



Machine Learning and Vision Based Techniques for Detecting and Recognizing Indian Sign Language



Navaneetha Krishna Bose Duraimutharasan^{1*}, Kumaravelu Sangeetha²

¹ Dean-Faculty of Advanced Computing Sciences, AMET University, Kanathur, Chennai 603112, Tamilnadu, India

² Department of Computer Science & Engineering, AMET University, Kanathur, Chennai 603112, Tamilnadu, India

Corresponding Author Email: duraibose@gmail.com

<https://doi.org/10.18280/ria.370529>

ABSTRACT

Received: 12 May 2023

Revised: 25 July 2023

Accepted: 3 August 2023

Available online: 31 October 2023

Keywords:

detection, CNN, AlexNet, ResNet, comparative performance

Despite rapid global advancement in technology, the persistent challenge of hearing impairments affects a significant proportion of the global population. Individuals with these impairments face complex communication barriers daily. The Indian Sign Language (ISL) has emerged as a universal communication tool for individuals with hearing impairments in India, playing a vital role in educational institutions and bridging societal gaps in a country marked by rich cultural and linguistic diversity. This work presents an innovative Supervised Learning approach for ISL recognition that extends beyond traditional classification techniques. The method employs advanced algorithms designed to classify new observations. Utilizing an expansive dataset, intricate patterns and nuances are identified, fostering accurate and adaptive decision-making. A Convolutional Neural Network (CNN) algorithm is applied, not only for data classification but also for iterative learning and refinement of classification boundaries. In the vast expanse of n-dimensional space, the CNN strives to identify optimal hyperplanes, establishing dynamic decision boundaries to adeptly categorize diverse data points. This approach transcends traditional classification boundaries, offering a more nuanced and effective data-driven decision-making process. This research heralds a new direction in addressing communication barriers, with potential applications extending beyond the realm of ISL. The assimilation of numerous regional languages into the sign language matrix, whilst challenging, is key to fostering a life of normalcy and integration.

1. INTRODUCTION

World Health Organization claims just a little over one-fourth of the world's deaf population resides in South Asia, which includes India's deaf minority (WHO). The largest of the eight nations that make up South Asia is India. The greatest deaf population would also follow this; however, the numbers are hazy [1]. There are sixty-three million individuals in the globe, according to estimates. Since communication is the sole means by which we may share our thoughts or disseminate a message, a person with a disability (such as someone who is deaf) finds it difficult to convey their feelings and emotions. India has significant population of 2.42M persons who are deaf and dumb. The development of gesture recognition faces challenges with everything from image capturing to classification [2-4]. The ideal approach for acquiring images is still being explored by researchers. The challenges of picture pre-processing are presented when gathering photos with a camera. obtaining data, collecting data (such as data from early studies or one's own research), recognizing methods lately employed by researchers, and the results of earlier studies [5-8].

Cameras are the primary input method in Indian Sign Language Recognition (ISLR) [9, 10]. The different gesture are stored in the memory. When the user shows the hand gesture the information. The speech is converted to sign language [11-14] explains how the percentage can be

calculated for the person's affected with disabilities, Static and dynamic are the two main categories in sign language. Continuous hand and facial motions are seen as static signals, whereas isolated and continuous signs are regarded as dynamic signs. For example, "welcome," "care," and many other words, are examples of isolated signs. Continuous signs, on the other hand, the series of distinct gestures that include both physical movements and expression of emotions, such as "how are you". 200+ different sign languages, including American, Spanish, Arabic, Italian, and Indian sign languages. The scholars utilized Indian sign gestures as an instance of suggested method [15]. There are 20 static indications and 635 samples in the specifically designed dataset for this system. Commonly used terms like hello etc. were utilized on the signage. The system was developed to transform impromptu gestures into sign gestures, which are then sent through a series of additional processing and preparatory tasks. These tasks ascertain the intended word corresponding to the motion.

2. LITERATURE REVIEW

The visual-spatial language known as Indian Sign Language was created in India. A real language having its own structure, syntax, and phonology is Indian Sign Language. It employs body/head movements, hand gestures, face expressions, and arm motions to create semantic information that conveys

words and emotions [15].

