Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 12
December 2023

Review on Real Time Conversion of Sign Language to Text in Context of Vision Based and Deep Learning

Prof. Chirag R. Prajapati [1], Prof.(Dr.) Atul M Gonsai [2]
[1]Research Scholar,Saurashtra University, Rajkot, Gujarat.
[2]Computer Science Dept, Saurashtra University, Rajkot, Gujarat.

Abstract. In their daily contacts, those people have a hard time hearing or speaking extensively relies on sign language. To communicate the opinions of this community to people who may not be sign language experts, sign language alphabets have been devised. Nevertheless, it can be difficult for many individuals without disabilities to comprehend the motions and signals used by the deaf and mute [37]. Techniques have been developed to convert the sign language made By those with disabilities into a format that is understandable to others without disabilities in order to close this communication gap. This study is based on a number of procedures, For instance image Obtaining, initial processing, and dividing of hand gestures, Extracting characteristics and categorization techniques.
In an effort to develop systems that can recognize sign language more accurately, researchers are investigating improved techniques for doing so.

Keywords: Recognition of sign language, SLR, Computer vision, Machine Learning, Deep learning, CNN

1 Introduction

The use of sign language as a powerful and incredibly successful form of communication by both hearing-impaired people and the broader public stands out. The most efficient approach for enhancing social interaction and closing the communication divide between deaf-mute individuals and those with regular speech and hearing capabilities is embodied in this approach. Sign language interpreters are instrumental in bridging this gap by translating sign language into spoken language and vice versa, allowing meaningful interactions with the hearing-impaired. It's evident from a review of several research articles that the predominant focus of current research endeavors lies in recognizing static sign language signs from images or video sequences captured in controlled settings. This method frequently entails the use of glove sensors or gloves with unique designs worn by signers. To standardize the task, these gloves are used during the segmentation process. The fact that participants must use the communication device while wearing sensor components and gloves is a significant disadvantage of this methodology.

Vision-based approaches in computer interfaces take advantage of how people perceive and engage with their environment to provide non-intrusive solutions. Although building a vision-based user interface for regulated situations can be difficult, doing so is nonetheless doable [2]. Gesture recognition primarily hinges on the selection of features due to the unique aspects of hand movements, including shape variations, materials, and motion. Separating factors like finger orientations, individual finger positions, color of skin, and hand shapes can simplify the identification of hand postures in static hand gestures. [36]. However, changes in lighting and image backdrops can impair the availability and dependability of these functionalities. Additionally, non-geometric components including color, texture, and silhouette do not always convey enough details for recognition. The objective of this work is to critically review and assess the research procedures used in earlier studies in this field. Additionally, it tries to suggest the best strategy for upcoming studies in this area..

2 Analysis of the SLR system

Since sign language does not involve hearing or vocalization, it serves as the primary mode of communication for the deaf community. Without the use of spoken words, sign language is a unique means of promoting interaction. Sign

languages use a variety of face emotions, hand and palm movements, torso movements, as well as different hand forms, orientations, and shapes to dynamically represent concepts. People can effectively communicate their thoughts and emotions thanks to this rich and diverse method.

2.1 Methods for Sign Capture

For the sign language recognition system to function, the signs must be recorded. After our examination, we discovered that various devices have been utilized to capture multimodal data, such as Microsoft Kinect sensors [6, 14], the Microsoft Kinect (RGB-D) sensor managed by the Nui Capture Analyze application [7], front and mobile cameras [5, 8, 11], cameras of Sony video [9], and RGB videos from the Canon 600 D camera [10]. These devices are employed to record images of hand gestures, with the Microsoft Kinect camera being particularly effective in capturing single-hand signs, double-hand signs, and finger-spelling [4, 12, 13]. Computer vision applications such as gesture recognition, motion recognition, robotics, and virtual reality are used by Microsoft Kinect sensors.

3 Techniques of SLR

There are two separate approaches to sign language recognition through visual means: appearance-based and deep learning-based methods. Several 2D intensity images are used to model appearance-based strategies. A series of views are then used to model gestures.

