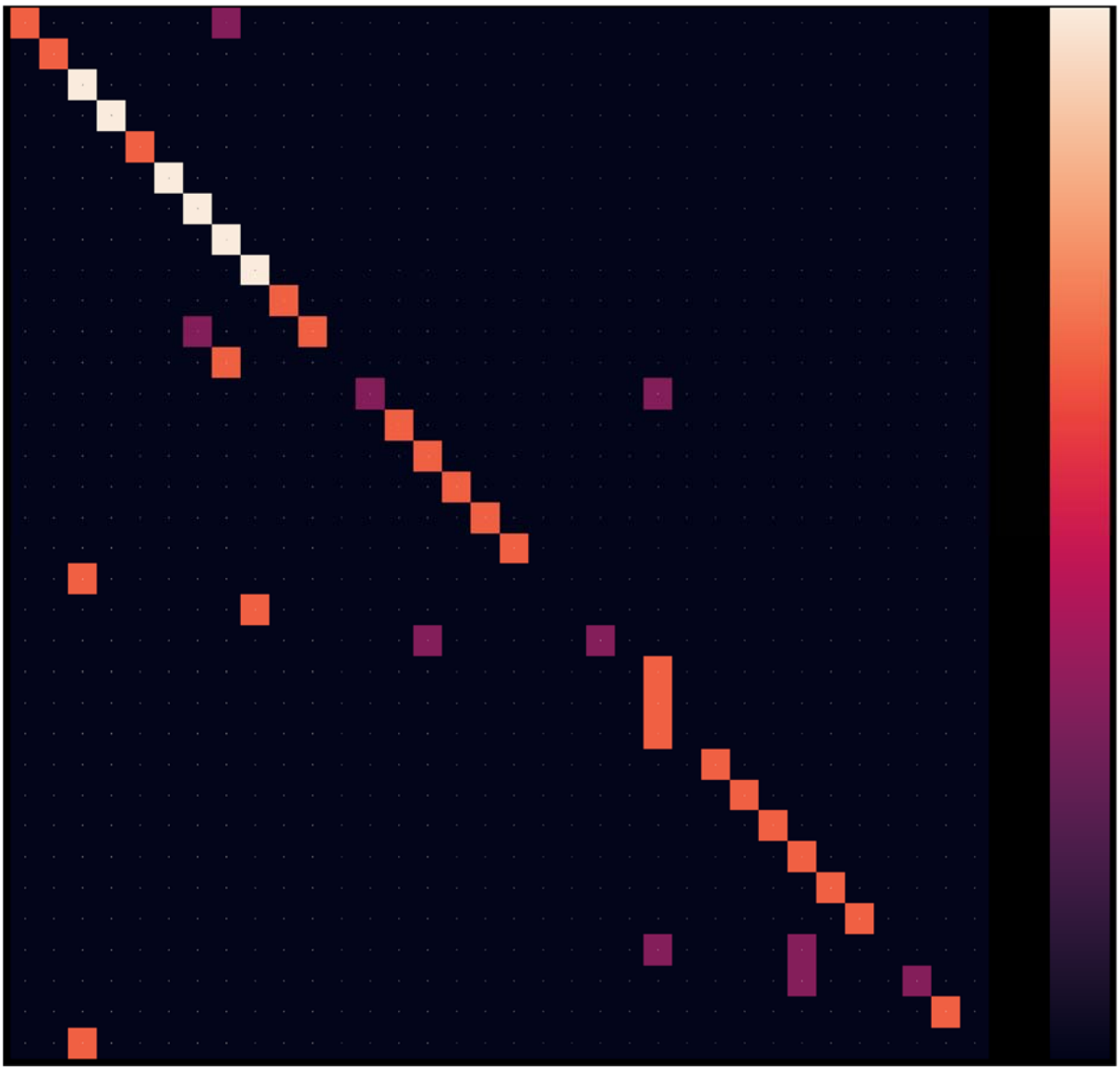

Fig. 8 Confusion Matrix

True Positive (TP) refers to the number of predictions where the classifier correctly predicts the positive class as positive.
True Negative (TN) refers to the number of predictions where the classifier correctly predicts the negative class as negative.
False Positive (FP): It refers to the number of predictions where the classifier incorrectly predicts the negative class as positive.
False Negative (FN) refers to the number of predictions where the classifier incorrectly predicts the positive class as negative.