A method for recognizing and identifying Indian Sign gestures motions from outlined photographs was put out by Nandy et al. [16]. Their method involves converting a video source with signing movements into frames that are gray scaled, where features are then retrieved using the focused image. Finally, clustering is utilized to group the signs into one of the pre-defined classes based on their attributes. Using a thirty six bin histogram, which shown to be superior to the eighteen bin histogram approach by the authors, who also noted that their study had a 100% sign recognition rate.

For instant text creation in a video transmission with sign language monitoring and identification, Authors in the study [17] developed a neural network design. The system architecture is made up of a number of phases, including framing, picture pre-processing, feature extraction based on hand position and movement, etc. A hand point of interest (POI) it represents these hand characteristics. This technique retrieved 55 distinct characteristics, which were then inputted into the authors' proposed feature-prediction CNN-layered neural network architecture. In their study, they claimed to have achieved 48% noise tolerance and 100% rate of detection after the English letters for model training and testing from A to Z.

A self-created getures of 3230 samples of 10 prestored gestures or characters, Chen [18] suggested a model. Pre-processing was first carried out via edge segmentation to find the hand's edges, and for skin colour segmentation, the technique involved converting RGB pictures to YUQ or YIQ colour spaces. The convex hull approach was then used to identify the fingers from the hand that had already been recognised. Finally, the categorization technique was based on neural networks. This model's ultimate accuracy score was 98.2%.

Sharma et al.'s [19] technique for conversing with people who have speech or hearing impairments was based on Indian Sign Language. After the image was captured, the data were first pre-processed by transforming them utilising the Matlab it is possible to convert Colour to grayscale. The edges of the picture were then discovered using a Sobel detector and a 3 3 filter. The 600 Element reduced picture was then subjected to the hierarchical centroid technique, which produced 124 features. KNN and neural networks were the classification approaches that were utilized. 97.10% accuracy was achieved using this practise.

By employing glove which can sense the gestures analysing signals, then exhibiting the output form to comprehensible phrase, Agarwal et al.'s [20] goal was to close the communication gap between those with speech difficulty and those with typical speaking ability. The individuals used the sensor gloves to make the movements. The dataset was compared, recognised were then sent on to the processed to produce a phrase. The application's accuracy in version 1 was less than 50%. Whereas the other which introduced a keyword for the necessary tense, resulting in cent percent correctness when dealing with simple and continuous tenses.

A technique for deriving sign language translation from raw video by dividing the layout was proposed by Wazalwar and Shrawankar [21]. They employed the P2DHMM which uses the hand detecting and CamShift technique for detecting. Using a Haar Cascade classifier, the indications were identified. Once the sign was identified, Using the parser detect the phrase, which is gives the output as text, which forms comprehensible phrases. Tagger is assigned as the sign

is recognized.

For the purpose of identifying signals and motions, Shivashankara and Srinath [22] has created a gesture for sign language interpretation. The authors put up a model that made use of YCbCr to improve the efficiency of skin colour clustering. The pre-processing of images was done using this model. The pre-processed image's centroid was used to recognise the sign language; after that, gesture's peak offset was discovered. This model's total accuracy was 93.05%.

The method that demonstrated that the translation of sign language which involve vocal communication, rather a one to one. The processes of common translation were replicated in authors' novel vision technique. The CNN design [23] is integrated with encoding and decoding technique and the conditional which will produce a vocal based video which exhibit videos in the translation stage, which converts sign gestures to vocal communication. Starting with word embedding, the authors use the vector techniques to convert to meaningful words which were located. Probability was increased by using encoder-decoder phase. During encoding, a sign gesture is being produced. It suggested a dataset from different type of people for training data set Mariappan and Gomathi [24]. It suggested to use a network to identify sign language from facial expression by De Coster et al. [25]. It proposes graph techniques to identify sign language by Jiang et al. [26]. It suggested dynamic method for identifying sign language. Its uses gestures for Chinese sign gestures.

