American Sign Language detection using Multi-Modal CNN

Sai Nalla
700763475
dept.Computer Science
University of Central Missouri
sxn34750@ucmo.edu

Anish Koppula
700759641
dept.Computer Science
University of Central Missouri
axk96410@ucmo.edu

Deekshitha Gaddameedhi
700755765
dept.Computer Science
University of Central Missouri
dxg57650@ucmo.edu

Bala RishikMarneni
700746746
dept.Computer Science
University of Central Missouri
Bxm67460@ucmo.edu

***Abstract*—Sign languages serve as vital communication tools for the deaf community, relying on visual and manual expressions when verbal communication is challenging. One prominent example is Indian Sign Language (ISL), utilized within educational settings to impart academic knowledge to deaf students. Sign Language plays a crucial role in fostering communication and inclusivity within the deaf community. However, sign language gestures are intricate, involving intricate hand shapes and movements that RGB images alone may not fully capture. This limitation can hinder accurate recognition and interpretation, particularly when gestures bear resemblance to one another. Depth information, on the other hand, offers valuable insights into the spatial relationships among various hand components and between the hand and surrounding elements. In this paper, we are proposing a Multi-Modal Convolutional Neural Network(CNN) which takes the color and depth as input. This architecture improves the classification accuracy compared to the existing models. The model achieved 75 percent accuracy.**
[1]

***Index Terms*—ASL(American Sign Language),Multi-modalCNN,RGB-D,ISL(Indian Sign Language),Hand gestures.**

I. Introduction

Sign languages are conveyed through visual and manual methods when speaking the language through the mouth is difficult. Initially, signed languages were developed to assist people with hearing impairment. However, it has other advantages as well. Sign language is a helpful tool for children who are affected by Autism Spectrum Disorder (ASD). These children struggle to communicate verbally, and sign language helps them to communicate effectively. Children affected by this disease struggle to communicate verbally so sign language helps them communicate.

Existing systems have robust sign language detection systems but the difficulty is to find the names, and brand names traditional methods use hand tracing but they are not accurate and visually challenging. The best method is to use fingerspelling to communicate names, brand names, and other words that are difficult to explain through hand gestures. Current methods are not accurate in generalizing gesture recognitions due to limited data availability.

Recent works on sign language recognition suggest that using multimodal data like combining the sign language data improves the generalization of results. For example, combining video sequences with descriptions of the gestures.

Accurate results of sign language systems depend on the two major tasks one is segmenting the hand part and the other one is after segmenting extracting the features of the image. In this study, a deep

[1]https://github.com/BalaRishik001/American-Sign-Language-Recognition-A-Deep-Learning-Approach

learning approach PCANet is proposed to classify the finger spelling that is the alphabet.PCANet is the unsupervised learning deep learning architecture. Experimental analysis is conducted on the publicly available real-depth images to validate the image's leave-one-out approach.

In this study, quiz-based sign language recognition is implemented in the form of a web application. The application allows the users to learn and evaluate sign language. In the learning phase, they learn sign language and how to pose and make gestures. In the testing phase, they can check their learning.

In this paper, we are proposing finger spelling recognition using various deep learning models like CNN as a baseline model and DenseNet models. The experimental analysis will be conducted on the ASL (American Sign Language) alphabet dataset which consists of alphabets. To enhance the accuracy of the model data augmentation methods are implemented. The applications of this project can be extended to finger spelling.

In the first step, the hand gesture image is preprocessed to remove noise from the images. In the second step, a preprocessed image is passed to the classifier to recognize the alphabet. Different data augmentation techniques consider background variations and hand articulation variations. In this step, various deep learning models are built and tested to check the performance. This is used to generate the fingerspelling of various names. To develop this project Python programming language is used. The major frameworks used in the project are TensorFlow and Opencv.

Project contributions: Our research contributes to the field of finger spelling recognition by implementing a diverse set of deep learning models and employing effective data augmentation strategies. We aim to enhance accuracy and robustness in recognizing finger-spelled alphabets, particularly in the context of American Sign Language

II. Motivation

Sign languages play a crucial role in facilitating communication within the deaf community. However, the intricate nature of sign language gestures poses significant challenges to accurate recognition and interpretation. Depending on RGB images will not capture those intricacies. Recognizing these challenges, our research is motivated by the need to address them effectively. We aim to propose a novel approach to sign language recognition by leveraging Multimodal Convolutional Neural Network (CNN) architecture. By integrating both color (RGB) and depth information, our goal is to enhance the accuracy and robustness of sign language recognition systems.