$$\text{Accuracy} = (TP + TN)/(TN + FP + TP + FN) \quad (1)$$

$$\text{Recall} = TP/(TP+FN). \quad (2)$$

$$\text{Precision} = TP/(TP+FP). \quad (3)$$

$$F1 - \text{score} = (2*TP)/(2*TP + FP + FN) \quad (4)$$

Class Name	Precision	Recall	F1-Score	Support
Above	1.00	0.67	0.80	3
Adult	1.00	1.00	1.00	2
Afraid	0.43	1.00	0.60	2
Airplane	1.00	1.00	1.00	3
All Done	1.00	1.00	1.00	2
Alligator	1.00	1.00	1.00	3
Ambulance	0.75	1.00	0.86	3
Angry	0.50	1.00	0.67	3
Animal	0.60	1.00	0.75	3
Art	1.00	1.00	1.00	2
Aunt	1.00	0.67	0.80	3
Baby	0.00	0.00	0.00	2
Back	1.00	0.50	0.67	2
backpack	1.00	1.00	1.00	2
ball	0.67	1.00	0.80	2
banana	1.00	1.00	1.00	2
birthday	1.00	1.00	1.00	2
book	1.00	1.00	1.00	2
bye-bye	0.00	0.00	0.00	2
cake	0.00	0.00	0.00	2
candy	1.00	0.50	0.67	2
cat	0.00	0.00	0.00	2
cheese	0.25	1.00	0.40	2
chocolate	0.00	0.00	0.00	2
cookie	1.00	1.00	1.00	2
cousin	1.00	1.00	1.00	2
cry	1.00	1.00	1.00	2
daddy	0.50	1.00	0.67	2
dog	1.00	1.00	1.00	2
dont	1.00	1.00	1.00	2
dont hit	0.00	0.00	0.00	2
dont want	1.00	0.50	0.67	2
door	1.00	1.00	1.00	2
dress	0.00	0.00	0.00	2
accuracy		0.75	76	
macro avg	0.70	0.73	0.69	76
weighted avg	0.71	0.75	0.70	76

Table1 Class Wise performance evaluation

Further, the proposed system was analyzed using SGD and Adam optimizers which gave a comparatively better result. The following table shows the precision in percentage of each optimizer class-wise. We can easily observe Adam Optimizer classify every sign correctly except Cat Baby Sign.

Sr. No.	Baby Sign	Optimizer	
		SGD	Adam
1	Above	100	100
2	Adult	100	100
3	Afraid	42.86	100
4	Airplane	100	100
5	All Done	100	100
6	Alligator	100	100
7	Ambulance	75	100
8	Angry	50	100
9	Animal	60	100
10	Art	100	100
11	Aunt	100	100
12	Baby	33.33	100
13	Back	0	100
14	Backpack	100	100
15	Ball	66.67	100
16	Banana	100	100
17	Birthday	100	100
18	Book	100	100
19	Bye-Bye	0	100
20	Cake	0	100
21	Candy	100	100
22	Cat	0	66.67
23	Cheese	25	100
24	Chocolate	0	100
25	Cookie	100	100
26	Cousin	100	100
27	Cry	100	100
28	Daddy	50	100
29	Dog	100	100
30	Don’t	100	100
31	Don’t Hit	0	100
32	Don’t Want	100	100
33	Door	100	100
34	Dress	0	100

Table 2 Comparative result of SGD and Adam optimizer class wise

The analysis shows that the Adam optimizer gives a better result.

4.3. Conclusion and Future work

A vision-based Dynamic Baby Sign language recognition system is devised and built for 34 random gestures for words used to bond children with their parents. The suggested technique employs a neural-based deep model based on the CNN-LSTM model. In the current study, our model classified the signs with 99% accuracy & 75% accuracy using Adam and SGD optimizer respectively.

It is possible to improve the dataset by getting more samples that take into account lighting conditions, increasing the number of signers to increase variety in the dataset, and decreasing the distance between signers in future.

References

[1] A. Das, S. Gawde, K. Suratwala and D. Kalbande (2018): Sign Language Recognition Using Deep Learning on Custom Processed Static Gesture Images , International Conference on Smart City and Emerging Technology (ICSCET), Mumbai,.