3. METHOD

Deep learning is currently proven to be quite effective in solving issues that were previously considered to be unsolvable or extremely complicated, especially related to techniques of problem solving. The Indian Sign Language recognition involves important five step as mentioned in Figure 1. The sign gestures Involves image collection, it is accomplished using data set. Preprocessing is the second stage, when undesirable noise is removed and quality enhanced. The second is the detection is to take the important values of the gestures. A gesture region involves extraction and next step is extraction of information. The characteristics of new sign gestures of characteristics recorded in the dataset to identify provided gestures is referred to as classification [27, 28].

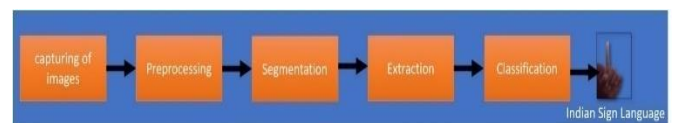


Figure 1. Indian sign language recognition

3.1 Capturing devices

To extract the pictures which has to be classified, utilized different sign gesture. The image capturing tools, hand glove, jump controller are some of instruments used. A image capturing is used by most of the researchers than hand glove which provide accurate information. Data collection with a data glove has shown to be more accurate, however it is quite expensive and cumbersome for consumers. Kinect has more accuracy and extensively used. It has deep visuals and colour streaming. It enables the 3D image where the background

noise and other associated information are removed. Figure 2 show the different capturing Devices.

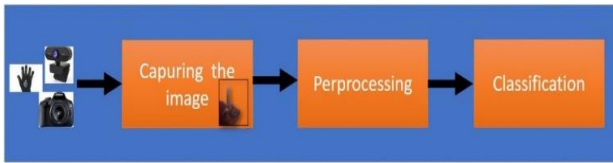


Figure 2. Capturing devices

3.2 Preprocessing

An image's quality is improved and undesirable noise is removed using preprocessing techniques. Resizing, colour conversion, eliminating undesired noise, where the actual image is restored. The accurate preprocessing techniques, output generated is accurate. The methods are enhancement and restoration of an image. There are many techniques involved in image preprocessing. Figure 3 shows the different techniques applied while the images being preprocessed.

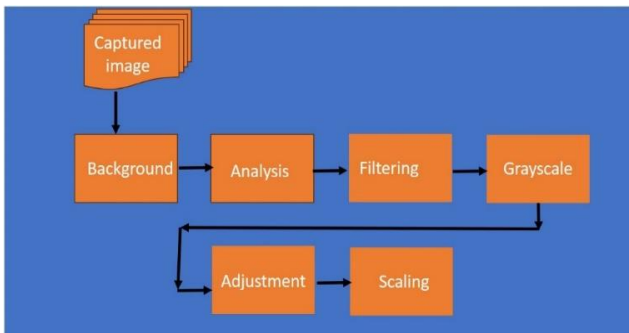


Figure 3. Preprocessing images

3.3 Segmentation and extraction

Image segmentation, a critical process in the digitalization of images, is employed to group pixels that share similar attributes. Given the significance of pixels in the digital representation of images, their grouping forms the foundation for further image processing tasks. In the methodology of this study, the Bresenham algorithm is initially applied for this segmentation task. Subsequent to the segmentation phase, the process advances to classification. This stage is instrumental in extracting salient image components, effectively isolating them from the background. Figure 4 illustrates this segmentation and subsequent classification, demonstrating the extraction of key image elements.

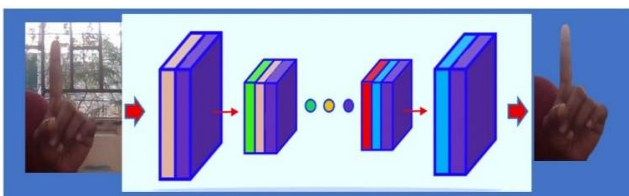


Figure 4. Segmentation of images

3.4 Image classification techniques

The classification methods use the supervised learning method to find the labels for the input data given. The data are

being trained so that when information comes dynamically it has to be predicted. Figure 5 shows how the data are being trained.

