[1]*Dr. G. B. Sambare

[1]Shailesh B. Galande

[1]Tejas Ippar

[1]Purnesh Joshi

[1]Yash Maddiwar

[1]Abdul Waasi S Mulla

Translating Gestures: Utilizing CNN to Enhance ASL Communication and Understanding

Journal of Electrical Systems

Abstract: - This research introduces a system for translating American Sign Language (ASL) utilising Convolutional Neural Networks. (CNNs). The system under consideration has the capability to identify manual gestures executed by persons communicating in American Sign Language (ASL) and subsequently convert them into written language, thereby facilitating uninterrupted communication between the deaf and hearing populations. The CNN model was validated and trained using a set of data made up of 78,300 pictures of movements in the ASL Alphabet written in American Sign Language (ASL). In order to improve the performance of the model, the pre-processing of the data included a number of stages, such as converting the photos to grayscale, normalising the pixel values, and enhancing performance of the model. In order to ease classification, fully connected layers were added after a succession of pooling and convolutional layers in CNN's design. After 15 epochs of training, the model attained a validation accuracy of 99.85%. The findings of this research demonstrate the viability of employing Convolutional Neural Networks (CNNs) in the creation of precise and effective American Sign Language (ASL) translation systems, which can serve as a means of communication for people with auditory impairments.

Keywords: Convolutional Neural Networks, American Sign Language, Deep Learning, Gesture Recognition.

I. INTRODUCTION

The deaf population in the United States communicates with one another through the use of a visible language known as American Sign Language (ASL). The meaning of what is being communicated in ASL is conveyed through a complicated system of hand movements, facial expressions, and body language. The American Sign Language (ASL) is acknowledged as a language, but it is not a written language. Because of this, it can be challenging for members of the deaf community to communicate with hearing people. Because of this communication impediment, their access to information, educational possibilities, and employment prospects has been restricted.

Machine translation systems and software that recognises speech are just two examples of the many different technologies that have been developed to interpret spoken languages. However, due to the visual character of ASL, its transformation has proven to be a more difficult undertaking than originally anticipated. The traditional approaches to interpreting American Sign Language require the utilisation of physical labour, such as the hiring of sign language interpreters, which can be labour-intensive, as well as expensive. As a consequence of this, there is a requirement for automated translation systems that are able to enhance the precision and effectiveness of ASL translation.

Recent developments in deep learning and computer vision have shown considerable potential for increasing the precision of American Sign Language (ASL) translation. Convolutional Neural Networks (CNNs), in particular, have shown their efficacy in a variety of computer vision tasks, including object identification and recognition, picture categorization, and character recognition, among others. CNNs have also been used to recognise American Sign Language movements, an area in which they perform significantly better than conventional machine learning algorithms.

In this research, we investigate how the application of CNNs can result in a more accurate translation of ASL. Our investigation is centred on the creation of a CNN-based ASL translation system that is capable of recognising ASL

[1] Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India.

*Corresponding author e-mail: santosh.sambare@pccoepune.org

movements and converting them into text. The goal of this research is to answer the following question: "Can a CNN improve the accuracy of ASL translation?" In order to train our translation system, we make use of a CNN model and an ASL dataset that is accessible to the public. We also evaluate the effectiveness of our system using a variety of measures and compare it to the different technologies that are already available for ASL translation.

The primary purpose of this research paper is to make a contribution to the advancement of technology for the transmission of American Sign Language (ASL) and to enhance the accessibility of communication and information for the deaf population. The present article is organised in the following manner: Initially, a literature review was conducted on the topic of American Sign Language (ASL) translation and Convolutional Neural Networks (CNNs), which has already been explored.

After that, we move on to a discussion of the methodology that was applied throughout the course of our research. This includes the dataset, the CNN architecture, and the assessment criteria. We explore the ramifications of our discoveries before presenting the outcomes of our investigations. In the final section of the article, we will briefly summarise our most important contributions and make some suggestions for where future research should go.

