**Chapter**

**SIGN LANGUAGE DETECTION USING MODIFIED ALEXNET ARCHITECTURE**

***Dr Sornavalli Ganesan Ph.D,***

Department of Information Technology,

***R Goutham,***

Department of Electrical and Electronics Engineering,

***Aditya Jyosyula***

Department of Computer Science Engineering,

Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamilnadu, India

**ABSTRACT**

Hand Gestures are the prominent ways in which the hearing and speech impaired use to communicate with one another. It is a mode of communication which bridges the gap between the abled and disabled and thus need the modern implementation of Hand gesture recognition by computer systems. This Human Computer Interaction should be error free and classify every character correctly. We use a Deep Learning Convolutional Neural Network for robust character recognition with static images. We use a CNN for hand character recognition for the American Sign Language (ASL) dataset considering both alphabets and numerals. The CNN architecture employed is modified AlexNet architecture which is tailored for this specific dataset and needs. The CNN model is evaluated based on its features for high accuracy and performance. The training is done using GOOGLE COLAB CPU/GPU session to emphasize accessibility and verifiability by reviewers. The proposed system has achieved training accuracy 94.43% and validation accuracy 98.09% .

**KEYWORDS**

Sign language recognition, CNN, ASL, Hand gesture recognition.

**INTRODUCTION**

As technology advances, computation ability has greatly increased and become cheaper. Hand gestures have been in use since the dawn of human civilization. Sign Language was developed to enable the deaf and mute to communicate with the general population. This sign language needs to be recognized by computer systems for many applications. Vision based approach is mostly used due to its advantages like non interaction, higher range, easy implementation, etc.

There are various methods that use different modes of data like images (2D/3D), sensor data, videos, etc. These image data in most cases are noisy and thus require more processing to use them for further evaluation and training models. The online datasets are taken in a controlled environment and can be used for model training. These sign languages are a way to integrate disabled people with the greater population by combining the body language and facial expression to express ideas and thoughts. A notable disadvantage is many people are still unaware and are unable to interpret sign language and thus there is a greater need to use modern recognition systems to aid them.

Machine learning offers a wide array of methods to solve this problem, like Naïve bayes, SVM, Decision trees, Artificial Neural Networks, etc. These methods are widely researched and evaluated for their effectiveness, accuracy, robustness and many more. This paper uses a Neural Network methodology, especially CNN to solve this classification problem.

This paper uses the CNN approach on an American Sign Language gesture which includes the alphabets and numerals. The methodology is explained, and results are discussed in the subsequent sections.

**LITERATURE SURVEY**

Convolution Neural networks have been used in Deep Neural Network architecture for a long time and have given impressive achievements in the field. They are a widely researched and used methodology both in academia and industry. The CNN architectures have proven highly successful like AlexNet, InceptionV3, GoogleNet, etc and have also developed in 1D and Multidimensional convolutional networks. Activation functions, like sigmoid and tanh, introduce non-linearity and are chosen based on the specific task. Loss functions, such as mean squared error or cross-entropy, measure how well the model performs. Finally, optimizers, including stochastic gradient descent and ADAM, adjust the network's internal parameters to minimize the loss function and improve performance. [1]

Researchers have also used CNN in medical image classification like in breast cancer detection. The ultrasound images are used to train a model that accurately (92.31%) detects cancerous breast tissues to non-cancerous ones. This is achieved without the need of mammograms which do not explicitly show cancerous breast tissues and employs ionising radiation to obtain. [2]

The authors of this paper dealing with Indian Sign Language Gestures have experimented with various parameters and have compared the accuracy between them. The paper compared SGD, Adam, Adam prop optimisers and between grayscale and RGB colour images. It also compared with various convolution kernel sizes and evaluated the various scores. It is found that in this setting SGD outperformed the other optimisers with 16 filters, 4 layers at 99.72% training accuracy and 98.56% validation accuracy. The dataset included 35,000 images classified into alphabets, numerals and 67 different commonly used words. [3]

A novel CNN architecture named G-CNN is proposed which effectively finds the features of the image is proposed in this paper. It is trained on Indian Sign Language dataset which is classified into 43 gestures. The architectures are made up of 4 convolution, 3 pooling layers, 2 dropout layers, 2 fully connected layers, 1 soft max layer. It has an input image size of 256\*256 and tanh activation function. It achieves an accuracy of 94.86% and 99.96% for Indian Sign Language finger spelling and isolated word recognition respectively. [4]