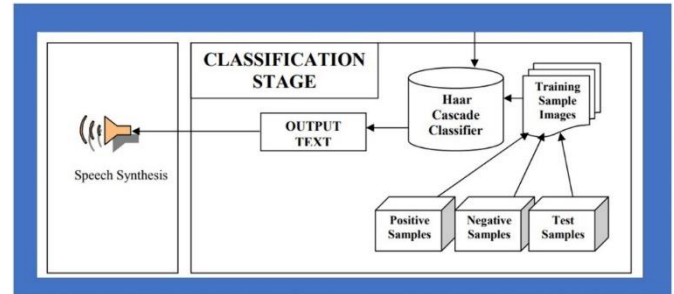


Figure 5. Classification of images

4. EXPERIMENTAL METHODOLOGY

As implied by the name, is a type of neural network which are similar to human sixth sense. It consists of a network of neurons, or learning cells. Automated recognition is built on the ability of these neurons to translate input signals (such as an image of an Indian Sign Language) into equivalent output signals (such as the label "One"). Repeated neural network building blocks which generated over pictures which can be images videos etc. in CNNs. Neural building techniques for pictures can be done by 3D Convolutional Neural Networks which can be a repetitive to all the image path. The Convolutional Neural Network can be generated for various text to speech. The repeated information can be shared and trained. Figure 6 shows Convolutional Neural Networks where Sign Language is used as image.

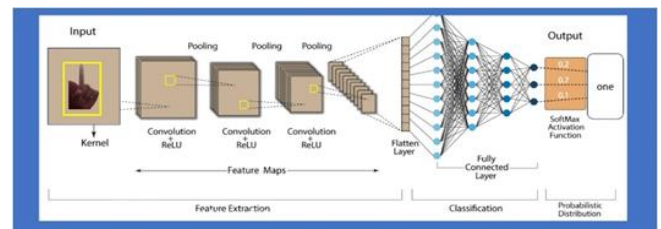


Figure 6. Classification of images

When a signal is generated a Convolution Neural Network starts the initial process. During convolution, the network attempts to identify the trained data which is being stored. The sign gesture which shows a symbol "Welcome" if it matches the different signals stored. Figure 7 shows the filter reacts to eye.

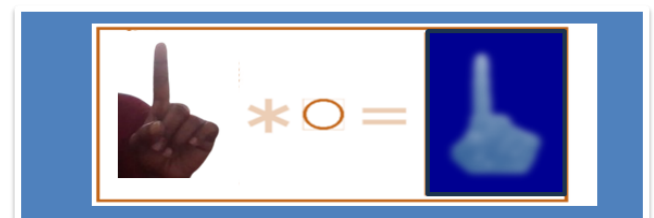


Figure 7. The circular filter reacts to the eyes significantly

Convolution neural network has the invariance. It filters the feature such as (such as hair), and the information which is

being trained. The output does not affect. Therefore, image of Sign Language with slight change also will be identified. We can an example which has a 3D filter $5 \times 5 \times 3$ and has a picture of $32 \times 32 \times 3$ dimension. We can dimension of $5 \times 5 \times 3$ apply it to the whole picture, it takes only dot matrix and filter is being applied. Figure 8 shows the Dot product.

Convolution Layer are being used, six separate blocks in the convolution layer. Figure 9 is individually convolved with each filter, and the result is six feature maps with the dimensions $32 \times 32 \times 1$.

The picture is individually convolved with each filter, and the result is six feature maps with the dimensions $32 \times 32 \times 1$. Figure 10 shows the Convolutional Layer of 6 Independent filter.

The Convolutional layers, which has a Comparable Filters, which points to CNN. For example, the diagram shows two levels of convolutional filters, 6 and 10. Figure 11 shows the Convolutional layer with 3, 6 and 10 layers.

Convolution Neural Network minimizes the noise. The maximum values of the signals are calculated and this smoothing technique known as subsampling is accomplished. The size is reduced and colour resolution is minimized which are the strategies for subsampling (picture inputs). Figure 12 shows how a lower resolution is created.

