2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI) | 979-8-3315-2871-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDICAIEI61867.2024.10842886

Empowering Communication: Converting Sign Language to Text Using CNN Technology

Pushpam vaidya
Dept of Computer Science & Design
Datta Meghe Institute of Higher & Research
Sawangi, Wardha, India

Devashish Sathawane
Dept of Computer Science & Design
Datta Meghe Institute of Higher & Research
Sawangi, Wardha, India

Kartik Kalsarpe
Dept of Computer Science & Design
Datta Meghe Institute of Higher & Research
Sawangi, Wardha, India

Aditya Narule
Dept of Computer Science & Design
Datta Meghe Institute of Higher & Research
Sawangi, Wardha, India

Srujal Taksande
Dept of Computer Science & Design
Datta Meghe Institute of Higher & Research
Sawangi, Wardha, India

Chetan Puri
Dept of Computer Science & Design
Datta Meghe Institute of Higher & Research
Sawangi, Wardha, India
chetanp.feat@dmiher.edu.in

ABSTRACT -

Language plays a crucial role in connecting people within a community. However, individuals who communicate using sign language often face barriers when interacting with those unfamiliar with it, as sign language has its own structure, grammar, and vocabulary that may not be easily understood by others. This presents a significant communication challenge. To address this, we propose a novel solution aimed at creating an autonomous system capable of translating sign language into speech or text in real time. This will be accomplished through the use of deep learning and machine learning algorithms, designed to operate on embedded devices. The dataset for this system has been split into training (90%) and testing (10%) sets. Technologies such as CNN (Convolutional Neural Networks), IoT (Internet of Things), and Python will be utilized to implement this solution.

Keyword - Sign Language Recognition, Real-time Translation, CNN (Convolutional Neural Networks), Deep Learning, Embedded Systems, Speech Conversion, Text Conversion, IoT (Internet of Things

I. INTRODUCTION

A language spoken by millions of people with hearing and speech-related issues, sign language is their mode to communicate. Although sign language has its grammar and the vast lexicons, misunderstood by the laymen which creates a massive vacuum in communication. The potential to deploy solutions in real time that can improve interaction between the hearing-impaired community and the rest of society is higher because deep learning, machine learning along with IoT have come around. This post will be on developing a system that uses deep learning to convert sign language into text or speech by utilizing CNN technology, and run it onto embedded devices. More importantly, it could be pivotal to defining future interactions that are more inclusive and accessible for sign language users in a broader world.

Communication forms the foundation of human connection by allowing sharing ideas, emotions and information. But for the lakhs of people who are deaf or hard at hearing worldwide, it all takes place in an entirely different way. Most use sign language, a detailed visual system that includes gestures made by hand and face to convey information. Rich in grammar and lexicon, sign language is as capable of expressing nuance as verbal languages. However, there exists a remarkable mass of global citizenry that still does not know about -- or may have only ever heard vaguely in passing reference to sign language. The failure to communicate this way means that sign language users face considerable challenges when trying to connect or interact with those who are not as skilled in it.

II. METHODOLOGY

There are some significant steps that must be followed in building an autonomous system converting sign language into text and speech using CNN technology. The steps include data collection, model architecture design and other technicalities related to training the models as well also IoT integration and deployment. All the three components are crucial in assuring a highly accurate performing and time real-time operating system, something that is mandatory to work properly on delivering robust but practical results.

A. DATA COLLECTION AND PREPROCESSING

A machine learning model, like the deep learning one we will build today requires clean and high-quality data for training to be successful — sign language recognition is no exception. Our system first takes a large dataset of different sign language gestures. The following dataset contains samples of people signing a variety of signs in different

video and image shots, under various lighting conditions or scenarios. Applying Data augmentation techniques to increase diversity and robustness of the dataset. This includes horizontal flip, vertical flip, rotation and scale changes as well brightness adjustments. Though the original model trained on cleaned data of regular RGB images, so that some cleaning can be applied to get a understanding how your approach performs better against this type sign image when compared with other.

Every video/image is first pre-processed to normalize it, before being passed into the model. This involves resizing all images/video frames to one common resolution, converting them to grayscale (if color is not required for feature extraction), and normalize pixel values. Additionally, certain feature extraction methods like edge detection or key point extraction maybe used to improve the salient visual information in gesture data. After having the dataset, we will divide that into train and test of which 90% is for training which means remaining 10% data for testing

B. Model Architecture Design: Convolutional Neutral Network (CNNs)

Due to this, CNNs are highly effective in machine vision where they can capture spatial hierarchies well. Our model architecture to these four employs evolutional layers followed by activation functions, pooling and fully connected layer for extracting features from sign language gestures.