This paper analyses the accuracy between SVM, single- and double-layer CNN with MNIST Kaggle American sign language dataset. The input layer is 28\*28 pixels grayscale. The SVM with ‘poly’ kernel approach yields an accuracy of 81.49% which is highly outperformed by the CNN approaches. The single- and double-layer approaches gave 97.344% and 98.581% respectively. [5]

A 3DCNN architecture is used in this paper for sign language detection from 3 datasets – KSU-SSL, ArSL, RVL-SLLL. The first step involved video processing where linear sampling where the videos are normalized to fixed length of 16 frames. The 3DCNN deep learning technique is used here to learn the features. Transfer learning is used to overcome the rarity of large, labelled datasets. In another approach, the temporal contributions of the extracted frames are enhanced by preprocessing methods, like linear sampling to select 32 frames instead of 16. The model utilizes three parallel 3D convolutional neural networks (3DCNNs) to obtain both spatial and time series information throughout the video. To achieve this, the video is segmented into overlapping sections of three short clips, each containing 16 frames. Each 3DCNN independently analyzes a specific video segment, extracting features. These features are then combined using different fusion techniques, such as MLP and LSTM networks. The study reveals that MLP fusion performs better in this scenario and using parallel 3DCNNs overall leads to improved video classification accuracy. [6]

A 3DRCNN is used with a ASL dataset for sign language detection in video sequences which uses C3D along with RNN architecture. The input data is in 3 modes – RGB, optical flow and depth. It achieves modest accuracy when trained in multiple datasets. [7]

This paper proposes an ASL learning application that utilizes a combination of LSTM networks and the KNN algorithm. This approach is chosen because the dataset incorporates both still images and sequences of moving images. Key features are extracted from the data, including the radius of a virtual sphere around the hand, angles between fingers, and distances between fingertips. The system uses these extracted features to identify signs. Researchers trained the model with a dataset containing 2600 samples, 100 for each letter of the alphabet. Their evaluation using 5-fold cross-validation with a Leap Motion controller yielded impressive results, achieving an average accuracy of 99.44% and 91.82%. [8]

This paper describes a complete system implemented on an FPGA (Field-Programmable Gate Array) that classifies 24 American Sign Language (ASL) hand gestures. The system analyzes images captured by a neuromorphic sensor, focusing on the hand's outer contour. It functions in three main stages. First, it communicates with the camera and sensor to receive event data. Second, it processes the image by applying noise filtering and edge detection techniques. Finally, it utilizes a Static Artificial Neural Network (ANN) with data adjustment algorithms and a comparator to recognize the gestures. The system achieves an impressive accuracy of 80% in classifying the ASL gestures. [9]

This research focuses on Bhutanese Sign Language (BSL), a domain with limited existing research. To address the challenge of a small dataset, they employed data augmentation techniques to bolster its size and enhance its robustness. the experiments revealed that a Convolutional Neural Network (CNN) modeled after VGGNet architecture achieved the best performance, reaching an accuracy of 97.62% while requiring the least training time. Additionally, it is observed that Logistic Regression achieved the fastest testing time (175 microseconds) while the Support Vector Machine (SVM) model had the slowest training time (825 seconds). [10]

The authors propose a framework for Arabic Sign Language (ArSL) recognition and evaluate it using a non-signer-dependent approach on a dedicated database. The system first isolates the hands in video clips through DeepLabv3+, a semantic segmentation network. These hand regions are then cropped, resized to a uniform size (64x64 pixels), and converted to grayscale to account for variations in skin color across different signers. To capture hand shape features, a 2-dimensional convolutional Self-Organizing Map (SOM) is employed. Finally, a deep Bidirectional Long Short Term Memory (BiLSTM) network classifies each input video. The study focused on optimizing the parameters of DeepLabv3+, CSOM, and BiLSTM for maximizing signer-independent recognition accuracy. Evaluations on the Arabic benchmark dataset achieved an average accuracy of 89.5% when DeepLabv3+ was used for hand segmentation. Notably, this accuracy dropped significantly to 69.0% when the hand segmentation module was excluded, highlighting its crucial role in the overall performance. [11]

