"CNN-BASED SIGN LANGUAGE RECOGNITION SYSTEM USING MACHINE LEARNING"

1.Mandage Nikita 2.Sasane Shreya, 3.Ransing Sakshi 4.Pawar Poonam

5.Prof. Bhosale S.S.

Dept. of Computer Engineering, Hon. Shri. Babanrao Pachpute Vichardhara Trust & Parikrama College of Engineering, Kashti, India.

ABSTRACT

Dynamic head and hand movements, as well as their continually shifting shapes, are regarded a difficult topic in computer vision. Hand segmentation, hand form feature illustration, and gesture sequences recognition are the three key issues in developing an effective sign language system that can distinguish dynamic standalone motions. These methods use colour scheme hand different algorithms to divide hands into segments, hand-crafted features for hand form representations, and Hidden Markov Model (HMM) sequence recognition for traditional sign language identification. a Convolutional neural network (CNN) can be used to recognise Indian sign language motions, according to this article (CNN). It's never easy to have a meaningful conversation with someone who has hearing loss. People with speech and hearing disabilities can use sign language to communicate their thoughts and feelings to the world, making it the ultimate remedy. It facilitates and simplifies the process of integrating them with the rest of society. It is not enough, however, that sign language has been invented. As a result, there are a lot of strings connected. For people who have never learned sign language or who are fluent in a different language, the sign movements can be difficult to decipher. Various strategies for automating the identification of sign motions have made it possible to close the long-standing communication gap.

Keywords: *Convolutional Neural Networks, Sign language recognition, Signer dependent and Signer independent.*

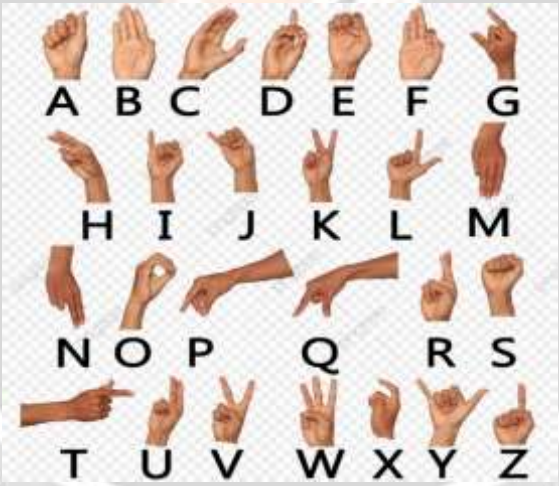
INTRODUCTION:

There is a growing interest in computer vision research in the field of sign language recognition (SLR). Sign camera background modelling, feature representation, and sign classification are some of the issues in SLR. To date, all previous attempts at solving problem have been a great success and have contributed significantly to the advancement of SLR's algorithmic state. The challenge is transformed into a two-dimensional natural language problem for machine translation. There has been minimal progress in bringing a model near to real-time implementation of 1D/2D/3D models that have been proposed in the literature [2]. Indian sign language signs will be recognised using 2D selfies video taken by a smartphone front camera in this project. The goal is to replicate algorithms that perform well on a mobile environment, even though this is a long way off. The major function of this module is to minimise the amount of video data each frame by extracting information frames. The precision and computation time of the [3]-proposed visual attention-based framework were important considerations. We will use the model only for conventional video settings because it works well for them. Selfie mode does not have any benchmark datasets. We were inspired to construct our own dataset because of our interest in Indian sign language (ISL). At a rate of 30 frames per second, five sets of native ISL speakers perform 200 commonly used words in ISL from five different viewing angles (angles dependent on the user). With 3 distinct batch sizes, training can begin at any one of three different times. It's only possible to train a single set of 200 signs in Batch-I; that is, a maximum of 200 1 5 2 30 60000 sign images, each completed by a single user from five different viewing angles over the course of two seconds each. Using a total of 200, 252, 302, 120000 = sign pictures, batch-II of learning is completed. 3 types of sign pictures were used in Batch-III of training. Varied signers and different camera angles are used to test the trained CNNs on two separate video sets. Two different scenarios are tested for robustness. To test various datasets in case-I and to test the same dataset in case-II are two alternative ways of doing so.

CNN ALGORITHM:

One of the most well-known deep learning models is the CNN (Convolutional Neural Network). It is used extensively in computer vision and other neural learning tasks. To limit the number of parameters, using model ever strategy. When training a CNN, the same parameter is used many times, resulting in a network that is not fully connected. Consequently, it's a sensible way to cut training time. Since its conception, CNNs have quickly progressed from image processing to text to speech recognition to become the gold standard in these and many other fields of artificial intelligence.