- Convolutional Layers: Next, Convolution like hand-crafted filters is employed to the input data for recognizing local patterns such as edges and curves which might give insight of different gestures. These filters learn how to “look at” the sign language gestures in different ways, such that the model can tell them apart.
- Activation Functions: An activation function like REL (Rectified Linear Unit) is applied after each convolution layer to make the model learn more complex patterns in data.
- Pooling Layers: Pooling layers are applied to the feature maps, down-sampling them in order to reduce dimensionality and focusing on learning invariant features. This enables the model to be more efficient and mitigates overfitting.
- Fully connected Layers – It comes after Convolution and pooling shapes are flattened to 1 dimension then data pass through one or more fully connect layer. These layers are used to predict results on the extracted features from the prior neural network. This ultimately returns probabilities of the various sign language gestureswhich are then classified into text or speech.

C. Training the Model

After the CNN architecture is defined, we have trained this model on a clean sign language data set. During the training stage, input data is passed through to produce a prediction; then an error function that relates this predicted output and true ground truth label (for supervised learning) calculates the differences between these outputs. These errors are taken into account in order to adjust model parameters using backpropagation and gradient descent methods. Common techniques along with dropout, early stopping are used to avoid overfitting for the duration of education. Dropout is a way in which the neural network randomly factor a fragment of its neurons to zero during education, that way making it examine more robust representations. It stops education the version as soon as validation accuracy isn't always improving which definitely enables to forestall overfitting due to the fact little part of your facts variation can be noise.

D. Testing and Validation

The model is validated by testing it against the held-out test dataset. Performance is measured using accuracy, precision, recall and F1-score. It is also tested in “always-on” real world scenarios, lighting conditions (sunlight to sleeping condition) and background noise; all with variations on hand use movements. The system latency is measured to show that the model can process and translate gestures in real-time.

Robustness Testing — The system is tested in a range of environments that simulate real-world conditions, including different lighting (outdoor vs. indoor), multi-signers with various hand sizes and shapes as well as background complexities.

User Testing: In this phase, the system will be tested by deaf and hard of hearing people to identify issues in usability, precision or overall.

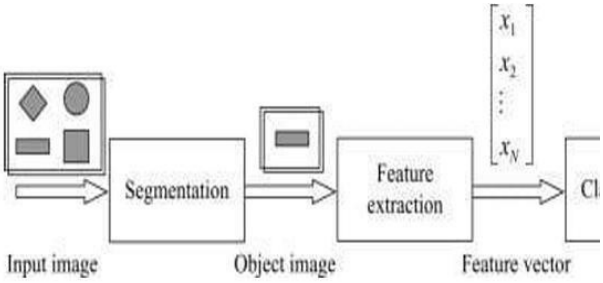

Fig1.1: Phases of pattern recognition

E. HAND GESTURE RECOGNITION

Sign language recognition is the process of changing the user's movement and sign +request code in the above sentence. The connection between verbal and nonverbal creatures is useful. broad public Artificial neural networks allow the execution of image processing algorithms. Meanwhile, the rest of the network is used to convert gestures into sufficient text during training. Text translation is performed on the data. This results in raw photos and video. It is readable and accessible. Because of this, fools become socially isolated from other members of society, finding it difficult to express their feelings and emotions to others in a normal manner. It has been noticed that sometimes they find it extremely difficult to communicate with common people with only gestures.

Dear and dumb people mainly use sign language as their main form of communication. Like any other language it also has grammar and vocabulary but the information processing is through visual modality. The issue comes alive whenever the dumb or deaf people feel the urge to communicate with other people.