II. Literature review

. The research paper referenced as [1] outlines a fresh methodology for ongoing recognition of sign language that takes into account contextual factors and employs a generative adversarial network framework. The second paper presents a methodology based on deep learning to detect hand movements and recognise common sign language gestures. The third paper presents the utilisation of Inflated 3D (I3D) Convolutional Neural Networks as a means of achieving sign language recognition on a large scale, independent of the signer. (SLR). The research paper referenced as [4] introduces a self-attention mechanism in a fully-inception (SAFI) network to enable vision-based end-to-end continuous recognition of sign language. The authors of Paper [5] present a new methodology for the recognition of sign language in a continuous manner. This approach employs a hybrid architecture comprising of CNN & RNN. Meanwhile, Paper [6] presents an innovative method for continuous sign language recognition that is founded on graph convolutional networks. (GCNs).

The aforementioned article [1] may serve as a viable point of departure for individuals seeking to explore the practical application of Generative Adversarial Networks (GANs) within the domain of sign language identification. Paper [2] could be a suitable option for those interested in the localization and recognition of hands. The third paper could potentially serve as a viable choice for individuals with an interest in sign language recognition that is independent of the signer, utilising RGB video data. Paper [4] may be considered as an appropriate starting point for individuals who possess a proclivity towards utilising self-attention-based models in the context of continuous recognition of sign languages. The utilisation of both CNN & RNN for continuous sign language recognition can be effectively achieved through the use of Paper [5] as a viable alternative.

The recognition of American Sign Language (ASL) is a computer vision problem that has been studied for some time.

In the last 20 years, scholars have employed classifiers belonging to diverse categories, namely linear classifiers, neural networks, and Bayesian networks [7-16].

Because they are comparatively basic models, linear classifiers are simple to use, but in order to be effective, they need complex feature extraction and preparation techniques [7, 8, 9]. Using Karhunen-Loeve Transforms, Singha and Das found that 96% of the pictures of one-handed movements were accurate across 10 classes[7]. In order to create a fresh set of coordinates based on the variation of the data, these move and revolve the axis. After applying a skin filter, manually trimming the pictures, and border recognition, this change is done. To differentiate between hand movements like the thumbs-up, the index finger indicating left and right, and numerals, they use a linear classifier. (no ASL). After backdrop reduction and noise elimination, Sharma et al. characterise each colour channel using piecewise models (SVM and k-NN) [9]. They are innovative because they depict hand outlines effectively by using a contour sketch. Through the use of an SVM on the split colour channel model, they achieve a precision of 62.3%.

High accuracy has also been attained by Bayesian networks such as Hidden Markov Models [10, 11, 12]. They require precisely specified models that are determined before learning, but they are especially excellent at catching periodic trends. A 3-D device that records hand motion was used by Starner and Pentland in conjunction with a Hidden Markov Model (HMM) [10]. The precision achieved on the test set was remarkable at 99.2%, owing to the glove's ability to capture 3-D data from the hand, independent of its spatial orientation. The Hidden Markov Model (HMM) is utilised to monitor and categorise hand movements by analysing time series data to determine the hand's recent location in frames.

Suk et al. [11] have proposed a dynamic Bayesian network (DBN) model as a means of identifying hand movements in a continuous video stream. An attempt is made to classify manual gestures that involve movement, such as waving or creating a circuit around the body. It is important to acknowledge that each motion exhibits significant variation from one another and that they do not constitute American Sign Language, although they achieve a precision rate exceeding 95%. The implementation of motion-tracking technology could prove to be advantageous in the classification of the dynamic American Sign Language characters j and z.