The palm's position and the joint angles are inferred using appearance-based techniques[18,19]. Appearance-based systems utilize videos or photographs as their data sources. They interpret these videos and photos straightforwardly, without employing a spatial model of the body.

In Traditional Machine Learning-based Approaches, CNN has garnered the most attention among these methods for categorizing gestures.

4 System for SLR Based on Visual Appearance

Fig. 1 depicts a straightforward block schematic of the System for Sign Language Recognition Based on Visual Appearance.

4.1 Image capture

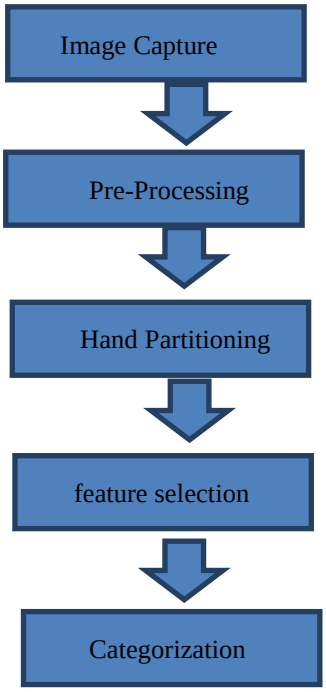

Fig. 1. System for SLR Based on Visual Appearance

The camera is a crucial component of the SLR method used as method of input. The input data for the Sign Language Recognition (SLR) system consists of moving images that can be readily captured by a camera. However, some researchers [1,2,4,7,11] take photographs using standard cameras. To lessen the difficulty of employing sensor-based gloves, few of Researchers claim that they are opting for cameras in lieu of gloves. Since cameras frequently offer a variety of video formats, we must specify the preferred format as well as the default format by utilising a Digitizer Configuration Format (DCF) file. Because the web camera's image is hazy, several researchers have employed better cameras.

Another camera used to take pictures is called a Microsoft Kinect [21]. Nowadays, researchers frequently employ Kinect because to its functionality. Both colour and depth video streams can be present at once on the Kinect. With depth data, background segmentation is simple to perform, and signal language recognition on Kinect can be used to accomplish this.

Several research studies have utilized established datasets, including ASL Gesture Dataset 2012 [23], the American Sign Language Image Dataset (ASLID) [29], ChaLearn Looking at People 2014,ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) – 2010, RWTH-Phoenix-Weather 12 [14], and RWTH-Phoenix-Weather Multisigner 2014. Few scientists produce their own data for data training due to the dearth of datasets for sign language in specific regional languages. American Sign Language represents the characters of the English alphabets signs in Fig. 2, while Fig. 3 displays Indian Sign Language two-hand signs.

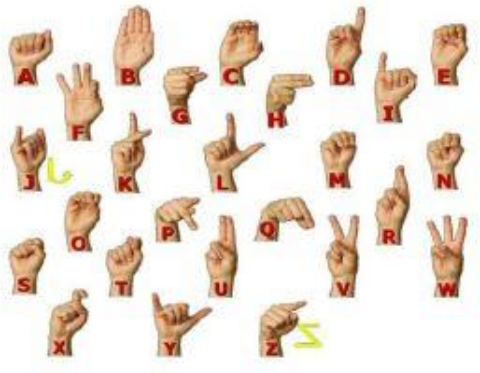

Fig. 2.Sign of ASL [39]

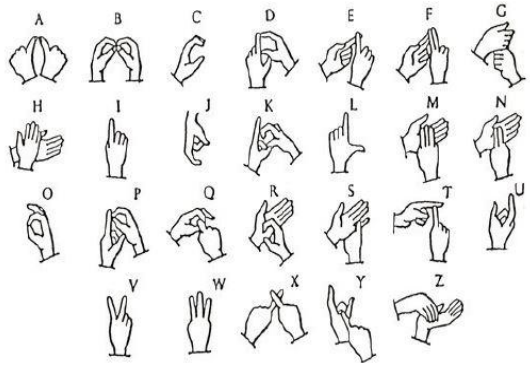

Fig. 3.Sign of ISL[40]

4.2 Pre-processing and segmentation

The pre-processing stage for images enhances the system input for editing pictures and films. Some of the most popular techniques for removing noise from raw photos

or videos are median and Gaussian filters. In research, median filtering is exclusively applied for the initial image preprocessing stage [7, 9], and morphological techniques [20] are commonly employed to eliminate superfluous data from the input.