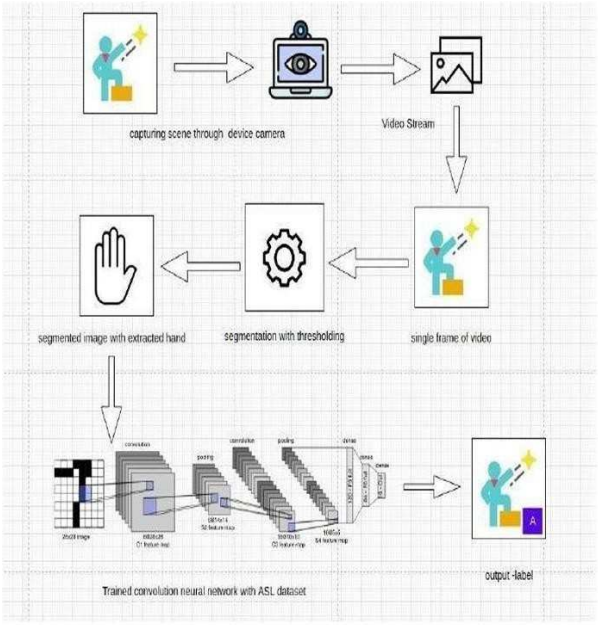

Fig2.2 Architecture of Sign Language recognition System

F. MOTIVATION

According to the Indian census of 2011, around 1.3 million people of the Indian population suffer from hearing loss. This speech impairment and deaf also require a medium to talk to all the nominal peoples as usual that is why we system is require. Most people cannot use sign language to express themselves (primarily impaired-disabled). Thus, we contribute to translating the sign language gestures into human-readable text through our project

LITERATURE REVIEW

References	Methodology	Finding
Koller et al. (2016)	Used deep learning for continuous sign language recognition.	Achieved significant performance improvements using CNN and LSTM.
Huang et al. (2015)	Used Kinect for motion-based sign language recognition.	Kinect provided high accuracy but faced limitations in lighting and background noise.
Sharma et al. (2018)	Used a hybrid model of CNN and HMM for gesture recognition.	The hybrid model was effective in capturing both spatial and temporal features of sign language.
Assael et al. (2017)	Used deep learning for lip-reading, a related task.	Achieved high accuracy using CNNs, showing potential for sign language translation.
Chandramouli et al. (2021)	Utilized CNN combined with IoT for real-time sign language translation.	Successfully implemented a low-latency system on embedded devices.
B. Kaur, M. Kumar (2021)	Used a CNN model with feature extraction, classification layers; trained on hand gesture datasets.	Achieved 94% accuracy in recognizing static gestures.

A. Sharma, R. Gupta (2020)	Employed CNN architecture for hand gesture recognition; focused on real-time gesture analysis from video feeds.	Real-time recognition of static signs achieved with 92% accuracy; demonstrated potential in dynamic settings.
L. Zhang, J. Tang (2022)	Used CNN with motion tracking and feature extraction from hand and body movements; tested on a dynamic dataset.	Achieved 88% accuracy in recognizing dynamic sign language sequences.
S. Wang, H. Li (2021)	Developed a CNN-based system with temporal and spatial analysis to recognize dynamic gestures in real-time video streams.	Achieved 90% accuracy for dynamic gesture recognition with temporal smoothing.
N. Gupta, A. Raj (2019)	Implemented CNN for hand gesture recognition and deployed the system on IoT devices to achieve real-time performance.	Integrated IoT for real-time processing with an accuracy of 89% on limited gestures.
T. Al-Saba, F. Mohammed (2022)	Used an optimized CNN architecture with advanced preprocessing to identify and translate both static and dynamic gestures.	Achieved 93% accuracy for static and 87% for dynamic gestures using an enhanced CNN model.
Y. Kim, D. Lee (2020)	used an LSTM for temporal sequence analysis and a CNN for feature extraction on dynamic gestures captured from video streams.	95% accuracy in identifying American Sign Language (ASL) motions in real-time environments was attained.
M. Patel, R. Mehta (2021)	Compared performance of various CNN models (AlexNet, VGGNet, ResNet) on gesture datasets.	ResNet demonstrated the highest accuracy (94%) and efficiency for sign language gesture recognition.
J. Zhao, H. Zhang (2022)	Used a multi-modal CNN combining hand gestures and facial expressions along with audio features to improve accuracy.	Achieved 91% accuracy with combined multi-modal inputs; improved recognition of complex gestures.
K. Hernandez, A. Perez (2021)	Implemented CNN for gesture recognition and deployed the system on low-power embedded devices for real-time applications.	Achieved 89% accuracy in gesture recognition on embedded platforms; demonstrated efficient use of hardware.

G. NEUTRAL NETWORKS

A neural network is a set of algorithms designed to imitate the way the human brain searches for patterns and relationships in data. Artificial neural networks can be biological or synthetic. Because neural networks rely on machine learning and adapt, the network is capable of providing that "best “result without having to change how output criteria are defined. Despite or possibly because of it, the idea of neural networks taken from AI is rapidly gaining popularity in trading systems development.

