Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 6, No.5, June : 2024

SIGN LANGUAGE RECOGNITION USING DEEP LEARNING: AN OVERVIEW

Shubha Chaturvedi, Research Scholar, Sage University, Indore
Dr. Manoj Ramaiya, Sage University, Indore

ABSTRACT
Speaking is how most people communicate with each other and express themselves. Speech poses a considerable challenge for individuals with certain disabilities, such as those who are hearing-impaired, mute, or nonverbal. Such individuals those are hearing-impaired they usually communicate through sign language or hand gesture as their only means of communication with others. Language ensures these individuals are included in spoken communication groups using a system of gestures and visual clues. But it's also critical to understand these people's language semantics and gestures. To close this gap, the field of sign language study has seen a noticeable expansion. In this study, we emphasize the importance of translating and interpreting sign language and provide an extensive review of relevant research conducted in this field. We analyze from several perspectives, such as sign language recognition, and dataset accessibility. Our goal is to make a positive impact on the field of sign language literature and its practical applications by investigating these facets.

Keywords: Recurrent Neural Network (RNN) **,** You Only Look Once (YOLO),Image Recognition , Feature Extraction , Pointnet.

Introduction
The hand plays an essential function as an interaction medium in human-computer interaction (HCI) [1]. Forms of communication can be broadly classified as linguistic and non-linguistic. A hand motion is a form of body language or physical communication that conveys a message through motion of body language or physical communication through different actions of body [2]. The most frequently used part of physical communication to create gestures that helps in commutating is by the motion of various body gestures which mainly includes the hand gesture. Hand gestures come in two varieties: static or fixed and dynamic or moving. Static or stationary hand gestures describe postures where the hands are in stable shapes, and dynamic hand gestures consist of a sequence of images. Visual communication in chaotic situations is a key task for various automated image analysis applications [3], including virtual reality, touch less or gesture-based interaction, sign or gesture language processing, hand activity analysis, and hand-movement surveillance for drivers. This research's primary objective [4] is to pinpoint the difficulty in identifying and detecting static or fixed hand gestures involves recognizing when the body poses or finger of hand takes specific positions to convey specific message to intended recipient. Reconstructing the body pose or position of figure in hand is particularly challenging due to the wide range of configurations of postures and angles in relation to the image or video sensor, which contribute to the high level of complexity of this challenge. Otherwise, recognizing static hand gestures is essential for numerous applications, including hand gesture commands and driver hand monitoring to reduce driver attention, SLR for hearing impaired. SLR is vital for the deaf and speech-challenged communities, among others. As for the interaction required for the system and for that purpose there is need of the exact location on the finger tips [28].

Understanding hand gestures is just as important as understanding sign language [5]. People with hearing impairments often face challenges in fulfilling basic human needs such as learning, writing, instructing, interacting, and literacy, which may not come naturally to them. People connect with one other all across the world using a variety of communication methods. Sign language is one of these communication methods.

Individual with hearing impairments frequently utilize body posture sign language as a natural means of communication. Occasionally, technology needs to be updated to recognize sign language in order to enable communication among individuals with hearing impairments and the general public [6]. The technology known as sign language recognition enables a computer to identify the sign language used by the signer and, with the aid of certain algorithms, convert it to text. Spoken languages are communicated using hand gestures, body language, and facial expressions rather than words or written words. Sign languages are unique in their visual presentation and interpretation of concepts, setting them apart from other languages, as they rely on visual rather than auditory communication. This also creates a unique language barrier.