The segmentation method can either be contextual or non-contextual. In a non-contextual segmentation, spatial relationships are not considered, and pixel grouping relies on global characteristics. On the other hand, contextual segmentation takes into consideration the spatial connections between features, such as the use of edge detection algorithms. To produce more precise outcomes, the integration of hand movement tracking with skin detection is employed in skin detection. Much like skin detection, hand segmentation benefits from the use of colored gloves to distinguish the hands by their unique features.

There are different color models available which are used to process skin colour segmentation such as RGB Model, HSV Model, although colour segmentation is more challenging due to potential sensitivity issues with lighting, cameras, and skin tone.In research [9]HSV and YCbCrcolour models were used to segment people's palms and faces. Ahmed et al. [28] segmentation of hand skin colour was done using the RGB colour model. Studies have shown that the YCbCr color model proves effective for color segmentation across various lighting conditions. [22].

According to study [31], K-means clustering in the YCbCr color space can differentiate the foreground area of an image from the background. For tracking hands in the video, Badhe et al. gives a tracking hand movements in Ref. [1].Hand segmentation is a method used in vision-based systems to separate hands and other elements from the background of the image.

4.3 Feature Selection

The process of deriving various characteristics from an image is referred to as feature selection. Image background, image translation, scaling, shape, rotation, angle, and coordinates are among the features. Fourier descriptors [1, 7, 9] are utilized to capture the outer boundary of objects in images. To identify items in an image, boundaries are created by the coordinate sequence. Each frame's tracking points for both arms are extracted by the HS Optical Flow method [5].

In Almeida et al.'s 2014 study, they used the Speeded Up Robust Features (SURF) algorithm as a feature selection technique. [38]. In the Hough transform, when identifying lines using polar directions, elements are paired as (q, h). [38]. For the purpose of SL recognition, it is utilised to identify two-hand communication aspects.

4.4Classification

The last step, classification, is crucial to the popularity of gestures. When the gadget deals with loud information and an uncontrolled environment, some issues arise. The way to gauge popularity is to choose the outfit from a group of options that best suits the phrase sequence. Two different types of gesture popularity procedures exist.

Certain studies have applied the derived features for recognizing gestures, employing techniques such as template matching, while others have employed machine learning classifiers based on Hidden Markov models. (HMM).

4.4.1 Machine learning classifiers We opted for the Hidden Markov model (HMM) [28], which offers the highest probability of generating the observed dataset and the highest likelihood of producing the sign.. The way to gauge popularity is to choose the outfit from a group of options that best suits the phrase sequence. These characteristics and linguistic components are used to classify signs using support vector machines (SVMs) [38,17]. A multi-class classifier, the SVM classifier, aims to find an optimal hyperplane to serve as the decision function. The SVM classifier can decide on the sign once it has been trained on photos with specific gestures. Regression problems can be categorised using a machine learning technique called random forest (RF). At the commencement of a new classification process, a set of characteristics related to the object is selected as the reference. BTA [5], Sugeno-type fuzzy inference systems [7], AdaBoost multiclass [6], ANNs [20], and also MPCNNs [22] are additional machine learning-based classifiers.

4.4.2 Pattern Matching In order to construct matching between test and reference signs, a suitable symbolic similarity measure is investigated. A simple nearest neighbor classifier is subsequently employed to recognize an unknown sign by considering it as one of the recognized signs, based on a chosen threshold level. Euclidean distance [1, 4] is used to calculate the distance, comparing each gesture image in the training dataset with each gesture in the testing dataset.

5. Conventional machine learning methods

5.1 Image capture

The camera plays a vital role as a component of the sign language recognition (SLR) input system. A camera can effortlessly capture a dynamic image to serve as the input data for the Sign Language Recognition (SLR) system. Nevertheless, some researchers take pictures with straightforward cameras. Simple cameras are still used by certain researchers to take pictures [8,10,12]. Microsoft Kinect is yet another gadget that can be used to take pictures. Due of its characteristics, Kinect is now frequently employed by researchers. The Kinect has the capability to simultaneously deliver depth and color video streams.