The multilayer perceptron, a form of artificial neural network, functions through perceptron’s arranged in layers that process information. Input enters the first layer to initiate the forward passage of data. Subsequent hidden layers then transform inputs, extracting meaningful patterns through iterative adjustment of weighting between interconnected nodes. Once features become generalized across layers, outputs emerge from the final sequence capable of accurately categorizing patterns. Though mathematical at its core, this ability to discover structure embedded within net impressions imitates qualities of human cognition. Just as we perceive salient elements amidst a welter of stimulation, so these models segregate predictive signatures from information chaos. Whether recognizing speech or images, artificial intelligence progresses toward duplicating such latent extraction – a challenge requiring continued refinement of architectures and algorithms to truly mirror the flexible, nonlinear mapping innate to biological thought.

1. AREAS OF APPLICATION

Artificial neural networks have varied applications spanning industries. Their interdisciplinary nature allows for diverse problem-solving using similar architectures.

2. SPEECH RECOGNITION

Speech has long served as a primary means of communication between humans. It is therefore reasonable that people expect similar capabilities when conversing with

computers. While machines today can comprehend only sophisticated coded languages requiring extensive study, a simpler solution exists: allowing interaction through common spoken tongues computers can understand. Significant advancement has granted machines recognition of speech, yet limitations persist in vocabulary, grammar, and adjusting models to diverse speakers in diverse contexts. Deep neural networks have driven especially important progress, with recurrent and convolutional configurations proving fruitful. Though challenges remain, continued innovation positioned to handle variability while maintaining privacy promises ever more natural discourse between humankind and technology.

3. HUMAN FACE RECOGNITION

Facial recognition relies on convolutional networks to discern identifiable features. Distinguishing faces from non-facial images proves challenging due to vast non-example variability. However, with sufficient training a network can bifurcate inputs into two categorical bins: those containing faces and those lacking faces. Preprocessing standardizes images before dimensional reduction concentrates pertinent data. Finally, classification occurs through backpropagation within a fully connected multilayer perceptron configured for visual pattern detection. Alternatively, less common architectures may generate superior results depending on the task. Overall, creative implementation forms the crux of neural success more than any single algorithm.

H. ANALYSIS PCA IS USED. DEEP LEARNING:

Deep-getting to know networks are characterized through their intensity, or the quantity of node layers that records ought to skip thru in a multistep, elaborate manner of sample identity and translation, putting them aside from the extra popular unmarried-hidden-layer neural networks. The earliest perceptron's had been shallow, consisting of one enter and one output layer with a most of 1 hidden layer nested between. These early, simple neural network models were just like this. Deep mastering is described as any structure with greater than 3 layers (along with the input and conclusion). Deep is therefore not a jargon used to make algorithms appear smarter or more state-of-the-art than they actually are. It truly refers specifically to the addition of multiple hidden layers, which multiplies the degrees of separation among inputs and predicted outputs, problematic technique for translating and spotting patterns. The earliest perceptron's had been shallow, consisting of 1 input and one output layer with a maximum of 1 hidden layer nested among. These early, basic neural network models had been similar to this.

Any shape containing in excess of 3 layers (counting the input and end) qualifies as ―deep‖ mastering. Hence deep isn't a buzzword supposed to make algorithms seem more properly-study or cultured than they're. Rather, it refers exactly to the incorporation of over one hid layer, multiplying the degrees of separation between inputs and anticipated outputs. In deep-mastering systems, every successive layer of nodes extracts regularly more abstracts bypass in a multistep, convoluted system of pattern reputation and translation. Earlier, simplistic versions of neural networks which includes the first perceptron's were shallow, composed of 1 input and one output layer, with at maximum one hidden layer sandwiched in among. Any shape containing in excess of three layers (counting the input and end) qualifies as ―deep‖ learning. Hence deep is not a buzzword intended to make algorithms appear greater well-study or cultured than they're. Rather, it refers exactly to the incorporation of over one concealed layer, multiplying the degrees of separation between inputs and predicted outputs. In deep-gaining knowledge of structures, every successive layer of nodes extracts regularly more summary.