Literature Review

Body gesture or hand key point movements can be recognized in several data sources, such as images or real time motion video and wearable sensors. Hand gesture recognition research has evolved. In early ages or in the beginning of this era, sensor gloves, Illuminated markers, or similar devices were employed. Hand classification is a difficult task that requires stable illumination conditions for precise findings. Hand motions can be detected using several parameters such as skin tone and velocity, as well as articulated models, spatiotemporal descriptors, and trajectory information. Deep feature-based techniques have been pivotal in propelling the success of convolutional neural networks, sparking innovation across diverse domains, leading to the development of various object identification and recognition algorithms. Here, a study has been done on the various techniques

In a recent study authors, Mariusz Oszusta ,Jakub Krupskia[6] introduced a method that segments uses gestures of sign language in smaller regions , or cells, and utilizes numerical metrics to characterize these regions. To effectively compare gestures of varying lengths, DTW technique is often paired with the nearest neighbor (NN) rule. This approach aligns and matches temporal sequences, allowing for accurate recognition of gestures even when their durations vary. DTW handles the variability in timing, while the NN rule ensures robust classification by comparing each gesture to the closest example in the training set. This combination enhances the performance of gesture recognition systems by accommodating differences in gesture execution speedNevertheless, the computational demands of DTW pose a limitation on the classifier's practicality due to its time-consuming nature.

Here another author Radu Mirsu et.al.[7]proposed an approach where 3D time of flight data used and PCA is used for segmentation algorithm which provides the normalized data and neighbourhood decision can be made on histogram with the architecture of Pointnet.

In another work of Jyotishman Bora et.al. presented a novel dataset preparation method by using fusion of bidimensional and tridimensional Assamese gestures. The MediaPipe framework was utilized for landmark detection in these images. So this initiative leads to the development of an Assamese Sign Language dataset containing more than 2000 data points. It includes 9 stationary gestures representing various vowels and consonants.

Yanqiu Liao et.al. [9] introduces a multimodal dynamic or changing SLR method called the BLSTM-3D Residual Network (B3D ResNet). This approach leverages a deep three -dimensional Residual ConvNet and bi-directional LSTM networks to automatically extract spatial-temporal features from moving images or video sequences. After feature extraction, B3D ResNet assigns intermediate scores or marked values to each action or gesture within the image or video sequence, enhancing the accuracy of SLR by analyzing both spatial or physical and temporal or sequential aspects of the gestures.

In another work of Yulius Obia et.al. [10] introduce an application capable of real-time language recognition and conversion. This research utilizes ASL datasets in combination with a CNN classification system. The process starts by filtering the hand image, which is then fed into the classifier. The CNN processes the filtered image to predict the class of the hand gestures, enhancing the accuracy and efficiency of SLR.

Razieh Rastgoo et al. [11] conducted an analysis of visual or optical SLR models utilizing DL methodologies within the past 5 years. Here, the analysis of various models represents an improvement

in recognition accuracy for SLR, more research is needed to address remaining challenges and optimize performance further.
In the recent work, Sumaya Siddique et al. [12] employed Detectron2, EfficientDet-D0 with TensorFlow, and PyTorch-based YOLOv7 models for implementation. The project also implement the Jetson Nano, a compact and powerful NVIDIA microprocessor, to create an object detection model capable of inferring from test images in real-time.
Qazi Mohammad Areeb et al. [13] introduce combination of VGG-16 and LSTM model as classification model for ISL data set with with approx 98% accuracy and YOLO v5 as detection model with almost 99.6% MAP value. Sign language serves as an emergency communication tool for deaf people, assisting them in dealing with challenging situations.
Muneer Al-Hammadi et.al[14] introduces an effective architecture for SLR utilizing a GCN. The architecture features a series of separable 3DGCN layers, augmented by a spatial attention mechanism. By incorporating only a few layers, this proposed architecture sidesteps the common issue of over-smoothing encountered in DGNN. Attention mechanism plays here a major role for representation of the gestures or hand signs.
Bayan Ibrahimm et.al.15] developed a method to process dynamic hand movements by converting them into frames, removing noise, and adjusting intensity for feature extraction. The method identifies hand gestures, generates a skeleton using mathematical calculations, and extracts features such as joint color cloud, and many more other specialized feature. These features are fine-tuned, and a subset is fed into a RNN classifier for improved accuracy in classification results.
Guoyu Zhou, Zhenchao CuiJing Qi [16] presents BLSNet, a tri-branch portable sub-division network for gesture sub-division. It uses three specialized branches to capture neighborhood characteristics or regional features, boundaries, and contextual hand properties. The edge segment features a MDSC module to enhance contour precision, while a boundary weight module refines boundaries. The semantic branch employs continuous downsampling for accurate hand localization in complex backgrounds. The Ghost bottleneck is used throughout BLSNet as the core building block for efficient and precise segmentation.
Zhen Liu et.al. [17] used a attention mechanism of spatial nature to pick images of different exposure whereas of lower dynamic range and also aligns images at feature level using various alignment module such as Pyramid etc.
Tuan et.l. [18] has devised a various components of gesture recognition system where Feature Extraction module is mostly used for extracting keypoint or human poses , as well as hand bounding boxes. The extracted features are then standardized and refined within the transformation module. This transformed data is fed into the Classification Module, which uses an cutting-edge two-pipeline architecture for gesture classification

