IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|

Sign Language Recognition Using Convolutional Neural Network

Suyog Gandhare[1], Niranjan Bhokare[1], Vaibhav Khandagale[1], Sugam Ingle[1], Prof. Ashish Gaigol[2]

UG Student, Dept. of Computer Engineering, JSPM's Imperial College of Engineering & Research Wagholi Pune, Maharashtra, India[1]

Assistant Professor, Dept. of Computer Engineering, JSPM's Imperial College of Engineering & Research Wagholi Pune, Maharashtra, India[2]

ABSTRACT: Indian Sign Language Recognition has transpire as one of the important area of research in Computer Vision .The real-time sign language recognition system is developed for identification the gestures of Indian Sign Language. Sign Language is the most expressive form of communication for speech and hearing diminish people to communicate with normal person but a normal person cannot grip sign language. we propose a new system based on the convolutional Neural Networks, nourish with a real dataset, this system will identify automatically numbers and letters of Indian sign language.The project is software-based which can be installed on any computer with good specifications. The training and prediction of hand gestures are performed by applying Convolutional Neural Network clustering Deep learning algorithm. Deep learning is among the best known techniques for such speculate. Deep learning is a data analytics technique that provides machine the potential to learn without being comprehensively programmed.

KEYWORDS: Convolutional Neural Network, Sign Language Recognition, HandSegmentation, Deep Leaning, karas, OpenCV.

I. INTRODUCTION

As well stipulated by Nelson Mandela, "Talk to man in a language he understands that goes to his head. Talk to him in his own language, that goes in his heart." Language is undoubtedly essential to human interaction and has existed since human civilization began. It is a medium humans use to communicate to express themselves and understand notions of the real world. Without it, no books, no cell phones and definitely not any word I am writing would have any meaning. It is so deeply embedded in our everyday routine that we often take it for granted and don't realize its importance. Sadly, in the fast-changing society we live in, people with hearing impairment are usually forgotten and left out. They have to struggle to bring up their ideas, voice out their opinions and express themselves to people who are different to them. Sign language, although being a medium of communication to deaf people, still have no meaning when conveyed to a non-sign language user. Hence, broadening the communication gap. To prevent this from happening, we are putting forward a sign language recognition system. It will be an ultimate tool for people with hearing disability to communicate their thoughts as well as a very good interpretation for non-sign language user to understand what the latter is saying. Many countries have their own standard and interpretation of sign gestures. For instance, an alphabet in Korean sign language will not mean the same thing as in Indian sign language. While this highlights diversity, it also pinpoints the complexity of sign languages. Deep learning must be well versed with the gestures so that we can get a decent accuracy. In our proposed system, American Sign Language is used to create our datasets.

Figure 1 shows the American Sign Language (ASL) alphabets. Identification of sign gesture is performed with either of the two methods. First is a glove-based method whereby the signer wears a pair of data gloves during the capture of hand movements. Second is a vision-based method, further classified into static and dynamic recognition [2]. Static deals with the 2-dimensional representation of gestures while dynamic is a real time live capture of the gestures. And despite having an accuracy of over 90% [3], wearing of gloves are uncomfortable and cannot be utilized in rainy weather. They are not easily carried around since their use require computer as well. In this case, we have decided to go

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|

with the static recognition of hand gestures because it increases accuracy as compared to when including dynamic hand gestures like for the alphabets J and Z. We are proposing this research so we can improve on accuracy using Convolution Neural Network (CNN).

Motivation:

The 2011 Indian census cites roughly 1.3 million people with "hearing impairment". In contrast to that numbers from India's National Association of the Deaf estimates that 18 million people –roughly 1 per cent of Indian population are deaf. These statistics formed the motivation for our project. As these speech impairment and deaf people need a proper channel to communicate with normal people there is a need for a system. Not all normal people can understand sign language of impaired people. Our project hence is aimed at converting the sign language gestures into text that is readable for normal people.