The pooling layers are important features of CNN. In order to minimize the number of variables and calculations in the network, the geographic width of the representation is gradually decreased. Each feature is calculated by the layer and the most prevalent pooling method is maximum pooling, which chooses the largest value for the region as its representative. For example, greatest value is used to replace a 2×2 section in the following diagram. Figure 13 shows the Pooling layer to reduce size.

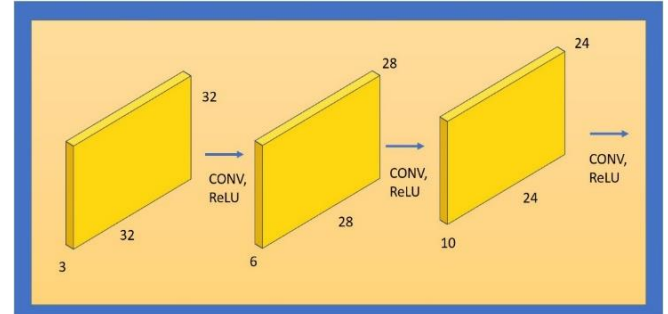


Figure 11. Convolutional layer with 3, 6 and 10 layers



Figure 12. Lower resolution is created

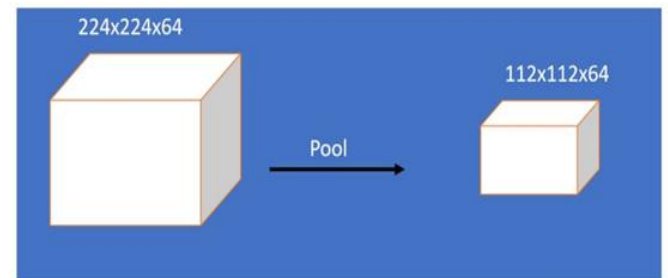


Figure 13. Pooling layer

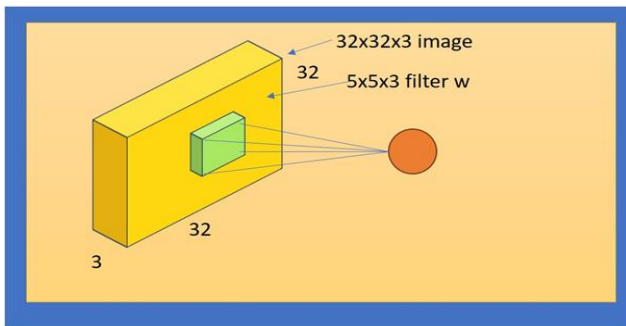


Figure 8. Dot product

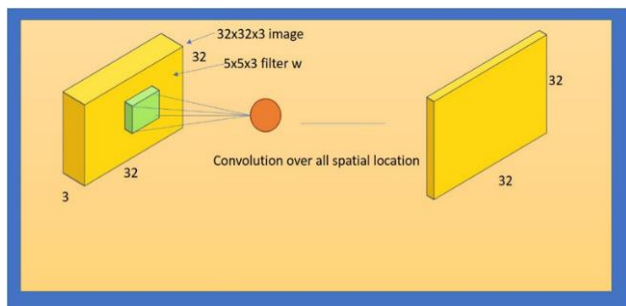


Figure 9. 5×5 Spatial location

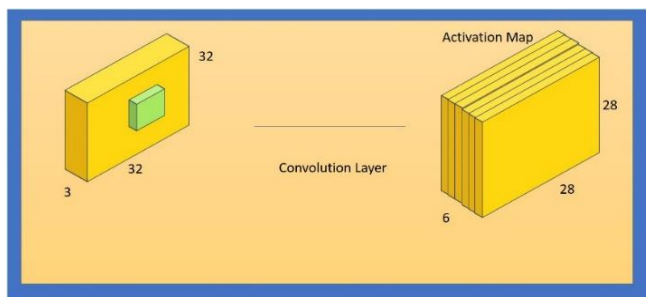


Figure 10. Convolutional layer of 6 independent filter

5. DISCUSSION

The convolutional Neural Network is between the neurons and artificial neural networks, is modelled after the way that the visual information is organized. The various sign language movements are converted into the relevant text to speech using 6DT smart gloves made of finger flex sensors [2]. Each letter in the India Sign Language has a unique combination that is communicated through Bluetooth to an Arduino Board. The power supply is made of solar so recharge could be done. Figure 14 and Figure 15 show the block diagram of training set and display output.