Methods &Material
The methods and resources utilized in this study to achieve the gesture tracking and classification, which is the focus of this paper, are outlined in this section. Here this Figure.1 represents the flow of the SLR Model

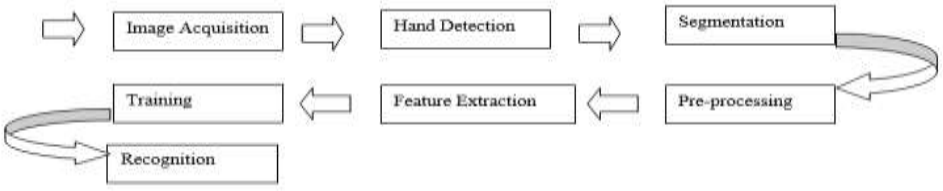

Figure 1: Hand Gesture Recognition system
i) Image Acquisition: Image acquisition entails recording video frames or photos that include sign language motions. This can be accomplished using a variety of devices, including webcams, depth sensors (e.g., Kinect), and specialized cameras. In real-time systems, image collection is continuous to capture a stream of input data, allowing for dynamic detection of sign language gestures.

ii) Hand detection: Hand detection seeks to locate and identify regions of interest that correspond to hands within collected pictures or video frames. HOG, Convolution Neural Network (CNN), You Only Look Once various versions, PointNet are all techniques that can be used to detect hands accurately and efficiently.
iii) Segmentation: Segmentation separates the hand region from the backdrop to allow for further analysis and processing. Thresholding, edge detection, and Semantic segmentation is applicable for segmenting the hand area based on colour, texture, or depth data.
iv) Pre-processing: Pre-processing processes improve the quality of the incoming data and prepare it for feature extraction and recognition. Noise reduction, normalization, scaling, and grayscale conversion are common pre-processing techniques used to standardize input data and make it more suitable for later processing steps.
v) Feature extraction: Feature extraction handles a major part in recognition sign language models impacts both the training process and the speed at which they can differentiate between different signs or words. Features are derived from raw data and correspond to crucial places in the hands and faces for sign language communication. Statistical processes are used to calculate features and weigh them depending on their discriminating value. The neural model learns the probability of feature association with specific classes by representing them as vectors in latent space. Various feature engineering approaches are presented, with some using specialized tools for extraction. The quantity and weight distribution of features are optimized to improve model accuracy and scalability.
Feature extraction is the process of extracting relevant information or features from pre-processed input data that can be used to recognize signs. Handcrafted features such as the Local Binary Patterns, Color Histograms, Principal Component analysis, Edge Detection, and Deep Learning-based features generated from convolutional neural networks (CNNs) might be utilized for this purpose.
vi) Training: Deep learning model is trained on labelled data (for example, photos or video frames tagged with associated sign language motions). Sign language recognition models are typically trained using supervised techniques (SVMs), Multi-layer perceptron or deep neural networks. During training or learning, the model learns to map input features to their associated sign language movements by iteratively optimizing model parameters using algorithms such as gradient descent.
vii) Recognition: In the recognition phase, the trained model is deployed to recognize sign language gestures in real-time or on static images or video sequences. The input data undergoes the same pre-processing steps as during training, followed by feature extraction using the trained model. The extracted features are then classified or matched to predefined sign language classes using techniques such as nearest neighbor classification, softmax classification, or sequence labeling for continuous sign language recognition.