II. LITERATURE SURVEY

Siming He [1] proposed a system having a dataset of 40 common words and 10,000 sign language images. To locate the hand regions in the video frame, Faster R-CNN with an embedded RPN module is used. It improves performance in terms of accuracy. Detection and template classification can be done at a higher speed as compared to single stage target detection algorithm such as YOLO. The detection accuracy of Faster R-CNN in the paper increases from 89.0% to 91.7% as compared to Fast-RCNN. A 3D CNN is used for feature extraction and a sign-language recognition framework consisting of long- and short-time memory (LSTM) coding and decoding network are built for the language image sequences. On the problem of RGB sign language image or video recognition in practical problems, the paper merges the hand locating network, 3D CNN feature extraction network and LSTM encoding and decoding to construct the algorithm for extraction.

This paper has achieved a recognition of 99% in common vocabulary dataset. Let's approach the research done by Rekha, J [2]. which made use of YCbCr skin model to detect and fragment the skin region of the hand gestures. Using Principal Curvature based Region Detector, the image features are extracted and classified with Multi class SVM, DTW and non-linear KNN. A dataset of 23 Indian Sign Language static alphabet signs were used for training and 25 videos for testing. The experimental result obtained were 94.4%for static and 86.4% for dynamic.

In [3], a low-cost approach has been used for image processing. The capture of images was done with a green background so that during processing, the green colour can be easily subtracted from the RGB colourspace and the image gets converted to black and white. The sign gestures were in Sinhala language. The method that thy have proposed in the study is to map the signs using centroid method. It can map the input gesture with a database irrespective of the hands size and position. The prototype has correctly recognised 92% of the sign gestures.

The paper by M. Geetha and U. C. Manjusha [4], make use of 50 specimens of all alphabets and digits in a vision-based recognition of Indian Sign Language characters and numerals using B-Spine approximations. The region of interest of the sign gesture is analysed and the boundary is removed. The boundary obtained is further transformed to a B-spline curve by using the Maximum Curvature Points (MCPs) as the Control points. The B-spline curve undergoes a series of smoothening process so features can be extracted. Support vector machine is used to classify the images and the accuracy is 90.00%.

In [8], Pigou used CLAP14 as his dataset [5]. It consists of 20 Italian sign gestures. After pre-processing the images, he used a Convolutional Neural network model having 6 layers for training. It is to be noted that his model is not a 3D CNN and all the kernels are in 2D. He has used Rectified linear Units (ReLU) as activation functions. Feature extraction is performed by the CNN while classification uses ANN or fully connected layer. His work has achieved an accuracy of 91.70% with an error rate of 8.30%.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

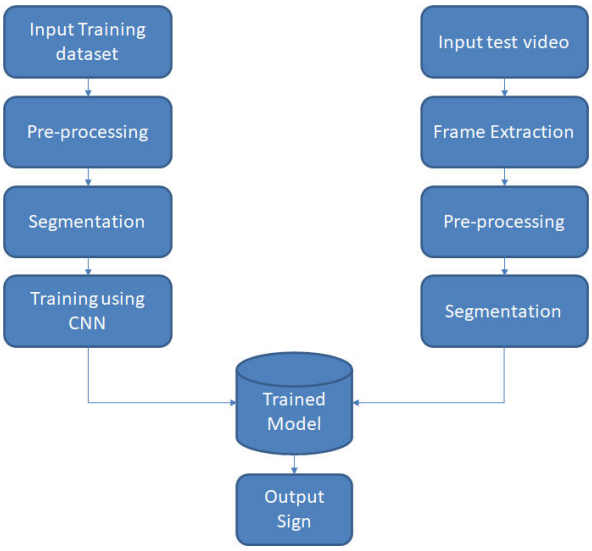
|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|

A similar work was done by J Huang [6]. He created his own dataset using Kinect and got a total a total of 25 vocabularies which are used in everyday lives. He then applied a 3D CNN in which all kernels are also in 3D. The input of his model consisted of 5 important channels which are colour-r, colour-b, colour-g, depth and body skeleton. He got an average accuracy of 94.2%.