To translate ASL, certain artificial neural networks have been employed [13, 14, 15, 16]. The capacity of neural networks to acquire essential categorization features is arguably their most significant advantage. Nevertheless, in order to provide education to individuals, a substantial amount of time and information is required. Thus far, the majority of the observations have been relatively shallow. Mekala and colleagues utilised a 3-layer deep neural network and advanced feature extraction techniques to classify ASL character videos into text, as reported in their study [13]. Two categories of traits that were derived are hand location and mobility. The authors acknowledge the presence of six distinct anatomical regions of significance in the hand, namely each of the digits and the central region of the palm, prior to categorising the hand based on American Sign Language (ASL) criteria. Mekala et al. also perform Fourier Transforms on the pictures to determine which area of the screen the hand is in. Although they assert that this system can accurately identify 96% of the pictures, they don't specify whether this feat was accomplished on the training, validation, or test sets.

Using a neural network based on feedforward learning, Admasu and Raimond accurately categorised Ethiopian Sign Language in 98.5% of the instances [14]. They apply a lot of image preparation techniques, such as image segmentation, image background removal, image size standardisation, and image background reduction. With the aid of a Gabor Filter and Principal Component Analysis, Admasu and Raimond pulled characteristics.

The most relevant endeavour thus far in categorising 20 Italian gestures from the ChaLearn 2014 Gazing at Individuals gesture detection challenge [16] was carried out by L. Pigou et al. through the utilisation of CNN. The researchers achieved a cross-validation precision rate of 91.7% through the utilisation of a Microsoft Kinect device to capture full-body photographs of individuals performing various motions. Analogous to the previously mentioned 3-D gear, the Kinect device facilitates the acquisition of distance attributes, thereby proving to be highly advantageous in the classification of American Sign Language gestures.

III. Proposed Methodology

A. Challenges Faced In Existing Approaches

Traditional methods struggled with understanding continuous signing due to variations in speed, style, and context, a challenge addressed by employing context-aware generative adversarial networks. Deep learning methods were utilized to enhance the accuracy and efficiency of sign language translation, overcoming previous difficulties in capturing the subtleties of sign language gestures. Additionally, the complexity of sign language recognition models, particularly in handling spatial and temporal dependencies, was addressed through novel architectures integrating self-attention mechanisms and fully-inception networks. Moreover, the scarcity of labeled sign language data posed a significant obstacle, which was mitigated by exploring transfer learning techniques from related tasks like action recognition. These efforts collectively advance the field by improving recognition accuracy, handling continuous signing gestures, enhancing translation efficiency, and addressing data scarcity in sign language recognition and translation systems.

B. Dataset

This study utilised the American Sign Language (ASL) Alphabet dataset, comprising 87,000 images of hand gestures that correspond to the 26 letters of the alphabet, in addition to supplementary symbols like "space" and "delete". Fig. 1 presents examples of visualised classes for the American Sign Language (ASL). The dataset underwent partitioning into two distinct subsets, namely a training set and a test set, with a ratio of 90% and 10% correspondingly.