3.1 Classifier (Naïve Bayes, Decision Tree, Random Forest)

Here Bayes' theorem is analyzed for the classifiers, assuming independence between features. It is simple, fast, and takes only a modest amount of training data. Naïve Bayes classifiers excel in vast feature spaces with conditional independence based on class labels. Naïve Bayes classifiers are effective for multi-class classification tasks in sign language recognition, with each class representing a distinct gesture.
Decision Trees are non-parametric supervised learning techniques used for classification and regression. They divide the feature space into regions depending on feature values and make decisions at each node to optimize class integrity. In sign language recognition, Decision Trees can be used to create models that directly map input data to sign language movements, providing insight into the decision-making process.
Random forests are ensemble learning approaches that use several decision trees to increase classification accuracy and robustness. They provide unpredictability to training by bootstrapping samples and randomly picking subsets of characteristics for each tree.
When compared to individual decision trees, Random Forests reduce overfitting while improving generalization performance.

Random Forests can be used in sign language identification to deal with complex and high-dimensional feature spaces, resulting in robust and accurate classification of movements.

Experimental Data

4.1 Dataset description

4.1.1 ChaLearn Looking at People (LAP) Dataset – The LAP Dataset provides extensive annotations on human gestures, covering hand movements, facial expressions, and body poses. These annotations offer detailed insights into the spatial and temporal aspects of gestures, supporting the development and assessment of gesture or body posture recognition algorithms.

4.1.2 NYU Depth V2 Dataset- This dataset offers a rich collection of RGB-D data captured with a Microsoft Kinect sensor in closed environment. It hold in excess of 200,000 frames showcasing diverse indoor scenes with detailed annotations, including object labels, semantic segmentation, and scene types. This dataset supports research in various fields like object recognition, scene understanding, and depth estimation.

4.1.3 MNIST Dataset- It is a dataset which consists of around seventy thousand grayscale images collection of hand-scripted numbers or manuscript from 0 to 9.Here, images consist of atmost 28x28 pixels, resulting in a total of approx 800 features. MNIST is widely utilized for training and testing machine learning models, especially in the area of image classification.

4.2 Integrating Deep Learning with Traditional Methods

Here, we have analyzed and discussed about the approaches of SLR with the different modalities like bidimensional or tridimensional or the combination of the both whether using specified sensors such as depth and also highlighted the various methodologies.

YOLO, or You Only Look Once, is a neural network technique designed to efficiently detect objects in images by examining the entire image in a single pass. Instead of scanning the image multiple times, YOLO divides the input picture into predetermined grid sizes and predicts the presence of recognized items within each grid. This approach essentially treats object detection as a regression problem, where the network makes predictions for each class and object in the image simultaneously in one iteration. (Redmon et al., 2015). Various methods can enhance the effectiveness of YOLO, including the implementation of batch normalization. This technique normalizes the output, minimizing fluctuations in the distribution of internal neurons. There are various implementations of YOLO approach from its intitial version YOLO then, YOLOv2, YOLOv3, YOLOv4

Table 1 shows the details of these models with the description of used methodology, analyzed dataset and its result