5.2Datasets

Many researchers generate their own datasets for training their data. Researchers capture data from the signer to create a dataset because sign language datasets aren't readily available in some places. Because there are typically not enough datasets available for research, researchers create their own datasets in sign language. Many research studies have utilized predefined datasets, such as those from the American Sign Language and ILSVR) - 2010 [33], Image Dataset (ASLID) [27], SIGNUM and The ArSL database.

Ref. [3] contains the most recent information about some pre-processing technique and studies with active sensors. Utilizing the information gathered with the use of a Leap Motion Controller (LMC), they recommended a feature extraction technique [24]. A gadget called an LMC can recognize hand gestures at 200 frames per second and provides identification each time it does. The specific LMC API may directly associate hand movements and fingertip detection with each other. LMC is continually evolving and is not without flaws. When hands are flipped over, there are several issues with how the API is implemented. The Leap Motion controller is a

compact and commercially available sensor designed for capturing hand and finger movements in a 3D environment, positioned about eight inches above the device [4].

5.2.3 Pre-trained model

A model that has already been trained on a sizable benchmark dataset to address a problem resembling the one we're trying to solve is known as a pre-trained model. AlexNet [8,16,19,27] Additionally, researchers are concentrating on developed networks that may be transferred via transfer learning. As a result of LeNet's development of AlexNet, it is one of the lights that lit up the deep learning explosion [87]. A neural network with depth in both the vertical and horizontal directions is GoogleNet [10]. In neural networks, a horizontal direction with width is also known as a "inception structure," which employs multiple filters, each of a different size, and ultimately merges their outputs. A main neuron structure could be pieced together using the "inception structure" to create a sparse, high-performance neural network structure.

VGG16 [33] Simonyan and Zisserman described the VGG network design in their 2014 paper. The only 3 X 3 convolutional layers placed on top of one another in increasing depth are used to characterise the VGG family of networks. Max pooling handles decreasing the volume size. Although ResNet model structure may be successfully trained at depths of 50–200 for ImageNet and over 1000 for CIFAR-10, VGG-16 networks were once thought to be exceedingly deep. VGG has two significant drawbacks i.e. training progress is really slow and there are a lot of weights in the network.

An important contribution to the literature on deep learning is the ResNet50 [32, 35] design, which shows how incredibly deep networks may be trained using conventional SGD (and a decent initialization function) by employing residual modules. Due to the use of global average pooling rather than fully linked layers, which reduces the size of the model down to 102 MB for ResNet50, ResNet is significantly deeper than both VGG19 and VGG16.

5.2.4 Preprocessing

The purpose of the image pre-processing stage is to alter the video or photo inputs in order to enhance the system's overall performance as a whole. Some of the most often used methods to minimise noise in captured photos or videos include median and Gaussian filters. In order to classify images in testing images, neural networks use the pre-processing method to extract features from the training image and store these characteristics.

Some of Pre-processing methods used in the reviewed paper are: Nearest neighbor interpolation [21], Haar feature classifier, Median filtering ,Bandpass filter [30], Gabor filter [34], Savitzky-Golay filter [24], HOG [29], Image background subtraction , RGB to HSV colour space [12].

5.4 Neural network model

5.4.1 CNN Convolutional neural networks (CNNs) are seen to be the most crucial deep learning neural network model to use when identifying and categorising images. It creates a hierarchical structure resembling models of human brain activity by extracting low-level features into pertinent features via multilayer superposition [3, 15]. Because the most recent features are passed from the previous layer, the laborious manual feature extraction process can be avoided. CNN combines classification and feature learning. In general, a convolutional neural network has multiple layers. In general, a convolutional neural network has multiple layers. A convolution process is performed in the convolution layer to extract information from an input layer or a preceding layer. The amount of characteristics and calculations can be continuously decreased thanks to the pooling layer. The completely connected layer in CNN acts as a "classifier."