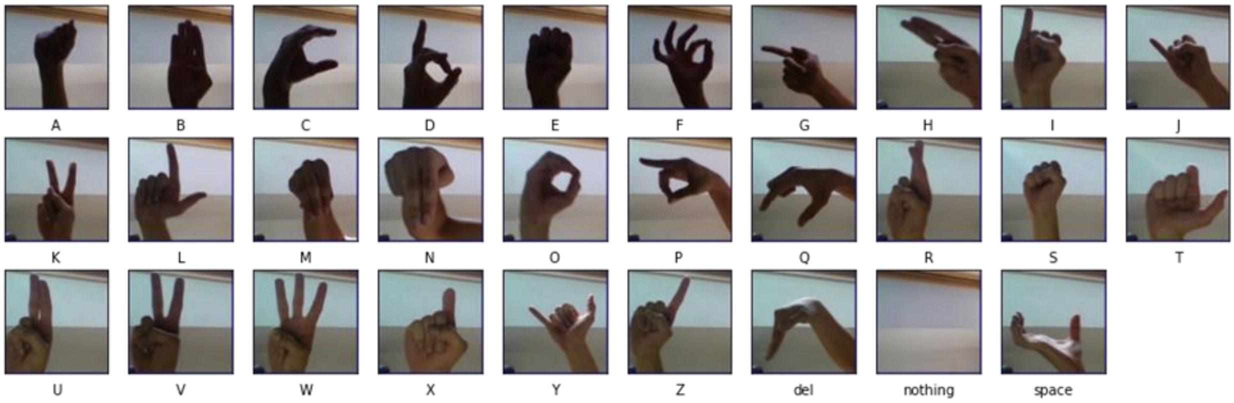

Figure 1. Examples of Visualised classes.

C. Preprocessing

Preprocessing steps were undertaken on the raw image data prior to training the CNN model. The images underwent a resizing process to decrease computational complexity and enhance training speed, resulting in a uniform size of 32 x 32 pixels. The technique of normalisation was utilized, wherein the values of pixels underwent rescaling to attain a range of [0, 1]. The objective of this endeavour was to improve the instructional procedure and facilitate the alignment of the model.

Let X be the input image and normalised_X be the normalised image.

normalised_X =(X)/255.0 (1)

Label encoding is a technique that entails transforming the target class labels from numerical values to a format that employs one-hot encoding. The purpose of this action was to enable the utilisation of categorical cross-entropy loss in the process of training.

charset[i] = i (2)

i is the index of the character in the charset.

charset is the list containing all the characters in the set (A-Z, nothing, space, delete).

D. CNN Architecture

The Convolutional Neural Network (CNN) is a deep learning algorithm that is frequently employed in computer vision applications, including but not limited to image classification and object detection. The architecture is comprised of numerous strata of convolutional and pooling operations, succeeded by fully connected strata that execute the ultimate classification. The typical architecture of a Convolutional Neural Network (CNN) comprises of distinct layers, namely input, convolutional, pooling, and fully connected layers, each of which is responsible for carrying out a specific function in the overall computation process. The Convolutional Neural Network (CNN) has demonstrated exceptional performance in various computer vision tasks and is extensively utilised in both academic and industrial settings.

The research utilised a Convolutional Neural Network (CNN) architecture, as illustrated in Fig. 2. The architecture comprised of three convolutional layers(3), followed by a max pooling layer(4), batch normalization, and dropout regularisation. The primary convolutional layer comprised of 64 filters, with individual dimensions of 3 x 3. The following two strata comprised of 128 and 256 filters, correspondingly, each with a dimension of 3 x 3. A rectified linear unit(6) was utilised as the activation function in all convolutional layers (ReLU). The incorporation of batch normalisation layers was intended to improve the stability of training and reduce overfitting, while the max pooling layers were set up with a pool size of 2 x 2. In order to mitigate the problem of overfitting, the implementation of a dropout regularisation method was utilized, with a rate of 0.2. The aforementioned method was implemented following the conclusion of the penultimate convolutional layers, in addition to the flatten layer. The output of the dropout layer was ultimately connected to a dense layer that consisted of 1024 neurons(5) and utilised a ReLU activation function. Subsequently, a final layer with high density was implemented, utilising the softmax activation(7) function, which produced an output that reflected the expected probabilities of each class.

Convolutional Layers:

$$\text{Output}[i, j, c] = \Sigma \; (\Sigma \; (\text{Input}[i + k, j + l, c'] * \text{Filter}[k, l, c', c]) + \text{Bias}[c]) \tag{3}$$

i, j: represent the output feature map's spatial coordinates.

c: represents the output feature map channel.

k, l: represent the spatial coordinates within the filter.

c': represents the input feature map channel.