[2] A. Krizhevsky, I. Sutskever, and G. E.Hinton (2012): Imagenet classification with deep convolutional neural networks, in *Proc. 25th Int. Conf. Neural Inform. Process. Syst.,* Lake Tahoe, Nevada,Dec. 2012, pp.1097-1105.

[3] Alnaim N., Abbod M., Albar A(2019): Hand Gesture Recognition Using Convolutional Neural Network for People Who Have Experienced A Stroke; Proceedings of the 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT); Ankara, Turkey. 11–13 October 2019; pp. 1–6.

[4] Ameen, Salem & Vadera, Sunil(2017): A convolutional neural network to classify American Sign Language fingerspelling from depth and colour images., Expert Systems.

[5] Baccouche M., Mamalet F., Wolf C., Garcia C., Baskurt A. (2011) : Sequential Deep Learning for Human Action Recognition. In: Salah A.A., Lepri B. (eds) Human Behavior Understanding. HBU 2011. Lecture Notes in Computer Science, vol 7065. Springer, Berlin, Heidelberg.

[6] Chong, Teak-Wei & Kim, Beom-Joon. (2020): American Sign Language Recognition System Using Wearable Sensors with Deep Learning Approach. The Journal of the Korea institute of electronic communication sciences. 15. 291-298.

[7] D. A. Kumar, P. V. V. Kishore, A. S. C. S. Sastry and P. R. G. Swamy(2016): Selfie continuous sign language recognition using neural network, 2016 IEEE Annual India Conference (INDICON), pp. 1-6, doi: 10.1109/INDICON.2016.7839069.

[8] Donoghue, Ellen C.(2014): Sign Language and Early Childhood Development, Rehabilitation, Human Resources and Communication Disorders Undergraduate Honors Theses. 20

[9] G. Ananth Rao, P.V.V. Kishore (2018):Selfie video based continuous Indian sign language recognition system,Ain Shams Engineering Journal,Volume 9, Issue 4,Pages 1929-1939

[10] Gangrade, Jayesh & Bharti, Jyoti. (2020):Vision-based Hand Gesture Recognition for Indian Sign Language Using Convolution Neural Network. IETE Journal of Research. 1-10.

[11] Gondu Anantha & Kishore, P.V.V. & Anil Kumar, D. & Sastry, A.. (2017): Neural network classifier for continuous sign language recognition with selfie video. Far East Journal of Electronics and Communications. 17. 49-71.

[12] H. -Y. Chung, Y. -L. Chung and W. -F. Tsai(2019): An Efficient Hand Gesture Recognition System Based on Deep CNN, 2019 IEEE International Conference on Industrial Technology (ICIT), 2019, pp. 853-858.

[13] J. Donahue., Long-term recurrent convolutional networks for visual recognition and description, *IEEE Trans. Pattern Anal. Mach.Intell. (PAMI)*, vol. 39, no. 4, pp. 677-691

[14] J. Huang, W. Zhou, Q. Zhang, H. Li, and W. Li, Video-based sign language recognition without temporal segmentation, in 32nd AAAI Conf. Artificial Intelligence (AAAI-18), New Orleans, Louisiana, USA, Feb. 2018.

[15] Morocho-Cayamcela, Manuel Eugenio & Lim, Wansu., Fine-tuning a pre-trained Convolutional Neural Network Model to translate American Sign Language in Real-time. 2019 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, pp. 100-104

[16] Mou, L., Bruzzone, L., & Zhu, X. X. (2018). Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 57(2), 924-935.

[17] N. C. Camgoz *et al.,* SubUNets: end-to-end hand shape and continuous sign language recognition, in *IEEE Int. Conf. Comput.Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3075-3084.

[18] N. K. Bhagat, Y. Vishnusai and G. N. Rathna, Indian Sign Language Gesture Recognition using Image Processing and Deep Learning,, Digital Image Computing: Techniques and Applications (DICTA), Perth, Australia, pp. 1-8

[19] N. Srivastava, E. Mansimov, and R. Salakhudinov, Unsupervised learning of video representations using lstms, in *Pro. 32th Int. Conf. Mach. Learn.*, Lille, France, Jul. 2015, pp. 843-852.

[20] Ng, Joe & Hausknecht, Matthew & Vijayanarasimhan, Sudheendra & Vinyals, Oriol & Monga, Rajat & Toderici, George. (2015). Beyond short snippets: Deep networks for video classification. 4694-4702.