III. PROPOSED SYSTEM

Block Diagram:



Input Dataset

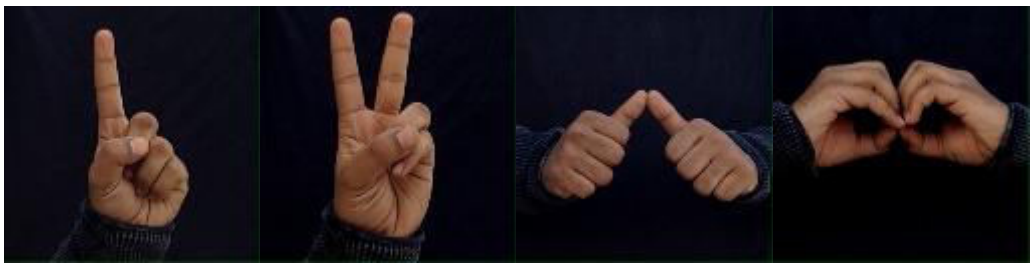


Fig.Sample images of ISL dataset

Preprocessing:

The database images are raw, noisy. Firstly, the images are in the RGB color format. The RGB color is converted into grayscale using the weighted average method. The camera captured images mostly affected by Rician and salt & pepper noise. The median filter is effective in the presence of unipolar and bipolar impulse noise and salt and pepper noise [21].

Segmentation

The segmentation is the important steps for extracting the region of interest. In this approach, thresholding is used to segment the Hand part. The preprocessed images I(x, y) is segmented using thresholding is defined as:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|

$$f_{g(x,y)} = \begin{cases} 1 & I(x,y) > T \\ 0 & else \end{cases} \quad (2)$$

where $I(x,y)$ is the grayscale value of the pixel and $f_{g(x,y)}$is the binary image. If the grayscale pixel value is greater than the defined threshold value then assign value 1 to that pixel otherwise set to 0. Then the threshold image again processed by a morphological operation such as erosion and dilation to get proper boundary and shape. Finally, the binary mask is convolved with the original image.

Classification

In this system Convolutional neural Network and Vgg16 transfer learning algorithm is used to train and test the Indian signs. Each algorithm is explained below.

CNN's are a category of Neural Networks that have proven very effective in areas such as image recognition and classification. CNN's are a type of feed-forward neural network made up of many layers. CNN's consist of filters or kernels or neurons that have learnable weights or parameters and biases. Each filter takes some inputs, performs convolution, and optionally follows it with a non-linearity[. A typical CNN architecture can be seen as shown in Fig.4. The structure of CNN contains Convolutional, pooling, Rectified Linear Unit (ReLU), and Fully Connected layers.

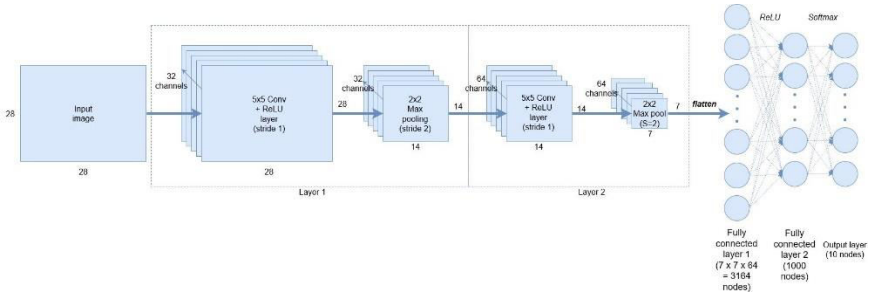
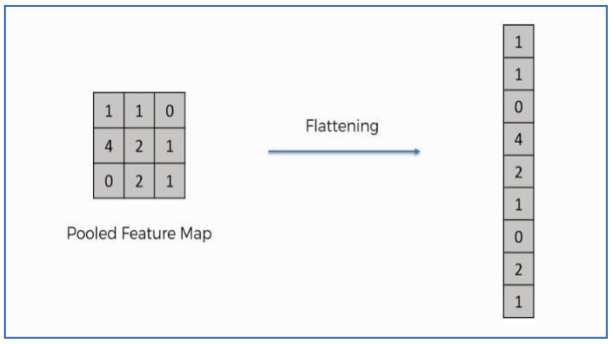