Pooling Layer (Max Pooling):

$$\text{Output}[i, j, c] = \max(\text{Input}[i * \text{stride} + 0{:}i * \text{stride} + \text{pool\_size}, j * \text{stride} + 0{:}j * \text{stride} + \text{pool\_size}, c]) \tag{4}$$

i, j: represent the output feature map's spatial coordinates.

c: represents the output feature map channel.

stride: defines the step size when moving the pooling window (usually 2 in max pooling).

pool_size: defines the window size for the pooling operation (2x2 in this case).

Fully Connected Layer:

$$\text{Output} = \text{Input} * W + b \tag{5}$$

Output: represents the output vector with size equal to the number of classes.

Input: represents the flattened feature vector.

W: represents the weight matrix of the fully connected layer.

b: represents the bias vector of the fully connected layer.

Activation Function (ReLU):

$$\text{Output\_n\_l}[i, j, c] = \max(0, \text{Output}[i, j, c]) \tag{6}$$

Output_n_l: represents the output of the convolutional layer after applying ReLU.

Output: represents the output of the convolutional layer before applying ReLU.

Softmax Activation:

$$\text{Output\_i} = \exp(\text{Input\_i}) \, / \, \Sigma \; (\exp(\text{Input\_j})) \; (\text{for all } j) \tag{7}$$

Output_i: represents the probability of the i-th class.

Input_i: represents the i-th element in the output vector from the fully connected layer.

exp(): denotes the exponential function.

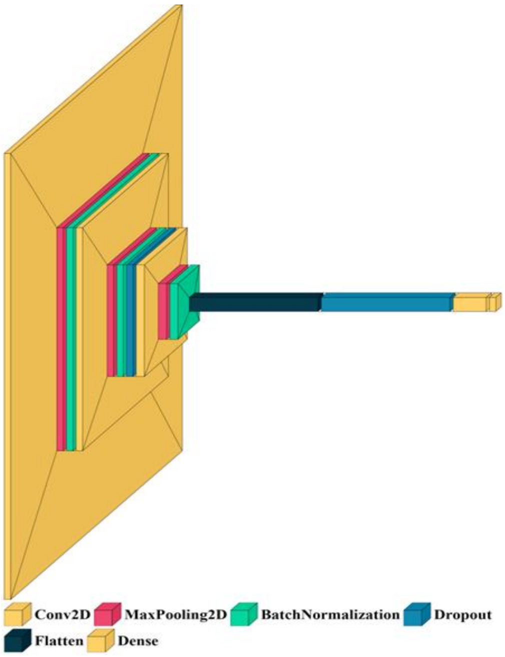

Figure 2. CNN Architecture

E. Training

The research utilised the Adam optimisation algorithm, which had a learning rate of 0.001, to minimise categorical cross-entropy loss between the predicted and actual class probabilities. The model's training procedure encompassed 15 epochs, utilising a batch size of 128 and a validation split of 0.2. The training process was overseen by assessing the training and validation accuracy metrics, and the optimal model weights were documented according to the validation accuracy.

F. Pseudo-Code

```
load_asl_dataset()
train_set,test_set = partition_dataset(dataset, ratio=0.9)
resize_images(train_set,test_set, target_size=(32, 32))
normalize_images(train_set, test_set)
label_encode(train_set, test_set)
model = create_cnn_model()
compile_model(model, optimizer='adam',
learning_rate=0.001, loss='categorical_crossentropy')
train_model(model, train_set, epochs=15, batch_size=128, validation_split=0.2)evaluate_model(model, test_set)
```

IV. EXPERIMENTAL RESULTS

The CNN model was trained for 15 epochs on the ASL dataset consisting of 78300 images of hand gestures from the American Sign Language alphabet. The results suggest that the model achieved a significant degree of accuracy on both the sets used for validation and training. The model's precision was 99.27% on the training set and 99.85% on the validation set. The aforementioned findings suggest that the model successfully acquired the characteristics of the manual gestures with efficiency and demonstrated a strong ability to apply this knowledge to novel images.