These layers are used by CNN to automatically learn the values. Applying convolution filters, nonlinear activation functions, pooling, and back propagation, our CNN may learn to recognise edges, recognise shapes, and assist boundaries in recognising higher-level features.

5.4.2 CNN-RNN A CNN architecture can be created using an LSTM (long-term memory) and

an RNN (recurrent neural network) model to collect spatial characteristics from the video for SLR and then extract temporal features from the video. LSTM recognises gesture classes using the SLR video's sequence data.

5.4.3 3D CNN Multiple layers of pooling and convolution can be combined to create a CNN architecture. Image datasets are classified and their spatial properties are extracted using 2D CNNs. In any case, both spatial and temporal data must be recorded for SLR in videos. Convolution is used by 3D CNN to extract both spatial and temporal data from films.

5.4.4 Deep-CNN To learn conditional probabilities for the existence of components and their spatial relationships inside image patches, deep convolutional neural networks (DCNNs) are used. video sequences where the temporal structure fills in the gaps or in ways that are less visible in static visuals. CNN-dynamic Bayesian network (DBN) [13], and attention-based RNN [30], Faster R-CNN [8], and stream CNN [26] are other neural network models that perform similarly to CNN and are used to extract spatial data from movies and lengthy sequences of the stance.

5.5 Loss function

How closely our predicted class labels agree with our ground truth labels is measured by a loss function. Our loss is lowered by the level of agreement between those sets of labels. The activation function employed in your neural network's output layer has a direct impact on your preference for loss functions.

5.6 Optimizer

By changing parameter values in the neural network model, an optimizer can lower the loss function. The optimizer uses the loss function as a guide to determine if it is going in the right or wrong path. An iterative technique known as gradient descent starts at a random location and proceeds down the slope of the goal to the function's lowest point.

A straightforward modification to the programming of the classic gradient descent method is stochastic gradient descent (SGD), which calculates the gradient and modifies the weight matrix value. Instead of using the entire training set, W on small batches of training data. Possibly the most important algorithm for deep neural network training is SGD. That is a modified form of the GD approach where each iteration updates the model parameters. Another effective method for calculating adaptive learning rates for each parameter is adaptive moment estimate (Adam) [14]. In addition, Adam keeps a momentum-like exponential decaying average of previous gradients. Adam operates like a heavy ball with friction, as opposed to momentum, which can be visualised as a ball strolling down a hill. As a result, Adam prefers flat minima on the error surface.

5.7 Classification

The process of classification locates a function to establish which category the incoming data falls within. It may be divided into two categories or more. The features of the data to be classified, the quantity of training samples, and the classification construction method are some of the factors that affect classification accuracy.

The multi-classification networks are implied by the Softmax classifier, which efficiently determines classes based on final output probabilities. In Softmax, the multiclass problem is given decimal probability for each class, and the probabilities add up to 1.0. The softmax activation function is typically used in conjunction with the categorical cross-entropy loss function. The logarithmic output value of the model makes its output the most useful. The model output is rescaled using the softmax activation to give it the desired attributes. Due to the way the ReLU [23, 25] appear when plotted, they are also referred to as "Ramp functions". It is important to note that the function is 0 for negative inputs and climbs linearly for positive values. The ReLU is extremely computationally efficient even though it isn't always saturable. When compared to nonlinear functions, ReLU has no vanishing gradient issue because the gradient is constant in the nonnegative area.

5 Outline of Previous work

This section provides a summary of earlier works on the methods used in various research as well as the results of work on gesture and sign language recognition. This section presents and tabulates information, including performance and methodologies used. Here the table is created which lists the techniques used and is categorised

by the image capturing, pre-processing and pre-trained model, neural network model, loss function, optimizer, classification, and accuracy. The accuracy column displays the best accuracy that the suggested approach produced.