Fig. Architecture of CNN

Flatten Layer

A Convolutional Neural Network takes high resolution data and effectively resolves that into representations of objects. The fully connected layer can therefore be thought of as attaching a standard classifier onto the information-rich output of the network, to "interpret" the results and finally produce a classification result. In order to attach this fully connected layer to the network, the dimensions of the output of the Convolutional Neural Network need to be flattened. After finishing the previous two steps, we're supposed to have a pooled feature map by now. As the name of this step implies, we are going to flatten our pooled feature map into a column like in the image below.



Fully Connected Layer

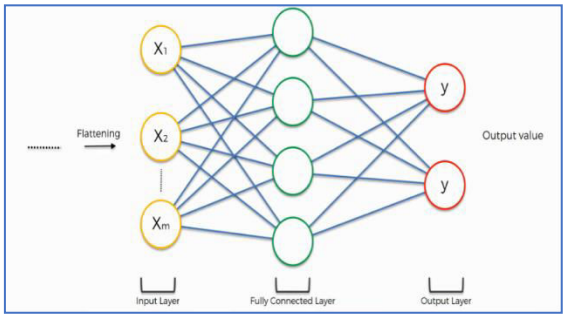
The goal of employing the FCL is to employ these features for classifying the input image into various classes based on the training dataset. FCL is regarded as the final pooling layer feeding the features to a classifier that uses the Softmax activation function. The sum of output probabilities from the Fully Connected Layer is 1. This is ensured by using the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|

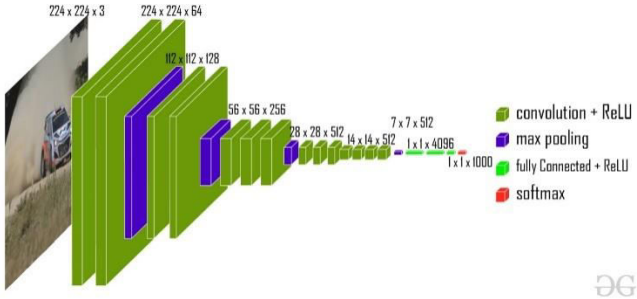
Softmax as the activation function. The Softmax function takes a vector of arbitrary real-valued scores and squashes it to a vector of values between zero and one.



Vgg16 :

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an annual computer vision competition. Each year, teams compete on two tasks. The first is to detect objects within an image coming from 200 classes, which is called object localization. The second is to classify images, each labeled with one of 1000 categories, which is called image classification. VGG 16 was proposed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group Lab of Oxford University in 2014 in the paper "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION". This model won the 1st and 2nd place on the above categories in 2014 ILSVRC challenge.

The architecture of vgg16 model is as shown in Fig.3.8. 13 convolutional layers and 2 Fully connected layers and 1 SoftMax classifier VGG-16 - Karen Simonyan and Andrew Zisserman introduced VGG-16 architecture in 2014 in their paper Very Deep Convolutional Network for Large Scale Image Recognition. Karen and Andrew created a 16-layer network comprised of convolutional and fully connected layers.