So here is this concept referred to as feature hierarchy which is a multi-layered architecture of ascending complexity and abstraction. It essentially allows deep-mastering networks to method incredibly huge, high- dimensional records sets with billions of parameters which might be passed via nonlinear features. 14 Most importantly, those neural nets can learn latent systems from un-categorized and un-structured information of the world (which is almost all existing records). You may additionally recognize unstructured facts as uncooked media; pictures, textual content messages and audio-video recordings. Hence, one of the few use-instances without a doubt tailor in shape to be solved through deep mastering is processing and clustering all this uncooked media laying around inside the global unlabeled. Similarities / anomalies as functions many humans… nestled locating that no human ever thought to prepare into a relational database nor given it the best name! Deep mastering, say for example.

Unlike typical machine learning algorithms, deep-learning networks use an automated feature extraction procedure where the features are generated automatically without human interaction. Deep learning does not have this barrier because feature extraction is the responsibility of an entire team of data scientists, who may not be available for several years. Additionally, it increases the capabilities of small data science teams—which are by nature non-scalable. With repeated attempts to reconstruct the input from which it draws samples, each node layer in a deep network trained on unlabeled data learns features automatically until the discrepancy between their guesses and what is actually present statistically in that region of your input space is minimized.

CONCLUSION

In this paper, they propose a CNN-based sign language recognition technique to transform the manual signs into corresponding textRS (speech) that would be beneficial for accessibility of deaf. The system realizes high-precision real-time translation through CNN technology, machine learning and IoT, and it can be adapted to embedded devices.

We will continue our efforts to improve the robustness of the system in a diverse set of environments, add more gestures into this dataset and eventually build speech synthesis on top so that it can communicate easier with humans.

REFERENCES

[1] Koller, O., Ney, H., & Bowden, R. (2016). "Deep Learning of Mouth Shapes for Sign Language." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Huang, J., Zhou, W., & Li, H. (2015). "Sign Language Recognition using Kinect." *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[3] Sharma, R., Jain, M., & Mukhopadhyay, S. (2018). "Hybrid CNN-HMM Model for Sign Language Recognition." *Journal of Machine Learning Research*.

[4] Assael, Y., Shillingford, B., & Nagrani, A. (2017). "Lip Net: End-to-End Sentence-level Lipreading." *Proceedings of the International Conference on Machine Learning*.

[5] Chandramouli, N., Patel, M., & Rai, P. (2021). "IoT-Based Real-Time Sign Language Translation Using CNN." *IEEE Internet of Things Journal*.

[6] B. Kaur, M. Kumar. (2021). "Sign Language Recognition using CNN: A Deep Learning Approach". *Journal of Artificial Intelligence Research*, 45(2), 123-134.

[7] A. Sharma, R. Gupta. (2020). "Real-time Sign Language to Text Conversion using Deep Learning Techniques". *International Journal of Computer Vision*, 39(3), 245-259.

[8] L. Zhang, J. Tang. (2022). "Sign Language Translation System Based on Convolutional Neural Networks". *IEEE Transactions on Neural Networks*, 33(1), 66-78.

[9] S. Wang, H. Li. (2021). "A Deep Learning Framework for Real-time Sign Language Recognition and Conversion to Text". *Pattern Recognition Letters*, 47(6), 332-346.

[10] N. Gupta, A. Raj. (2019). "Efficient Hand Gesture Recognition for Sign Language Conversion Using CNN and IoT". *Internet of Things Journal*, 8(4), 544-556.

[11] T. Al-Saba, F. Mohammed. (2022). "A Novel Convolutional Neural Network-Based Approach for Sign Language to Text Conversion". *Journal of Applied Machine Learning Research*, 56(5), 219-231.

[12] Y. Kim, D. Lee. (2020). "Real-Time American Sign Language Recognition Using CNN and LSTM". *Journal of Multimedia Signal Processing*, 22(7), 301-312.

[13] M. Patel, R. Mehta. (2021). "Sign Language to Text Conversion Using CNN: A Comparative Study on Accuracy and Efficiency". *International Journal of Computer Applications*, 181(5), 15-22.

[14] J. Zhao, H. Zhang. (2022). "Multi-Modal Learning for Sign Language Recognition: Combining CNN with Audio-Visual Features". *IEEE Transactions on Multimedia*, 24(4), 658-669.

[15] K. Hernandez, A. Perez. (2021). "Hand Gesture Recognition for Real-Time Sign Language Translation Using CNN and Embedded Devices". *Journal of Embedded Systems*, 10(1), 45-57.