The most common types of cameras used for data collection are Kinect cameras and regular cameras. The process of gathering data and sending it to the SLR system is crucial. Pre-processing is done after data collection and is necessary to increase accuracy and gather more data from the initial set of data. Gaussian and median filters are used in pre-processing to take the noise out of the input data. Images are compressed before segmentation in order to speed up calculation. The segmentation process uses skin colour segmentation the most. The backdrop colour and skin tone may be effectively distinguished from the HSV, YCbCr, and RGB colour spaces. The study shown that segmenting skin colour with additional parameters like threshold and edge identification enhances the segmentation outcome.

The appearance-based strategy's classification of motions is the last phase, and it takes features out of photos for quick identification.The most popular classification algorithms are ANN and SVM. SVM offers excellent performance when evaluated by researchers. The hidden Markov model (HMM), which is used in statistical techniques to obtain spatiotemporal information, is a general way to recognise sign language. Reviewing the available data, the majority of models rely on sensors to get information about the surroundings. While HMMs and SVMs are utilised for classification, neural networks are widely used in vision-based approaches to images and videos due to the increase in the availability of data sources.

Another crucial processing technique for the identification of sign language is the neural network model. To understand sign language, the CNN processing image convolves through convolution layers, pooling, activation functions, and fully linked layers. Utilizing applied 3D-CNNs, motion information is obtained from depth variation in frames and features of temporal and spatial correlation are acquired. Videos of sign language can provide simulation temporal sequence data using LSTM-based techniques. Pre-trained models resemble magic, and we may use them right away without providing any data or undergoing any training. Researchers have given pre-trained models a lot of thought in recent years due to their great potential. Additionally, it lowers the cost of training and datasets. To lower the cost of training, many researchers have employed pre-trained models like gesture-based VGG16, Google Net, and AlexNet.

Because it can self-learn and self-associate, the deep neural network (DNN) produces higher outcomes but calls for the greatest training dataset. It now has the computing power to run applications on massive datasets thanks to recent technological advancements in GPUs. The implementation of a new algorithm and enhancements to current ones allow for better programme execution. Increasing processing speed enables big data and cloud-based apps to run applications more quickly.

As per my research on deep learning base approach I have found some best results in which SsShivashankara and et all (2018) [10] uses Inception model as a pre-trained model using CNN-RNN, Categorical cross entropy as a Loss function and softmax function as a classifier and get 90% accuracy. While Wadhawan and et al. [15] in their research shows they have used AlexNet as pre-trained model using CNN and get 99.72% accuracy. Yongsen and et all (2020) [25] also used the same technique of CNN with same softmax as a classifier and get 98.01% accuracy. But B. Shi et al.[10] in his research uses Alexnet as a pre-trained model with Faster CNN and as a optimizer uses Stochastic Gradient Descent (SGD) where they get accuracy of 42%. So we can do more research in this area to make the result more accurate.

6 Conclusion

We presented a quantitative analysis of various sign language recognition techniques in this research. An examination using SLRs based on look and vision (deep learning) following conclusions were drawn from a review of papers:

A system for identifying solely static signs and alphabets in sign language has evolved into one that can accurately recognise dynamic actions that appear in continuous sequences of images.

Results from vision-based approaches are often superior to those from appearance-based approaches in published research. Making a broad vocabulary for sign language recognition systems is currently receiving greater attention from researchers. Access to more training for specific samples is made possible by the availability of datasets and advancements in computing speed. Many academics are creating their own tiny datasets to use in the development of their SLR. There are still certain nations and languages for which large databases are unavailable. Most nations' versions of sign language are entirely focused on their grammar and how each phrase is presented, such as by using words or complete sentences. In the sequence of pictures and video streams, deep learning-based techniques like CNN, RNN, LSTM, and Bi-Directional LSTM Models offer good recognition accuracy.

References

[1] P.C. Badhe, V. Kulkarni, Indian sign language translator using gesture recognition algorithm, in: 2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), Bhubaneswar, 2015, pp. 195–200.

[2] J. J. Raval and R. Gajjar, “Real-time Sign Language Recognition using Computer Vision,” in 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, May 2021, pp. 542–546. doi: 10.1109/ICSPC51351.2021.9451709.

[3] Nikhil Kasukurthi, BrijRokad, Shiv Bidani, AjuDennisan American Sign Language Alphabet Recognition Using Deep Learning, 2014.