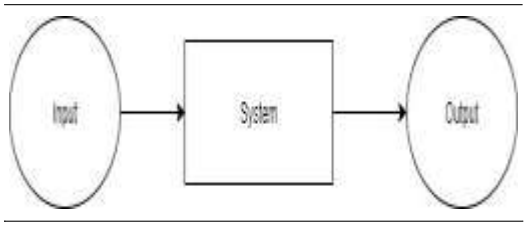
Nelson Mandela [5] said it best: "Speak to a man in a manner he knows, that goes straight to his head. Human communication would not be possible without language, which has existed since the dawn of civilization and is spoken by the majority of the world's population. Humans utilise it as a communication tool to convey their thoughts and ideas, as well as to make sense of the world around them. No literature, no cell phones, and certainly no words I write will have any meaning if it weren't for it. It is so ingrained in our daily lives that we frequently fail to appreciate its significance. It is unfortunate that individuals with hearing problems are often overlooked and left out of society's fast-paced transformations. As a result, they must battle with expressing themselves to those who are different from them. Even though sign language serves as a means of communication for the deaf, it has no meaning for those who do not know the sign language. As a result, the chasm between us has grown wider. We're working on a system that can recognise signs in order to keep this from happening. People with hearing disabilities will be able to utilise it as a means of expressing themselves, and non-sign language users will be able to understand what they're saying. Sign gestures are interpreted in a variety of ways in different countries. For example, a Korean sign language alphabet is not the same as an Indian sign language alphabet. This exemplifies both the diversity of sign languages and their intricacy. A good understanding of gestures is necessary for deep learning to be accurate. Our datasets are created using American Sign Language in our proposed system. As shown in Figure 1, the alphabets of American Sign Language (ASL) can be identified using either approach.



RESWEARCH METHODOLOGY:

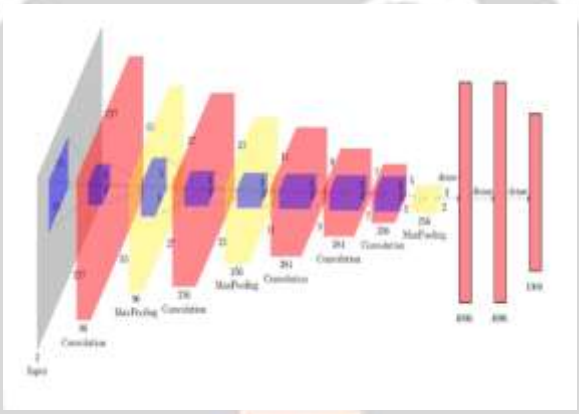
Data collection is the initial phase in the suggested system. Sensors or cameras have been utilised by many researchers to record the movements of the hands. We use a webcam to record the hand motions for our system. Backgrounds can be recognised and deleted using the HSV colour extraction technique through a number of processing steps (Hue, Saturation, Value). The region of skin tone is then detected using segmentation. Using morphological processes, a mask is placed to the images and elliptical kernel dilation and erosion are performed. With open CV, there is no distinction between photos of different gestures because the photographs are all altered to the same size. Images of American sign gestures are included in our dataset; 1600 of them are used for training purpose 400 for testing. 80% of the time, it's 80% of the time. For training and classification, a Deep Neural Network is used to each frame. Finally, the model is tested, and the system is possible to forecast the letters of the alphabet.

Data flow Diagram:



Proposed System:

Two methods for learning spatiotemporal features were used in this study, both utilising a 3DCNN architecture. Feature extraction from the video was done with 3DCNN and classification was done with SoftMax in the first method. There were two ways we tried to improve the link between video frames. As a result, the same 3DCNN framework was learned to extract features from various regions inside the video sample in order to accomplish this. We then looked into several methods for combining features.