The research builds a dataset of Indian Sign Language (ISL) signs specifically related to banking tasks. A CNN model is then trained using individual video frames from the signs as input data. The final, fully connected layers of the CNN act as a filter, classifying the image and preparing it for an LSTM network. This two-stage approach achieves an accuracy of 81% in correctly predicting the corresponding text representation for the input signs. [12]

In this paper, the ASL Lexicon Video Dataset consisting of 100 words is divided into two sections for preprocessing and training in cascaded CNN. The video sequences is converted to several frames and processed individually, converting the color frames to grayscale. Median filtering is used to remove unwanted noise and spots, Histogram equalization is used to reduce illumination variations and resized, normalized. The cascaded 3D CNN is used in the training stage. The work achieves a precision of 96%, recall of 97.1% and F measure of 96.4%. [13]

This paper leverages a large dataset named RKS\_PERSIANSIGN to tackle sign language recognition for Persian sign language. The model starts by estimating the 3D keypoints of the hand from regular 2D images. These points are then used to build a skeletal representation of the hand. To capture a more comprehensive picture of the hand, the model projects this skeleton onto three distinct surface regions. This creates a more distinctive representation for the network to analyze. 3D convolutional neural networks (3DCNNs) are then employed to extract features from these stacked representations. Finally, a Long Short-Term Memory (LSTM) network takes the output from the 3DCNNs and utilizes its ability to handle sequential data to model the dynamic aspects of sign language gestures. This innovative approach achieves an impressive accuracy of 96.02%. [14]

This study proposes a new American Sign Language (ASL) recognition system using a Leap Motion sensor. This sensor offers a more affordable and user-friendly approach compared to traditional methods like Cybergloves and Microsoft Kinect. The system aims to recognize the 26 ASL signs for the English alphabet. It utilizes two machine learning algorithms, k-Nearest Neighbors (kNN) and Support Vector Machine (SVM), to achieve this. Experiments showed the SVM classifier achieved the highest accuracy, reaching 79.83%, while kNN reached 72.78%. [15]

This paper explores a multidimensional Hidden Markov Model (HMM) based system for recognizing American Sign Language (ASL) gestures. The system leverages a Cyberlove for capturing hand shape data and a Flock of Birds motion tracker for capturing hand movement data. This combined approach provides a comprehensive understanding of both the static hand posture and the dynamic trajectory of the signs. The system utilizes an HMM to analyze the extracted features and achieves an impressive average recognition accuracy of 95%. [16]

This paper introduces a feature extraction method and emphasizes seven features and uses a Madaline neural network for classification. 30 images each for 26 alphabets are used to train and the performance is compared with other models and found to be better. [17]

This paper proposes a methodology for learning of hand shape embeddings for discriminative American sign language gestures. The data is manually labelled and high confidence predictions to train a CNN and the sequential gesture element is used as training data for RNN. The model achieves an accuracy of 93% for separate learning and 89% for joint learning. [18]

This paper uses a Microsoft Kinect depth sensor to capture depth images of ASL, and this method of data collection overcomes the limitations of color images like illumination and background. The local features are extracted and segmented and are trained on a unsupervised Principal Component Analysis network. The paper employs two strategies namely a single PCAnet for all the users and separate PCAnet for each user. The extracted features are trained on a SVM. The proposed user specific PCANet model gives 84.50% and proposed single PCANet model gives 88.70% accuracy respectively. [19]

This study explores two approaches for recognizing gestures using the Massey dataset. The first approach extracts informative features from each gesture. It utilizes two techniques: Histogram of Oriented Gradients (HOG) to capture edge directions and Local Binary Pattern (LBP) to analyze local spatial patterns. The extracted features are fed into a multi-class SVM classifier to categorize the data. To compare performance, a Convolutional Neural Network (CNN) is also trained on the same dataset. While the CNN achieved a respectable accuracy of 97.08%, the handcrafted HOG/LBP features combined with an SVM using Radial Basis Function (RBF) and polynomial kernels achieved an even higher average accuracy of 98.36%. This result highlights that carefully chosen handcrafted features, coupled with a powerful classifier like SVM, can outperform deep learning approaches like CNNs in certain tasks. [20]

**METHODOLOGY**

DATASET DESCRIPTION

We created a dataset of 48\*48-pixel sign language image dataset for the alphabets and blank class that captures the background. The dataset contains 5687 images of 27 classes and were taken in non-uniform background and lighting. Some images are shown below.