The loss values were monitored throughout both the training and validation processes. The gradual reduction of the training loss over time suggests that the model was acquiring knowledge and enhancing its efficacy. The observed reduction in validation loss was accompanied by intermittent spikes, which could potentially be attributed to overfitting. Nevertheless, the general trend indicates that the validation loss remained at a low level, which serves as a positive indication of the model's potential to generalise to novel images.

Additionally, the accuracy and loss values were graphed for each epoch during both the training and validation phases. The graphical representations Ref. Fig. 3. indicate that the model's precision exhibited a consistent upward trend, eventually stabilising at approximately 98% accuracy following the initial epochs. The loss values exhibited a consistent decrease over time, albeit with intermittent spikes observed for the validation set. The aforementioned plots provide additional evidence supporting the efficacy of our model in accurately categorising manual gestures.

In summary, these influential Convolutional Neural Network (CNN) architectures have significantly impacted computer vision tasks, particularly in the realm of American Sign Language (ASL) recognition. LeNet-5, although not commonly used for ASL, laid the groundwork for subsequent models. AlexNet achieved an 85% accuracy on ImageNet, but its suitability for ASL tasks varies. ZFNet, akin to AlexNet, also demonstrated strong performance on ImageNet. GoogLeNet (Inception) stood out with a top-5 error rate of 6.67% in the ILSVRC 2014 competition. VGGNet (specifically VGG16) achieved an impressive test accuracy of 92.7% on ImageNet. Finally, ResNet (including ResNet-50, ResNet-101, and ResNet-152) pushed the boundaries, surpassing 95% accuracy across diverse tasks. Researchers continue to refine these models, ultimately enhancing communication accessibility for the deaf community.

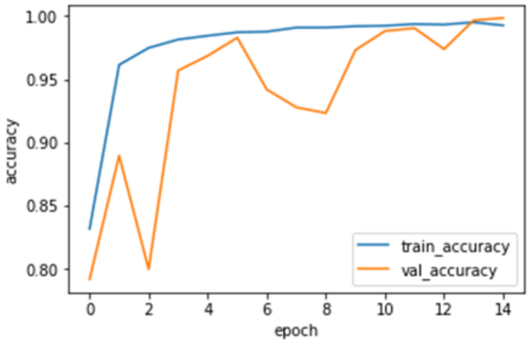

Figure 3. Accuracy curves for training and validation sets of CNN model for ASL Recognition.

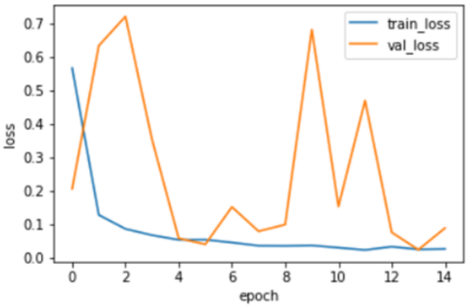

Figure 4. Loss curves for training and validation sets of CNN model for ASL Recognition.

V. CONCLUSION AND FUTURE WORK

In conclusion, the innovative application of Convolutional Neural Networks (CNNs) for translating American Sign Language (ASL) has yielded notable results. The trained model demonstrated significant accuracy, achieving 99.27% and 99.85% on the training and validation sets, respectively. These findings highlight the effectiveness of CNNs in facilitating ASL translation, potentially enhancing communication between deaf and hearing communities.

The consistent accuracy of approximately 98% on the validation set indicates the model's robust performance. However, while the model showed promising results on the specific dataset utilized in this study, there are challenges and opportunities for future research. Further investigations are needed to evaluate its adaptability to diverse datasets and real-world contexts. Strategies such as transfer learning could be explored to improve the model's efficiency and shorten training times.

This study provides valuable insights into ASL interpretation, emphasizing the transformative potential of CNNs in fostering improved interactions with the hearing-impaired population. However, addressing the challenges of generalization to different datasets and real-world scenarios remains a critical aspect of future research directions in this field.