Approach	Year	Ref.	Purpose schema	Dataset	Results
RGB,Depth	2021	Tasnim Ferdous Dima[22]	YOLO v5	(MU HandImages ASL)	95
RGB	2019	Yanqiu Liao [9]	BLSTM-3D	DEVISIGN_D	89.8
RGB	2020	Md. Zabirul Islam [24]	LSTM	X-ray image dataset COVID	99.4
RGB	2022	Pu Tu et al 2022[25]	VGG-16 network	-	99.7
RGB,Depth	2022	Sheng Xu [17]	YOLO v5	AIZOO dataset	95.2
RGB	2019	Kopukluetal.,[20]	CNN	NVIDIA benchmarks	83.83
RGB,Depth	2020	Elboushakietal[21]	CNN	isoGD	72.53
RGB,Depth	2020	Elboushakietal[21]	CNN	SBU	97.51

RGB,Depth	2021	Yuan et al[27]	CNN+LSTM	ASL/CSL	99.93 (ASL)/96.1 for(CSL)

Table1. Hybrid sign language Recognition models

Limitations and Future directions

In this extensive review, we investigated recent progress in optical-sensing models for SLR models, focusing on DL methodologies over the past few years. The most prominent output of this survey is to provide a perspective of the advancements in optical-sensing SLR models. It offers insights into their accomplishments, benefits, challenges, and potential future directions. We also acknowledge the consideration of sensor-based modalities and other data collection devices for potential usage in SLR models. Additionally, we discussed the application domain of SLR models and emphasized the need for further exploration to address real-world requirements, particularly for the deaf and speech-impaired communities. While most models showcased isolated SLR, there's a growing interest in transitioning towards continuous SLR, which presents challenges such as dataset availability, tokenization techniques, and multi-modal data modeling. This transition is anticipated to benefit from the integration of visual and linguistic models, leading to enhanced understanding of continuous sign language expressions.

References

[1] R. P. Sharma, G. K. Verma, “Human computer interaction using hand gesture”. Procedia Computer Science, vol. 54, pp.721-727, 2015.
[2] D. Phutela, “The importance of non-verbal communication”. IUP Journal of Soft Skills, vol. 9, no. 4, pp. 43.
[3] A. A. Q. Mohammed, J. Jiancheng and M. S. Islam, “A deep learning based end-to-end composite system for hand detection and gesture recognition”. Sensors, vol. 19, no. 23, pp. 5282, 2019.
[4] P. Nakjai, T. Katanyukul, “Hand Sign Recognition for Thai Finger Spelling: An Application of Convolution Neural Network”. Journal of Signal Processing Systems, vol 91, pp. 131–146, 2019.
[5] M. M. Alam, M. T. Islam, and S. M. Rahman, “A unified learning approach for hand gesture recognition and fingertip detection”. UMBC Student Collection, 2021.
[6]MariuszOszusta,∗ , Jakub Krupski,”Isolated Sign Language Recognition with Depth Cameras“, Procedia Computer Science Elsevier 2021.
[7]RaduMirsu *, Georgiana Simion, Catalin Daniel Caleanu and Ioana Monica Pop-Calimanu,”A PointNet-Based Solution for 3D Hand Gesture Recognition”, MDPI, 2020.
[8]Jyotishman Bora, SaineDehingia, Abhijit Boruah* AnuraagAnujChetia, DikhitGogoi,”Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning”,Procedia Computer Science Elsevier,2023.
[9]Yanqiuliao, pengwenxiong ,weidong min , (member, ieee), weiqiong min , and jiahaolu,”Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D ResidualNetworks, IEEE Access (2019).
[10]YuliusObia , Kent Samuel Claudioa , Vetri Marvel Budimana , Said Achmada,∗, Aditya Kurniawan,”Sign language recognition system for communicating to people with disabilities”,Procedia Computer Science Elsevier (2022).
[11]RaziehRastgoo ,KouroshKiani , Sergio Escalera,”Sign Language Recognition: A Deep Survey “,Elsevier (2021).
[12]Sumaya Siddique, Shafinul Islam, EmonEmtiyaz Neon, TajnoorSabbir, Intisar TahmidNaheen, Riasat Khan,”Deep Learning-based Bangla Sign Language Detection with an Edge Device”,Elsevier (2023).