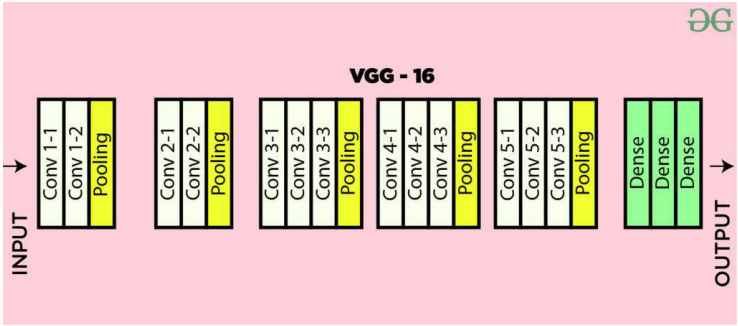


This model achieves 92.7% top-5 test accuracy on ImageNet dataset which contains 14 million images belonging to 1000 classes.The input to the network is image of dimensions (224, 224, 3). The first two layers have 64 channels of 3*3 filter size and same padding. Then after a max pool layer of stride (2, 2), two layers which have convolution layers of 256 filter size and filter size (3, 3). This followed by a max pooling layer of stride (2, 2) which is same as previous layer. Then there are 2 convolution layers of filter size (3, 3) and 256 filter. After that there are 2 sets of 3 convolution layer and a max pool layer. Each have 512 filters of (3, 3) size with same padding. This image is then passed to the stack of two convolution layers. In these convolution and max pooling layers, the filters we use is of the size 3*3 instead of 11*11 in AlexNet and 7*7 in ZF-Net. In some of the layers, it also uses 1*1 pixel which is used to manipulate the number of input channels. There is a padding of 1-pixel (same padding) done after each convolution layer to prevent the spatial feature of the image.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|



IV. RESULT

Qualitative Analysis

The aim of qualitative analysis is a complete, detailed description. No attempt is made to assign frequencies to the linguistic features which are identified in the data, and rare phenomena receive (or should receive) the same amount of attention as more frequent phenomena. Qualitative analysis allows for fine distinctions to be drawn because it is not necessary to shoehorn the data into a finite number of classifications. Ambiguities, which are inherent in human language, can be recognized in the analysis.

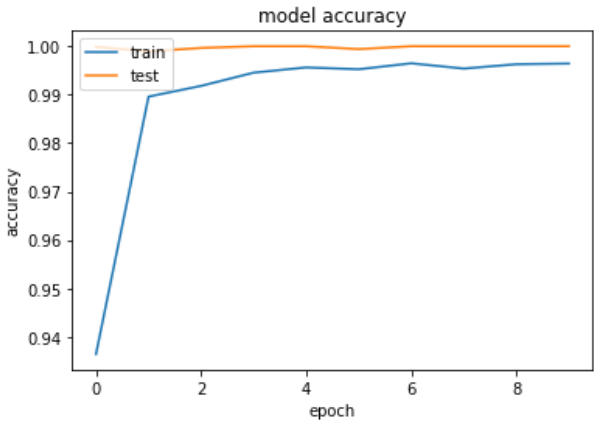
Quantitative Analysis

In quantitative research, the approach is to classify features, count them, and even construct more complex statistical models in an attempt to explain what is observed. Findings can be generalized to a larger population, and direct comparisons can be made between two corpora, so long as valid sampling and significance techniques have been used. Thus, quantitative analysis allows us to discover which phenomena are likely to be genuine reflections of the behavior of a language or variety, and which are merely chance occurrences. The more basic task of just looking at a single language variety allows one to get a precise picture of the frequency and rarity of particular phenomena, and thus their relative normality or abnormality.