Video pre-processing:

During the pre-processing phase, the input video was converted into a sequence of RGB frames. In order to fit the original model to video streams of 16 frames each, linear sampling was used to normalise all the video sequences to a set length of 16 frames. Using the below formula, the relevant indices for each of the 16 frames are determined.

$$index_i = round\left(\frac{len(input)}{16} * i\right), i \in \{1, 16\}$$

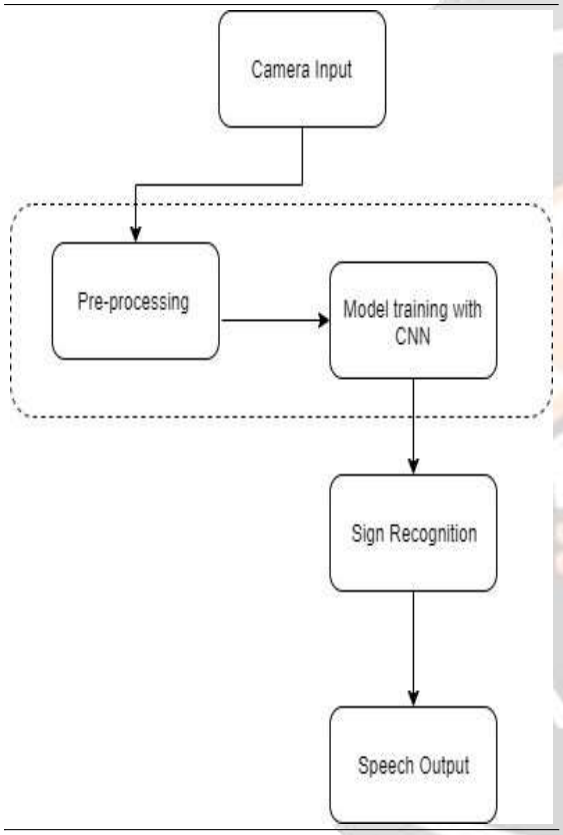
In addition to the Bag - Of - visual Words, other approaches to normalising the input films' temporal dimension have been described in the literature. To retain the sequence of the selected frames, linear sampling was used. The sequence in which the frames were chosen reveals key elements of gesture recognition discrimination. However, the signers' varying heights and distances from the camera need spatial dimension levelling. This normalisation was performed in two stages: During first frame of the sequence, we used a face detection technique to identify a signer's face. - Next, using the actual part of the body ratios to calculate the length and width of the gestures region to be trimmed in all frames, as shown in we used the identified face's position and height as a starting point. All of the input videos were resized to a constant size of 112 112 pixels, with the aspect ratio remaining unchanged, as the final pre-processing step. Each gesture sample's RGB channels were also adjusted then each channels had a mean of zero and a unit variance, as seen in the figure 5. Next, the model was resized and normalised, which lowered the computation complexity and teaching convergence of the model. The features learning phase received 112 112 16 3 volumes as final inputs.

the retrieved features are then fed into a Soft - max for classification. An activation function that output the likelihood of each class is called SoftMax. The most likely outcome is the one predicted.

$$SoftMax(x_i) = \frac{ex_i}{\sum_{j=1}^{k} ex_j}$$

Where xi is the class number, and k is the total number of classes. Training algorithm with inverse log-likelihood is used to fine-tune the suggested model for recognising hand motions. On datasets of movements, the final two blocks only optimised the convolutional and fully connected layers, while the rest of the architecture was frozen. A gradient descent (SGD) algorithm was used to find the best model parameters. The initial learning rate was set at 104, the decay rate was set at 106, and the momentum was set at 0.9. We added 50% dropouts after each fully - connected layers in order to avoid fitting problem and boost the model's generalisation on test data.

Activity Diagram:



Classification:

The retrieved features are then fed into a Soft - max for classification. An activation function that output the likelihood of each class is called SoftMax. The most likely outcome is the one predicted.

$$SoftMax(x_i) = \frac{ex_i}{\sum_{j=1}^{k} ex_j}$$

Where xi is the class number, and k is the total number of classes. Training algorithm with inverse log-likelihood is used to fine-tune the suggested model for recognising hand motions. On datasets of movements, the final two blocks only optimised the convolutional and fully connected layers, while the rest of the architecture was frozen. A gradient descent (SGD) algorithm was used to find the best model parameters. The initial learning rate was set at 104, the decay rate was set at 106, and the momentum was set at 0.9. We added 50% dropouts after each fully - connected layers in order to avoid.

Feature Learning:

The local spatiotemporal properties of gesture sequences can be extracted using a deep 3DCNN, which is recommended for feature learning. The lack of a big labelled dataset of gestures necessitated the use of transfer learning. we began with a 3DCNN structure that has been pre-trained. Millions of human activity recognition data had already been used to train this system [6]. A six-block structure was found after removing the output layer. Each of the first two phases has a single 3DCNN layer. Initially, there are 64 kernels, thereafter there are 128 kernels. There are 256 kernels in each of two 3DCNN layers in the third block of data. In the fourth block, 512-kernel 3DCNN layers are found. Two 3DCNN levels, each with 512 kernels and zero padding, make up the fifth block. The sixth block contains two 4096-neuron thick layers. The feature modelling is globalised with the help of these two layers. Max-pooling is used to move data from block to block in the initial four blocks. With such a stride of one to one, all of the kernels of the 3DCNN are three by three by three in size. Unless you count the initial block's max-pooling kernel that has a stride of 1 to preserve the early stage's temporal information, all max-pooling kernel are of size (2 2 2) with a length of 2 2 2. The activation was performed using a simple non - linear activation unit (RELU), as depicted in (2).