[13]Qazi Mohammad Areeb , Maryam , Mohammad Nadeem , RoobaeaAlroobaea , And Faisal Anwer,"Helping Hearing-Impaired in EmergencySituations: A Deep Learning-Based Approach,IEEE (2022).
[14]Muneer Al-Hammadi ,Mohamed A. Bencherif , Mansour Alsulaiman , Ghulam Muhammad, Mohamed Amine Mekhtiche , Wadood Abdul , Yousef A. Alohali , Tareq S. Alrayes , Hassan Mathkour , Mohammed Faisal , Mohammed Algabri , HamdiAltaheri , TahaAlfakih and Hamid Ghaleb ,"Spatial Attention-Based 3D Graph Convolutional Neural Network for Sign Language Recognition",MDPI.
[15]Bayan IbrahimmAlabdullah , Hira Ansar , Naif Al Mudawi , AbdulwahabAlazeb , Abdullah Alshahrani , Saud S. Alotaibi, and Ahmad Jalal," Smart Home Automation-Based Hand Gesture Recognition Using Feature Fusion and Recurrent Neural Network",MDPI.
[16]Guoyu Zhou, · Zhenchao ,Cui · Jing Qi(2023),"Blsnet: a tri-branch lightweight network for gesture segmentation against cluttered backgrounds",Elsevier.
[17] Sheng Xu1, Zhanyu Guo2, Yuchi Liu3, Jingwei Fan2, Xuxu Liu , An Improved Lightweight YOLOv5 Model Based on Attention Mechanism for Face Mask Detection, arXiv:2203.16506v3 [cs.CV] 11 Sep 2022 .
[18]YanqiuLiao ,PengwenXiong , Weidong Min , (Member, Ieee), Weiqiong Min , And Jiahao Lu,"Dynamic Sign Language Recognition Based on Video Sequence with BLSTM-3D Residual Networks",IEEE ,2019.
[19] Tuan LinhDang ,SyDat Tran a, Thuy Hang Nguyen , Suntae Kim , Nicolas Monet,"An improved hand gesture recognition system using keypoints and hand bounding boxes,Elsevier, 2022.
[20]Kopuklu, O., Gunduz, A., Kose, N., &Rigoll, G. (2019). Real-time hand gesture detection and classification using convolutional neural networks. arXiv:1901.10323.
[21]Abdessamad Elboushaki, Rachida Hannane, Karim Afdel, Lahcen Koutti,MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences, Expert System with Application, Vol 139, 2020.
[22] Tasnim Ferous Dima ,Md. Eleas Ahmed Using YOLOv5 Algorithm to Detect and Recognize American Sign Language Conference: 2021 International Conference on Information Technology (ICIT).
[23]Molchanov, P.; Yang, X.; Gupta, S.; Kim, K.; Tyree, S.; Kautz, J. Online Detection and Classification of DynamicHand Gestures with Recurrent 3D Convolutional Neural Networks; IEEE CVPR: Las Vegas, NV, USA, 2016; pp. 4207–4215.
[24] Md. Zabirul Islam, Md. Milon Islam * , Amanullah Asraf A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, Elsevier, 2020, Informatics in Medicine Unlocked 20 (2020) 10041.
[25]Zhang,L.; Zhu, G.M.; Mei, L. Attention in convolutional LSTM for gesture recognition. In Proceedings of theNeural Information Processing Systems, Montreal, QC, Canada, 4 December 2018.
[26]Kingkan, C.; Owoyemi, J.; Hashimoto, K. Point Attention Network for Gesture Recognition Using Point Cloud Data; BMVC: NewCastle, UK, 2018; p. 118.
[27]G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, "Hand gesture recognition using deep feature fusion network based on wearable sensors," IEEE Sensors Journal, vol. 21, no. 1, pp. 539–547, 2020.
[28] Shubha Chaturvedi, Dr. Manoj Rawat,"Hand Gesture Recognition: An Approach of Multiple Object Tracking ", International Conference on Recent Trends in Machine Learning and Image Processing (MLIP), November 2023.