The neural network consists data pertaining to training and testing. The CNN use signal's receiving and output layer which

finds the class of the input being given. Between the two Layer (input and output) consist of 200 perception which is hidden. The neural networks can be used to find errors, a collection of feature variables or logits is translated to distribution of probabilities.

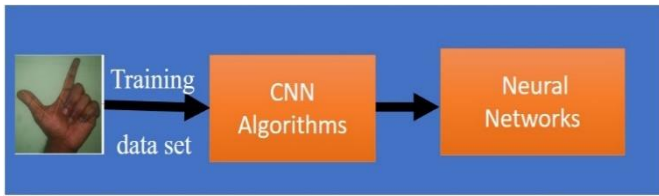


Figure 14. Testing set-block diagram

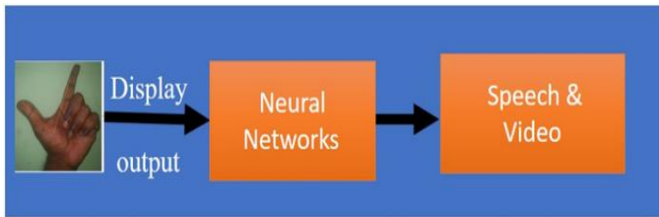


Figure 15. Display output-block diagram

CNN's capability to leverage temporal and spatial data integration is one of its most important attributes. There are numerous Nonlinear processing units, subsampling layers, and convolutional layers in each Convolution Neural Networks learning stage. The most widely used Convolution Neural Network design is the LeNet modelling, which made its debut in the year 1998. LeNet was initially created to classify the written in numerals starting at 0 in the Collection of MNIST. There are a total of seven levels, and every layer has a different set of trainable parameters. The network will take images up to 32x32 pixels in size, which is a relatively large image when compared to the images in the data sets that the network is constructed on. ReLU is the function of stimulation.

It is based on the LeNet Architecture, but unlike the original LeNet, it has a lot more filters, which allows you to categorize many more entities. Additionally, "Dropout" is used to correct as opposed to regularization. AlexNet design exhibits five convolutional layers and three linked layers in their entirety.

A building block of ResNets is referred to as a residual block. This uses batch-normalization extensively and is designed to successfully train hundreds of layers and is based on the idea of "skip-connections" over the time without giving up speed.

Convolutional neural network architecture known as VGG has been around for a while. VGG was developed with Especially in comparison to AlexNet and ZfNet, there are 19 deep Layer mimic the connection between network representational capabilities and depth. ZfNet, the effectiveness of CNN can be improved. Considering these findings, Very Deep Convolutional Neural Network substituted a stack of 3x3 filters for the 11x11 and 5x5 filters, demonstrating that the impact of a large-size filter may be accomplished by positioning small-and large-size filters next to one another (5x5 and 7x7). Tiny size filters have the added benefit of having minimal computational complexity because they use fewer parameters. As a result of these findings, CNN has started using smaller filters for its research.

PolyNet travels the dimension captures and looks through every region, making smart choices regarding weights and

layout that is it can Automates changes that can improve functionality and effectiveness with improved user outcomes. A first You can utilise PolyNet, an internal neural network training system, to ensure that your data never leaves the building. That is what distinguishes and makes PolyNet unique. Highway Networks and ResNet were also proposed, and DenseNet was made to deal with the vanishing gradient issue.

The accuracy is being checked for different CNN model. Figure 16 show the performance of the accuracy.