To accelerate up large-network training, this function was chosen since it has a straightforward derivative.

$$ReLU(x) = \{x, x \geq 0 | 0, \quad x < 0\}$$

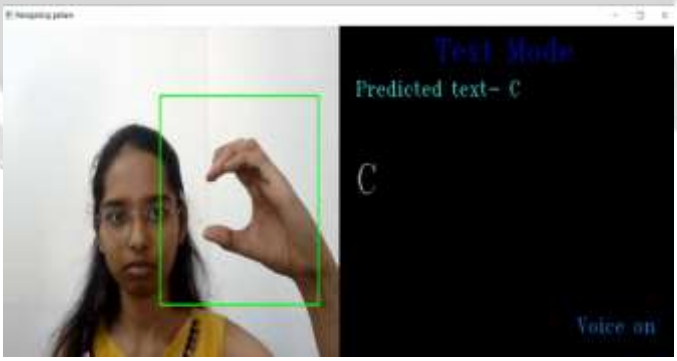
In the first layer, each 3D kernel is convolved. A temporal feature map is created by stacking the 16 incoming stacking frames. A volume of stacked image features produced by the preceding layers is condensed into the 3D kernel in the subsequent levels. An average value is calculated as follows for each coordinate on the Lth layer's Kth feature map:

$$V_{LK}^{xyz} = ReLU\left(b_{LK} + \sum_m \sum_{p=0}^{P_L-1} \sum_{q=0}^{Q_L-1} \sum_{r=0}^{R_L-1} W_{LKm}^{pqr} V_{(L-1)m}^{(x+p)(y+q)(z+r)}\right)$$

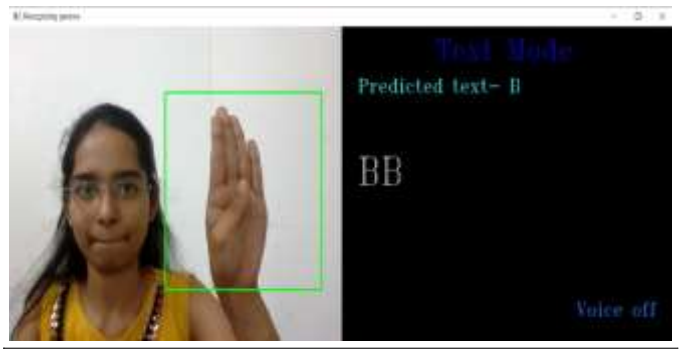
where PL, QL, and RL are the 3D kernel's dimensions and W pqr LKm is the kernel's (p, q, r) the value related to the mth convolution layer in the preceding layer. Each sample is represented as a 4096-character feature vector in the final layer, which is coupled to the other two levels.

RESULTS:

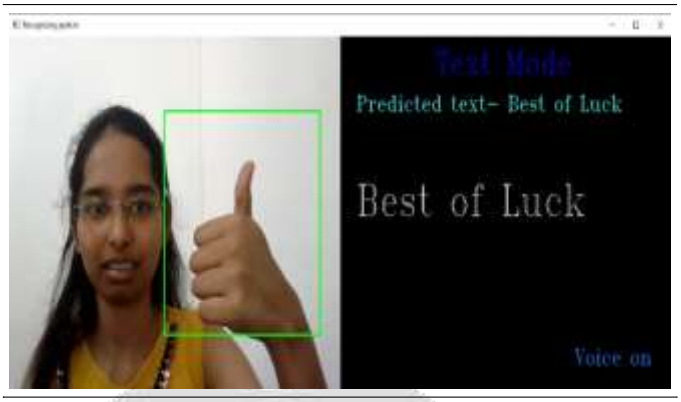
Output 1.



Output 2.



Output 3.