[21] O. Köpüklü, N. Köse, and G. Rigoll(2018): Motion fused frames: data level fusion strategy for hand gesture recognition, in *2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, 2018, pp. 2184-21848.
[22] P. V. V. Kishore, D. A. Kumar, Goutham E.N.D and M. Manikanta(2016): Continuous sign language recognition from tracking and shape features using Fuzzy Inference Engine, 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2016, pp. 2165-2170
[23] Pigou L., Dieleman S., Kindermans PJ., Schrauwen B.(2014): Sign Language Recognition Using Convolutional Neural Networks. In: Agapito L., Bronstein M., Rother C. (eds) Computer Vision - ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science, 8925: 572-578
[24] R. Cui, H. Liu, and C. Zhang(2017): Recurrent convolutional neural networks for continuous sign language recognition by staged optimization, in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 7361-7369
[25] Rathi, Dhruv(2018): Optimization of Transfer Learning for Sign Language Recognition Targeting Mobile Platform. International Journal on Recent and Innovation Trends in Computing and Communication, 6(4): 198 -03
[26] S. Nadgeri and A. Kumar(2019): An Image Texture based approach in understanding and classifying Baby Sign Language, 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, Kerala, India, pp. 854-858
[27] Sarkar A., Talukdar A.K., Sarma K.K.(2019): CNN-Based Real-Time Indian Sign Language Recognition System., In Chillarige R.,Distefano S., Rawat S. (eds) Advances in Computational Intelligence and Informatics. ICACII 2019. Lecture Notes in Networks and Systems, Singapore. 119: 71-79
[28] Srivastava, N., Mansimov, E. & Salakhudinov, R.. (2015): Unsupervised Learning of Video Representations using LSTMs. Proceedings of the 32nd International Conference on Machine Learning in Proceedings of Machine Learning Research 37:843-852
[29] Tao, Wenjin & Leu, Ming & Yin, Zhaozheng(2018): American Sign Language Alphabet Recognition Using Convolutional Neural Networks with Multiview Augmentation and Inference Fusion., Engineering Applications of Artificial Intelligence , 76: 202-213
[30] W. Yeh, T. Tseng, J. Hsieh and C. Tsai(2016): Sign language recognition system via Kinect: Number and english alphabet, International Conference on Machine Learning and Cybernetics (ICMLC), Jeju, pp. 660-665, 2016.
[31] Yale, M. E., Messinger, D. S., Cobo-Lewis, A. B., & Delgado, C. F. (2003): The temporal coordination of early infant communication. Developmental Psychology, 39(5), 815–824.
[32] Yong Du, W. Wang and L. Wang(2015): Hierarchical recurrent neural network for skeleton based action recognition, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1110-1118
[33] Z. Liu, X. Chai, Z. Liu, and X. Chen(2017): Continuous gesture recognition with hand-rienting spatiotemporal feature, in *Proc. IEEE Conf. Comput. Vis. Workshops*, Venice, Italy, Oct. 2017, pp.3056-3064.
[34] Zhan Felix. (2019): Hand Gesture Recognition with Convolution Neural Networks. 295-298.
[35] Sign Language Alphabets From Around The World Ai-Mediahttps://www.ai-media.tv/ai-media-blog/sign-language-alphabets-from-around-the-world/

Authors Profile

Sulochana Nadgeri, Research Scholar in the Department of Computer Science and Engineering at Sir Padampat Singhania University, Udaipur, Rajasthan. She completed Bachelor's degree in Computer Science and Engineering from Shivaji University Maharashtra and Master's degree from University of Mumbai. Her areas of interest are Image Processing, Artificial Intelligence, Machine Learning and Deep Learning. She got Minor research grant from University of Mumbai. She got 16 year teaching experience.

Prof Arun Kumar, presently working as Professor in the Department of Computer Science and Engineering at Sir Padampat Singhania University, Udaipur, Rajasthan. He is also shouldering the responsibility of Dean of the School of Engineering at SPSU. He holds a Bachelor's degree in Applied Electronics and Instrumentation Engineering from NIT Rourkela,a master's degree in Computer Science and Engineering from University of Madras and a Doctoral degree in the area of Computer Vision. He has interest in the development of applications in the area of Data Science, Recommender Systems, and Fake News Analysis. He holds two granted patents from the Govt of India. He is a registered PhD guide with SPSU and has four research scholars who have already graduated from the department of Computer Science and Engineering.