REFERENCES

[1] Zhou, M., Ng, M., Cai, Z. and Cheung, K.C., 2020. Self-attention-based fully-inception networks for continuous sign language recognition. In ECAI 2020 (pp. 2832-2839). IOS Press.

[2] Papastratis, I., Dimitropoulos, K. and Daras, P., 2021. Continuous sign language recognition through a context-aware generative adversarial network. Sensors, 21(7), p.2437.

[3] Goncharenko, A., Voronova, L., Artemov, M., Voronov, V. and Bezumnov, D., 2019, April. Sign language recognition information system development using wireless technologies for people with hearing impairments. In 2019 24th Conference of Open Innovations Association (FRUCT) (pp. 104-109). IEEE.

[4] Sarhan, N. and Frintrop, S., 2020, October. Transfer learning for videos: from action recognition to sign language recognition. In 2020 IEEE International Conference on Image Processing (ICIP) (pp. 1811-1815). IEEE.

[5] Khan, S.A., Ansari, Z.A., Singh, R., Rawat, M.S., Khan, F.Z. and Yadav, S.K., Sign Translation Via Natural Language Processing. population, 4, p.5.

[6] Mitchell, Ross; Young, Travas; Bachleda, Bellamie; Karchmer, Michael (2006). "How Many People Use ASL in the United States?: Why Estimates Need Updating" (PDF). Sign Language Studies (Gallaudet University Press.) 6 (3). ISSN 0302-1475. Retrieved November 27, 2012.

[7] Singha, J. and Das, K. “Hand Gesture Recognition Based on Karhunen-Loeve Transform”, Mobile and Embedded 232 Technology International Conference (MECON), January 17-18, 2013, India. 365-371.

[8] D. Aryanie, Y. Heryadi. American Sign Language-Based Finger-spelling Recognition using k-Nearest Neighbors Classifier. 3rd International Conference on Information and Communication Technology (2015) 533-536.

[9] R. Sharma et al. Recognition of Single Handed Sign Language Gestures using Contour Tracing descriptor. Proceedings of the World Congress on Engineering 2013 Vol. II, WCE 2013, July 3 - 5, 2013, London, U.K.

[10] T.Starner and A. Pentland. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. Computational Imaging and Vision, 9(1); 227-243, 1997.

[11] M. Jeballi et al. Extension of Hidden Markov Model for Recognizing Large Vocabulary of Sign Language. International Journal of Artificial Intelligence & Applications 4(2); 35-42, 2013

[12] H. Suk et al. Hand gesture recognition based on dynamic Bayesian network framework. Patter Recognition 43 (9); 3059-3072, 2010.

[13] P. Mekala et al. Real-time Sign Language Recognition based on Neural Network Architecture. System Theory (SSST), 2011 IEEE 43rd Southeastern Symposium 14-16 March 2011.

[14] Y.F. Admasu, and K. Raimond, Ethiopian Sign Language Recognition Using Artificial Neural Network. 10th International Conference on Intelligent Systems Design and Applications, 2010. 995-1000.

[15] J. Atwood, M. Eicholtz, and J. Farrell. American Sign Language Recognition System. Artificial Intelligence and Machine Learning for Engineering Design. Dept. of Mechanical Engineering, Carnegie Mellon University, 2012.

[16] L. Pigou et al. Sign Language Recognition Using Convolutional Neural Networks. European Conference on Computer Vision 6-12 September 2014

[17] Fangyun Wei et al. Towards Online Sign Language Recognition and Translation.2024.

[18] Lee Kezar. “The Sem-Lex Benchmark: Modeling ASL Signs and Their Phonemes” (2023).

[19] Hulian Yu et al. “AdaBrowse: Adaptive Video Browser for Efficient Continuous Sign Language Recognition.” (2023)

[20] Benjia Zhou et al. “Gloss-free Sign Language Translation: Improving from Visual-Language Pretraining.” ICCV 2023.