[4] A. Nandy, J.S. Prasad, S. Mondal, P. Chakraborty, G.C. Nandi, Recognition of isolated indian sign language gesture in real time, Inf. Process. Manag. (2010) 102–107.

[5] P.V.V. Kishore, M.V.D. Prasad, D.A. Kumar, A.S.C.S. Sastry, Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks, in: 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, 2016, pp. 346–351.

[6] G.A. Rao, P.V.V. Kishore, Selfie sign language recognition with multiple features on adaboostmultilabel multiclass classifier, J. Eng. Sci. Technol. 13 (8) (2018) 2352–2368.

[7] P.V.V. Kishore, P.R. Kumar, A video based Indian Sign Language Recognition System (INSLR) using wavelet transform and fuzzy logic, Int. J. Eng. Technol. 4 (5) (2012) 537.

[8] B. Shi, et al., American sign language fingerspelling recognition in the wild, in: 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 145–152.

[9] S.S. Shivashankara, S. Srinath, American sign language recognition system: an optimal approach, Int. J. Image Graph. Signal Process. (2018).

[10] KshitijBantupalli, Ying Xie, American sign language recognition using machine learning and computer vision, Master of Science in Computer Science Theses 21 (2019).

[11] M. Krishnaveni, V. Radha, Classifier fusion based on Bayes aggregation method for Indian sign language datasets, Procedia Eng. 30 (2012) 1110–1118.

[12] Yang Su, Qing Zhu, Continuous Chinese sign language recognition with CNN-LSTM, in: Proc. SPIE 10420, Ninth International Conference on Digital Image Processing (ICDIP 2017), 21 July 2017, p. 104200F, https://doi.org/10.1117/ 12.2281671.

[13] Q. Xiao, Y. Zhao, W. Huan, Multi-sensor data fusion for sign language recognition based on dynamic Bayesian network and convolutional neural network, Multimed. Tool. Appl. 78 (2019) 15335–15352, https://doi.org/10.1007/s11042-018-6939- 8

[14] Sylvie C.W. Ong, SurendraRanganath, Automatic sign language analysis: a survey and the future beyond lexical meaning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 6, https://doi.org/10.1109/TPAMI.2005.112 (June 2005), 873–891.

[15] A. Wadhawan, P. Kumar, Deep Learning-Based Sign Language Recognition System for Static Signs, Neural Comput&Applic, 2020,

https://doi.org/10.1007/s00521- 019-04691-y.

[16] N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden, Neural sign language translation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7784–7793.

[17] H. Lilha, D. Shivmurthy, Analysis of pixel level features in recognition of real life dual-handed sign language data set, in: Recent Trends in Information Systems (ReTIS), 2011 International Conference on, IEEE, 2011, December, pp. 246–251.

[18] N.C. Camgoz, S. Hadfield, O. Koller, R. Bowden, SubUNets: end-to-end hand shape and continuous sign language recognition, in: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3075–3084.

[19] A. Kika, A. Koni, Hand gesture recognition using convolutional neural network and histogram of oriented gradients features, in: CEUR Workshop Proceedings, vol. 2280, CEUR-WS, 2018, pp. 75–79.

[20] R. Akmeliawati, M.P. Ooi, Y.C. Kuang, Real-time Malaysian sign language translation using colour segmentation and neural network, in: 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007, Warsaw, 2007, pp. 1–6.

[21] Becky Sue Parton, sign language recognition and translation: a multidisciplined approach from the field of artificial intelligence, J. Deaf Stud. Deaf Educ. 11 (1) (2006) 94–101, https://doi.org/10.1093/deafed/enj003. Winter.

[22] J. Nagi, et al., Max-pooling convolutional neural networks for vision-based hand gesture recognition, in: 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, 2011, pp. 342–347, https://doi. org/10.1109/ICSIPA.2011.6144164.

[23] http://www.massey.ac.nz/~albarcza/gesture_dataset2012.html.

[24] Biyi Fang, Jillian Co, Mi Zhang, DeepASL: enabling ubiquitous and non-intrusive word and sentence-level sign language translation, in: Proceedings of the 15th ACM Conference on Embedded Networked Sensor Systems (SenSys '17), 2017.