A collage of hands making signs

Description automatically generated

PRE-PROCESSING OF DATASET

To prepare the images for the Convolutional Neural Network (CNN), we first converted them from colour to grayscale, removing unnecessary color information. Then, we resized all images to a uniform size of 48x48 pixels for efficient processing by the CNN. Finally, we normalized the pixel values from a range of 0-255 to 0-1. This normalization step ensures the CNN works more effectively with the data.

MODEL BUILDING

The model is built using 4 convolutional layers with kernel size 3\*3 and ReLU activation function. Dropout layers are used to increase the robustness of the model and is set to 40%. Max pooling layer is used in between the layers to reduce dimensionality of the data. A flattening layer is added to the last of convolutional layer. The neural layer has 4 dense layers and an output layer, with soft max activation. It is optimised with ADAM and categorical cross entropy loss function. The layers also have dropout layers to improve the robustness. The summary of model is shown below.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

TRAINING AND TESTING

To improve the model's robustness against overfitting, we employed data augmentation on the input images. This technique artificially creates variations of the original images by introducing randomness, such as rotations, shifts, and horizontal flips. We used the ImageDataGenerator library to perform these augmentations with specific parameters:

Rotation Range: Up to 40 degrees

Width Shift Range: Up to 20% horizontally

Height Shift Range: Up to 20% vertically

Shear Range: Up to a 20% shearing transformation.

Zoom Range: Up to a 20% zoom.

Horizontal Flip: Images were randomly flipped horizontally.

The model was then trained for 100 epochs, and the results are presented below.

A black and white rectangular labels

Description automatically generated with medium confidence

**RESULTS AND DISCUSSION**

The proposed model augments the generated sparse data and use them for training the model. This increases the robustness and decreases the chances of overfitting. The RGB images are converted to grayscale and rescaled to definite size. The CNN model has the feature of extracting convolution layer, which also helps in decreasing the dimension of the data. The layers are added with batch normalizing function to normalize the layer to unit mean unit variance data. The dropout layers randomly drop out neurons in the layer and strengthen the effect of neurons. The dense layers are progressively increased in number of neurons and are activated with ReLU function, and have dropout function. This model, when trained with the data achieved 94.43% accuracy, and 98.09% accuracy in validation data.

A screenshot of a graph

Description automatically generated

FIGURE 1 ACCURACY VS EPOCH

A graph of a graph

Description automatically generated with medium confidence

FIGURE 2 LOSS VS EPOCH

| A collage of hands with numbers  Description automatically generated | A collage of hands with different gestures  Description automatically generated | A collage of hands with numbers  Description automatically generated |
| --- | --- | --- |

FIGURE 3 SELECTED PREDICTIONS OF PROPOSED MODEL

**CONCLUSION**

This research proposes a new method for classifying and recognizing sign language. Sign language serves as a bridge, allowing deaf and hard-of-hearing people to communicate effectively and connect with the world.It has significant potential applications in various fields, including communication aids, human-computer interfaces, security systems, and advanced artificial intelligence. Sign language recognition has been a growing area of research, with scientists aiming to develop reliable, affordable, and accessible systems. These systems utilize various data sources like sensors, images, and videos. Current sign language datasets, often built in controlled lab environments, might not capture the natural variations of real-world signing. This work addresses this gap by creating a dataset that captures the natural variations of sign language encountered in everyday environments, ultimately leading to a more practical and robust recognition model. Using various techniques like converting image to grayscale, standardizing the image size, augmenting data to increase the size of dataset and its robustness, using batch normalization and dropout functions, and ReLU activation with ADAM optimizer along with Categorical cross entropy has yielded a robust model. The validation accuracy is 98.09% with the generated dataset.

There are some limitations in this approach. Data availability is a concern when using CNN since they generally need a higher amount of data for training, which can be partly solved using data augmentation techniques. The image processing techniques help in reducing the dimension of data and standardizing them but could also be improved by adding a region of interest identification system to further crop the image. Other feature identification methods like descriptors can be integrated in the CNN architecture to improve performance. These could be investigated in future works.