The quantitative analysis of the proposed system is calculated using an accuracy parameter. The accuracy of the sign language recognition system is given as (Eq.5.1.)

$$Accuracy = \frac{No\ of\ sample\ correctly\ detected}{Total\ no\ of\ samples}$$

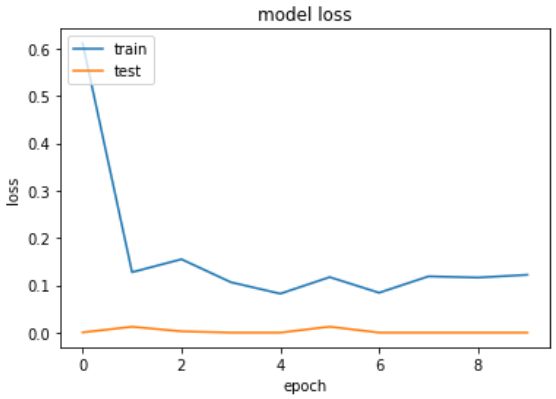
The progress of the CNN algorithm for fake currency detection is given below



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|



V. CONCLUSION

In this project, the recognition of Indian sign language using CNN and Vgg16 algorithm has been presented. The proposed system uses Indian sign language dataset from Kaggle. The dataset consists of alphabets and digit. From the qualitative and quantitative analysis, it is observed that the vgg16 algorithm outperforms than the CNN algorithms in terms of accuracy and execution time. The CNN algorithm achieved a training and validation accuracy of 99.91% and 100%, and loss of 0.0036 and 0.0000003 while Vgg16 achieved a training and validation accuracy of 99.64% and 100% and loss of 0.1221 and 100.

REFERENCES

1. He, Siming. (2019). Research of a Sign Language Translation System Based on Deep Learning. 392-396. 10.1109/AIAM48774.2019.00083.
2. International Conference on Trendz in Information Sciences and Computing Herath,
3. H.C.M. & W.A.L.V.Kumari, & Senevirathne, W.A.P.B & Dissanayake, Maheshi. (2013). IMAGE BASED SIGN LANGUAGE RECOGNITION SYSTEM FOR SINHALA SIGN LANGUAGE
4. M. Geetha and U. C. Manjusha, , "A Vision Based Recognition of Indian Sign Language Alphabets and Numerals Using B-Spline Approximation", Inter- national Journal on Computer Science and Engineering (IJCSE), vol. 4, no. 3, pp. 406-415. 2012.
5. Pigou L., Dieleman S., Kindermans PJ., Schrauwen B. (2015) 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, vol 8925. Springer, Cham. https://doi.org/10.1007/978-3-319-16178-5_40
6. Escalera, S., Baró, X., Gonzàlez, J., Bautista, M., Madadi, M., Reyes, M., . . . Guyon, I. (2014). ChaLearn Looking at People Challenge 2014: Dataset and Results. Workshop at the European Conference on Computer Vision (pp. 459-473). Springer, . Cham.
7. Huang, J., Zhou, W., & Li, H. (2015). Sign Language Recognition using 3D convolutional neural networks. IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). Turin: IEEE.
8. Jaoa Carriera, A. Z. (2018). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on (pp. 4724-4733). IEEE. Honolulu.
9. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 248-255). IEEE.
10. Miami, FL, USA . [13]Soomro, K., Zamir , A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. Computer Vision and Pattern Recognition, arXiv:1212.0402v1, 1-7.
11. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: a large video database for human motion recognition. Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 2556-2563). IEEE
12. T. Bohra, S. Sompura, K. Parekh and P. Raut, "Real-Time Two Way Communication System for Speech and Hearing Impaired Using Computer Vision and Deep Learning," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2019, pp. 734-739, doi: 10.1109/ICSSIT46314.2019.8987908.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005138|

13. Singha, Joyeeta & Das, Karen. (2013), "Recognition of Indian Sign Language in Live Video," International Journal of Computer Applications. 70. 10.5120/12174-7306.
14. H. Muthu Mariappan and V. Gomathi, "Real-Time Recognition of Indian Sign Language," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862125.
15. S. Hayani, M. Benaddy, O. El Meslouhi and M. Kardouchi, "Arab Sign language Recognition with Convolutional Neural Networks," 2019 International Conference of Computer Science and Renewable Energies (ICCSRE), Agadir, Morocco, 2019, pp. 1-4, doi: 10.1109/ICCSRE.2019.8807586.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details