[25] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, Woosub Jung, SignFi: sign language recognition using WiFi, Proc. ACM Interact. Mob. Wear. Ubiq. Technol. 2 (1) (2018) 21. Article 23 (Mar. 2018).

[26] P.V.V. Kishore, K.B.N.S.K. Chaitanya, G.S.S. Shravani, TejaMaddala, KiranEepuri, D. Anil Kumar, DSLR-net a Depth Based Sign Language Recognition Using Two Stream Convents, vol. 8, 2019, pp. 765–773.

[27] SrujanaGattupalli, Amir Ghaderi, VassilisAthitsos, Evaluation of deep learning based pose estimation for sign language recognition, in: Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '16), Association for Computing Machinery, New York, NY, USA, 2016, https://doi.org/10.1145/2910674.2910716. Article 12, 1–7.

[28] A.A. Ahmed, S. Aly, Appearance-based Arabic sign language recognition using hidden Markov models, in: 2014 International Conference on Engineering and Technology, ICET, Cairo, 2014, pp. 1–6.

[29] http://vlm1.uta.edu/~srujana/ASLID/ASL_Image_Dataset.html.

[30] W. Tao, M.C. Leu, Z. Yin, American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion, Eng. Appl. Artif. Intell. 76 (2018) 202–213.

[31] EriglenGani, AldaKika, Albanian sign language (AlbSL) number recognition from both hand's gestures acquired by Kinect sensors, Int. J. Adv. Comput. Sci. Appl. 7 (2016) 7, 2016.

[32] JunfuPu, Wengang Zhou, Houqiang Li, Dilated convolutional network with iterative optimization for coutinuous sign language recognition, in: International Joint Conference on Artificial Intelligence, IJCAI, 2018, pp. 885–891.

[33] S. Masood, H.C. Thuwal, A. Srivastava, S. Satapathy, V. Bhateja, S. Das, American sign language character recognition using convolution neural network, in: Smart Computing and Informatics. Smart Innovation Systems and Technologies, vol. 78, Springer, Singapore, 2018.

[34] Arif-Ul-Islam, S. Akhter, Orientation hashcode and articial neural network based combined approach to recognize sign language, in: 2018 21st International Conference of Computer and Information Technology, ICCIT, Dhaka, Bangladesh, 2018, pp. 1–5.

[35] PulkitRathi, Kuwar Gupta, Raj, SoumyaAgarwal, AnupamShukla, sign language recognition using ResNet50 deep neural network architecture. https://doi.org/1 0.2139/ssrn.3545064, February 27, 2020.

[36] I. A. Adeyanju, O. O. Bello, and M. A. Adegboye, “Machine learning methods for sign language recognition: A critical review and analysis,” Intelligent Systems with Applications, vol. 12, p. 200056, Nov. 2021, doi: 10.1016/j.iswa.2021.200056.

[37] S. Subburaj and S. Murugavalli, “Survey on sign language recognition in context of vision-based and deep learning,” Measurement: Sensors, vol. 23, p. 100385, Oct. 2022, doi: 10.1016/j.measen.2022.100385.

[38] S.G.M. Almeida, F.G. Guimar˜aes, J.A. Ramírez, Feature extraction in brazilian sign language recognition based on phonological structure and using RGB-d sensors, Expert Syst. Appl. 41 (16) (2014) 7259–7271, https://doi.org/10.1016/j. eswa.2014.

[39] G. G. S, P. N, A. Yaji, A. M, A. M. Dsilva, and C. S R, “Review on Text and Speech Conversion Techniques based on Hand Gesture,” in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, May 2021, pp. 1682–1689. doi: 10.1109/ICICCS51141.2021.9432277.

[40] V. Adithya, P. R. Vinod, and U. Gopalakrishnan, “Artificial neural network based method for Indian sign language recognition,” in *2013 IEEE CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES*, Thuckalay, Tamil Nadu, India, Apr. 2013, pp. 1080–1085. doi: 10.1109/CICT.2013.6558259.