MOTIVATION:

In today's world, interaction is the most crucial aspect of life. People who are deaf or dumb have a particularly difficult time interacting with others since they are unable to interact with them. Thus, this approach was a necessity since people who are deaf or mute need to be able to participate fully in society. Isolation, despair, and social isolation are all symptoms of their underlying weakness. It would be great if they could interact more socially and form stronger bonds with others. Rather than utilizing other languages by deaf individuals, why don't they write and present as a way of communication," one person suggests. This explanation may appear normal and inviting from the viewpoint of a regular human, but the people who are experiencing these challenges require human solutions to their problems. Writing alone isn't enough for many folks; they need to show their emotions and behaviours in person. As a result, we chose to broaden the scope of sign language for still another purpose. For those who are unable to talk or hear, the idea of presenting outcomes in written language helps us to communicate. For those who are deaf or hard of hearing, such an app would certainly make their lives a little easier. A bigger platform will make these folks happier as more and more technologies are adopted and technological advancements are made.

CONCLUSION:

In the area of artificial intelligence, machine learning, and computer vision, there have been numerous advancements. They've made a significant impact on how we see the world and how we use their methods in our daily lives. Different approaches, such as ANN, LSTM, and 3D CNN, have been used to study sign gesture recognition. Most of them, on the other hand, necessitate more processing power. Artificial intelligence (AI) is a potent tool for identifying patterns.

To classify selfie sign language motions, we developed a CNN architecture in this work, which we provide here. Four convolutional layers make up the architecture of the CNN. The faster and more accurate the recognition, the more convolutional layers with varying filtering window sizes are taken into account. It is developed a stochastic pooling strategy that incorporates the characteristics of the both max and mean pooling.

Hand gestures can be recognised using 3DCNNs in this study. Linear sampling was utilised to normalise the temporal ordering of hand motion samples throughout the pre-processing step. We used the ratio of the lengths of the face recognition system and human body parts to normalise the spatial dimension. As a last step, we used 3DCNNs for feature learning in two ways. Hand gesture features were extracted from the entire video using a single 3DCNN implementation. Hand gesture features were extracted from the beginning, endings of the video sample using materials of the 3DCNN structure.

Analyses of the proposed methods were conducted on a variety of data sets. Both in signer-dependent and signer-independent modes, the three datasets performed exceptionally well. Six other cutting-edge strategies from the literature were compared with the ones we proposed. Four of these strategies were outperformed by them, while the other two performed about as well as they did. As part of our ongoing research, we'll conduct a comprehensive search to optimise all of the hyperparameters. Using a live video feed, we'll put the

recommended strategy to the test in real time online. Edge-cloud computing can be used to distribute processing across network edge and the core cloud in this context.

REFERENCE:

[1] Becky Sue Parton. “Sign language recognition and translation: A multidisciplined approach from the ¿eld of arti¿cial intelligence”. Journal of deaf studies and deaf education, 11(1):94–101, 2005.
[2] Zhengzhe Liu, Fuyang Huang, Gladys Wai Lan Tang, Felix Yim Binh Sze, Jing Qin, Xiaogang Wang, and Qiang Xu. “Real-time sign language recognition with guided deep convolutional neural networks”. In Proceedings of the 2016 Symposium on Spatial User Interaction, pages 187–187. ACM, 2016.
[3] Mukul Singh Kushwah, Manish Sharma, Kunal Jain, and Anish Chopra. “Sign language interpretation using pseudo glove”. In Proceeding of International Conference on Intelligent Communication, Control and Devices, pages 9–18. Springer, 2017.

[4] S. Boutadghart, "Sign Language Digits Recognition using Deep CNN", *Salihbout.com*, 2022. [Online]. Available: https://www.salihbout.com/cnn-signs/. [Accessed: 03- May- 2022].
[5] K. Bantupalli and Y. Xie, "American Sign Language Recognition using Deep Learning and Computer Vision," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 4896-4899, doi: 10.1109/BigData.2018.8622141.
[6] CABRERA, MARIA & BOGADO, JUAN & Fermín, Leonardo & Acuña, Raul & RALEV, DIMITAR. (2012). GLOVE-BASED GESTURE RECOGNITION SYSTEM, doi: 10.1142/9789814415958_0095.
[7] S. Kausar and M. Y. Javed, ‘‘A survey on sign language recognition,’’ in Proc. Frontiers Inf. Technol., Islamabad, Pakistan, Dec. 2011, pp. 95–98.
[8] M. B. Waldron and S. Kim, ‘‘Isolated ASL sign recognition system for deaf persons,’’ IEEE Trans. Rehabil. Eng., vol. 3, no. 3, pp. 261–271, Sep. 1995.