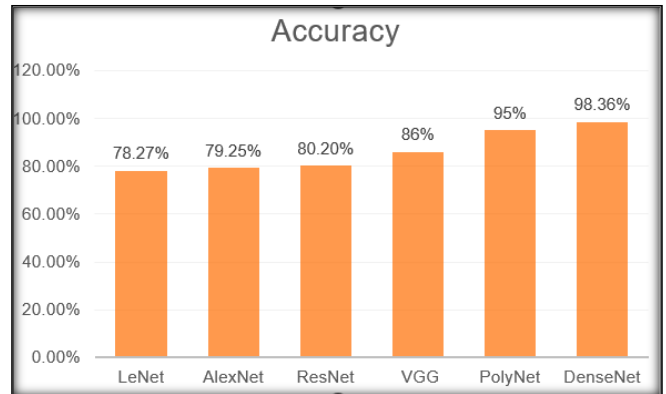


Figure 16. Accuracy of different CNN model

Integration of Glove and Convolution Neural Network has shown a great success as the hand gesture which could be converted into speech will be extremely useful for hearing impaired. The CNN will be helpful in creation of new datasets which could be in millions of storages. The DenseNet CNN model will give 98.2622% accuracy.

6. CONCLUSIONS

It was discovered that DenseNet CNN model had the maximum accuracy of 98.622% after constructing the suggested with the aid of a neural network. When networks are built with little training data, overfitting occurs. Due to the absence of a vast and diversified dataset used to train the neural network so built is only able to categorise the practise exceptionally high precision, and is therefore unable to categorise any test record besides learning records [7].

The object does not generate a generic answer because it is taught to work for a particular dataset. As a result, accuracy suffers when the dataset is less.

REFERENCES

- [1] <https://wecapable.com/disabled-population-india-data/>.
- [2] Elmahgiubi, M., Ennajar, M., Drawil, N., Elbuni, M.S. (2015). Sign language translator and gesture recognition. In 2015 Global Summit on Computer & Information Technology (GSCIT), IEEE, pp. 1-6. <https://doi.org/10.1109/GSCIT.2015.7353332>
- [3] Sze, V., Chen, Y.H., Yang, T.J., Emer, J.S. (2017). Efficient processing of deep neural networks: A tutorial and survey. Proceedings of the IEEE, 105(12): 2295-2329. <https://doi.org/10.1109/JPROC.2017.2761740>
- [4] Gao, B.L., Pavel, L. (2017). On the properties of the softmax function with application in game theory and reinforcement learning. arXiv Preprint arXiv:

- 1704.00805. <https://doi.org/10.48550/arXiv.1704.00805>
- [5] Different Types of CNN Models. (n.d.). IQ OpenGenus. <https://iq.opengenus.org/different-types-of-cnn-models/>.
- [6] Gupta, G. (2014). A self explanatory review of decision tree classifiers. In International Conference on Recent Advances and Innovations in Engineering (ICRAIE), IEEE, 2014: 1-7. <https://doi.org/10.1109/ICRAIE.2014.6909245>
- [7] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1): 1929-1958.
- [8] Mehdi, S.A., Khan, Y.N. (2002). Sign language recognition using sensor gloves. In Proceedings of the 9th International Conference on Neural Information Processing, ICONIP'02, Singapore, pp. 2204-2206. <https://doi.org/10.1109/ICONIP.2002.1201884>
- [9] Salian, S., Dokare, I., Serai, D., Suresh, A., Ganorkar, P. (2017). Proposed system for sign language recognition. In 2017 International Conference on Computation of Power, Energy Information and Commuication (ICCPEIC), Melmaruvathur, India, pp. 58-62. <https://doi.org/10.1109/ICCPEIC.2017.8290339>
- [10] Nikam, A.S., Ambekar, A.G. (2016). Sign language recognition using image based hand gesture recognition techniques. In 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, India, pp. 1-5. <https://doi.org/10.1109/GET.2016.7916786>
- [11] Piriadarshani, D., Sasikala, K., Sangeetha, K., Naveena, N.R. (2022). Analysis of characteristic roots of neutral delay differential equation using generalized lambert W function. *Advanced Engineering Science*, 54(2): 2337-2344.
- [12] Piriadarshani, D., Sasikala, K. (2022). Spectral legendre approximation for the population growth model of E. coli. *NeuroQuantology*, 20(10): 8497-8501. <https://doi.org/10.14704/nq.2022.20.10.NQ55834>
- [13] Piriadarshani, D., Sasikala, K., James, B., Narasimhan, S., Nishi, N.D. (2020). Stability of neutral delay differential equation using spectral approximations. *European Journal of Molecular & Clinical Medicine*, 7(2): 5006-5015.
- [14] Wadhawan, A., Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, 28: 785-813. <https://doi.org/10.1007/s11831-019-09384-2>
- [15] Papastratis, I., Chatzikonstantinou, C., Konstantinidis, D., Dimitropoulos, K., Daras, P. (2021). Artificial intelligence technologies for sign language. *Sensors*, 21(17): 5843. <https://doi.org/10.3390/s21175843>
- [16] Nandy, A., Prasad, J.S., Mondal, S., Chakraborty, P., Nandi, G.C. (2010). Recognition of isolated indian sign language gesture in real time. In Information Processing and Management: International Conference on Recent Trends in Business Administration and Information Processing, Springer Berlin Heidelberg, pp. 102-107. https://doi.org/10.1007/978-3-642-12214-9_18
- [17] Mekala, P., Gao, Y., Fan, J., Davari, A. (2011). Real-time sign language recognition based on neural network architecture. In 2011 IEEE 43rd Southeastern Symposium on System Theory, Auburn, AL, USA, pp. 195-199. <https://doi.org/10.1109/SSST.2011.5753805>
- [18] Chen, J.K. (2011). Sign language recognition with unsupervised feature learning; CS229 project final report. <https://cs229.stanford.edu/proj2011/ChenSenguptaSundaram-SignLanguageGestureRecognitionWithUnsupervisedFeatureLearning.pdf>.
- [19] Sharma, M., Pal, R., Sahoo, A.K. (2014). Indian sign language recognition using neural networks and KNN classifiers. *ARPN Journal of Engineering and Applied Sciences*, 9(8): 1255-1259.
- [20] Agarwal, S.R., Agrawal, S.B., Latif, A.M. (2015). Sentence formation in NLP engine on the basis of indian sign language using hand gestures. *International Journal of Computer Applications*, 116(17): 18-22. <https://doi.org/10.5120/20428-2757>
- [21] Wazalwar, S.S., Shrawankar, U. (2017). Interpretation of sign language into English using NLP techniques. *Journal of Information and Optimization Sciences*, 38(6): 895-910. <https://doi.org/10.1080/02522667.2017.1372136>
- [22] Shivashankara, S., Srinath, S. (2018). American sign language recognition system: An optimal approach. *International Journal of Image, Graphics and Signal Processing*, 10(8): 18-30. <https://doi.org/10.5815/ijgisp.2018.08.03>
- [23] Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R. (2018). Neural sign language translation. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7784-7793. <https://doi.org/10.1109/CVPR.2018.00812>
- [24] Mariappan, H.M., Gomathi, V. (2019). Real-time recognition of Indian sign language. In 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, pp. 1-6. <https://doi.org/10.1109/ICCIDS.2019.8862125>
- [25] De Coster, M., Van Herreweghe, M., Dambre, J. (2020). Sign language recognition with transformer networks. In 12th International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), pp. 6018-6024.
- [26] Jiang, S.Y., Sun, B., Wang, L.C., Bai, Y., Li, K.P., Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 3413-3423. <https://doi.org/10.1109/CVPRW53098.2021.00380>
- [27] Raj, R.D., Jasuja, A. (2018). British sign language recognition using HOG. In 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), pp. 1-4. <https://doi.org/10.1109/SCEECS.2018.8546967>
- [28] Pokharna, H. (2016). The best explanation of Convolutional Neural Networks on the internet. Medium. <https://medium.com/technologymadeeasy/the-best-explanation-of-convolutional-neural-networks-on-the-internet-fbb8b1ad5df8>.