**REFERENCES**

1. Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 6999-7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

2. Pathan RK, Alam FI, Yasmin S, Hamd ZY, Aljuaid H, Khandaker MU, Lau SL. Breast Cancer Classification by Using Multi-Headed Convolutional Neural Network Modeling. Healthcare. 2022; 10(12):2367. <https://doi.org/10.3390/healthcare10122367>

3. Wadhawan, A., Kumar, P. Deep learning-based sign language recognition system for static signs. Neural Comput & Applic 32, 7957–7968 (2020). <https://doi.org/10.1007/s00521-019-04691-y>

4. Sakshi Sharma, Sukhwinder Singh Vision-based hand gesture recognition using deep learning for the interpretation of sign language <https://doi.org/10.1016/j.eswa.2021.115657>

5. Jain, V., Jain, A., Chauhan, A. et al. American Sign Language recognition using Support Vector Machine and Convolutional Neural Network. Int. j. inf. tecnol. 13, 1193–1200 (2021). <https://doi.org/10.1007/s41870-021-00617-x>

6. M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif and M. A. Mekhtiche, "Hand Gesture Recognition for Sign Language Using 3DCNN," in IEEE Access, vol. 8, pp. 79491-79509, 2020, doi: 10.1109/ACCESS.2020.2990434.

7. Recognizing American Sign Language Gestures From Within Continuous Videos

Yuancheng Ye, Yingli Tian, Matt Huenerfauth, Jingya Liu; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 2064-2073

8. C.K.M. Lee , Kam K.H. Ng , Chun-Hsien Chen , H.C.W. Lau , S.Y. Chung , Tiffany Tsoi American sign language recognition and training method with recurrent neural network <https://doi.org/10.1016/j.eswa.2020.114403>

9. Rivera-Acosta M, Ortega-Cisneros S, Rivera J, Sandoval-Ibarra F. American Sign Language Alphabet Recognition Using a Neuromorphic Sensor and an Artificial Neural Network. Sensors. 2017; 17(10):2176. <https://doi.org/10.3390/s17102176>

10. Karma Wangchuk, Panomkhawn Riyamongkol, Rattapoom Waranusast Real-time Bhutanese Sign Language digits recognition system using Convolutional Neural Network <https://doi.org/10.1016/j.icte.2020.08.002>

11. S. Aly and W. Aly, "DeepArSLR: A Novel Signer-Independent Deep Learning Framework for Isolated Arabic Sign Language Gestures Recognition," in IEEE Access, vol. 8, pp. 83199-83212, 2020, doi: 10.1109/ACCESS.2020.2990699.

12. G. Jayadeep, N. V. Vishnupriya, V. Venugopal, S. Vishnu and M. Geetha, "Mudra: Convolutional Neural Network based Indian Sign Language Translator for Banks," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1228-1232, doi: 10.1109/ICICCS48265.2020.9121144.

13. Sharma, S., Kumar, K. ASL-3DCNN: American sign language recognition technique using 3-D convolutional neural networks. Multimed Tools Appl 80, 26319–26331 (2021). <https://doi.org/10.1007/s11042-021-10768-5>

14. Hand sign language recognition using multi-view hand skeleton

Author links open overlay panelRazieh Rastgoo a, Kourosh Kiani a, Sergio Escalera <https://doi.org/10.1016/j.eswa.2020.113336>

15. C. -H. Chuan, E. Regina and C. Guardino, "American Sign Language Recognition Using Leap Motion Sensor," 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, USA, 2014, pp. 541-544, doi: 10.1109/ICMLA.2014.110.

16. American Sign Language Recognition Using Multi-dimensional Hidden Markov Models

HONGGANG WANG , MING C. LEU AND CEMIL OZ

17. S. Saha, R. Lahiri, A. Konar and A. K. Nagar, "A novel approach to American sign language recognition using MAdaline neural network," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 2016, pp. 1-6, doi: 10.1109/SSCI.2016.7850121.

18. A. A. Hosain, P. S. Santhalingam, P. Pathak, H. Rangwala and J. Košecká, "FineHand: Learning Hand Shapes for American Sign Language Recognition," 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 2020, pp. 700-707, doi: 10.1109/FG47880.2020.00062.

19. W. Aly, S. Aly and S. Almotairi, "User-Independent American Sign Language Alphabet Recognition Based on Depth Image and PCANet Features," in IEEE Access, vol. 7, pp. 123138-123150, 2019, doi: 10.1109/ACCESS.2019.2938829.

20. H. B. D. Nguyen and H. N. Do, "Deep Learning for American Sign Language Fingerspelling Recognition System," 2019 26th International Conference on Telecommunications (ICT), Hanoi, Vietnam, 2019, pp. 314-318, doi: 10.1109/ICT.2019.8798856.