III. Objectives

The main objectives of this project are:

- Enhancing Classification Accuracy: The primary objective is to improve the accuracy of sign language gesture classification compared to existing models.
- Utilizing Multi-Modal Information: Another objective is effectively utilizing information from multiple modalities (RGB and depth) to capture the nuances of sign language gestures.
- Data gathering: In the data gathering step in addition to the ready-to-use parameters new parameters are also used to check the model performance and their importance in the prediction of attrition the features are: previous work experience, the reason for leaving, and previous company salary.Data preparation: Collected raw data is cleaned and prepared to match the requirements of the predictive models
- Promoting Accessibility: A broader objective may involve promoting accessibility and inclusion for the deaf community. By developing a robust prediction system.
- Scalability and Generalization: Another aim could be to design a model that is scalable and generalizable

IV. Dataset Description

These datasets feature recordings of 5 distinct non-native signers demonstrating each letter (excluding "X" and "Z" due to their reliance on movement) in the alphabet. Sample images are shown in Figure 1. The data is captured using a Microsoft Kinect sensor, providing comprehensive visual and depth information for accurate analysis and training of ASL recognition models. The depth images are saved as single-channel short unsigned int images.

Dataset A features 24 static signs. This was captured in 5 different sessions, with consistent lighting and background conditions. Dataset B Comprises depth-only data captured from 9 different individuals.

Figure 1. sample dataset

V. Related work

This article addresses the challenge of developing a high-quality sign language dataset for Thai fingerspelling. The researchers gathered information from 43 individuals with varied backgrounds, genders, and physical characteristics of 24 fundamental Thai fingerspelling signs. To tackle this challenge, the study explores six deep learning architectures: RGB-sequencing-based CNN LSTM and VGG-LSTM, LSTM, BiLSTM, and GRU ST-GCN model. The article investigates individual modalities as well as their combinations. The findings indicate that combining the RGB-sequencing modality from VGG-LSTM achieved the highest performance [18].

Indian Sign Language (ISL) is commonly used in India. Nowadays, online interpreters are available, but they often require experts for translation, making them uneconomical at times. Therefore, this work focuses on a vision-based fingerspelling system using Convolutional Neural Networks (CNNs). The proposed CNN model performs well with InceptionV3 and outperforms existing models such as ResNet and VGG16 [12].

This project focuses on real-time recognition of American Sign Language (ASL) finger-spelling, utilizing a Deep Convolutional Neural Network (CNN) for classification. The aim is to facilitate communication between hearing-impaired individuals and the general population. The dataset consists of over 50,000 images covering the ASL alphabet (A to Z) and numerals (0 to 9), capturing the distinct characteristics and variations in hand shape and position for each sign. The proposed model achieves an accuracy of 99.70 in classification [7].

Indian Sign Language(ISL) is used in educational institutions to teach academics to deaf students. This plays a pivotal role in communication and inclusion for the deaf community. This project aims to create unified alphabets in the Malayalam language. For this task, ResNet50 a transfer learning model is used. The proposed model achieved 97 percent training accuracy and 93 percent validation accuracy [10].

This study focuses on utilizing DNN models for recognizing hand gestures from the American Sign Language (ASL). Specifically, the research involves gathering a dataset comprising 24 ASL hand gestures and implementing the architectures of SqueezeNet and MobileNets. Through experiments conducted on the American Sign Language-Finger Spelling (ASL-FS) dataset, the findings demonstrate that the proposed approach has average accuracy compared to existing methods [11].

Understanding sign language is challenging. Researchers have spent several years trying to understand the patterns. If we consider the problem as a computer vision problem it becomes more challenging. There are several sign languages in the world: American, French, etc., In India, Bangladesh has around 9 million people with hearing impairment which is a high number. There is less amount of work done on Bangla sign language. This study focuses on building sign language detectors for Bangla speakers. The main challenge here is the amount of data available is very small. Deep learning models need a high amount of data to process. This study aims to address these problems [19].

The methodology involves a two-step process consisting of gesture refinement and classification. In the first step, preprocessing and refining of hand gestures include noise reduction. After preprocessing the gestures are classified using the CNN model which is a multi-class classification of 26 alphabets and one space sign. The American Sign Language on-screen text representation model achieved 95.7 percent accuracy [1].

Data for the experiment includes 117 videos explaining emergencies in Costa Rican Sign Language (LESCO). The data contains multiple frames and multiple hands in each frame. This makes the encoding process very challenging. The study overcomes this by leveraging Google MediaPipe.The methodology involves encoding videos into one-dimensional vectors, augmenting video data, and employing various machine learning methods for sign recognition [20].

The current landscape of sign language learning systems predominantly relies on costly external sensors, limiting accessibility. This research proposes feature extraction techniques and supervised learning algorithms novel with fourfold cross-validation. This study highlights the importance of robust validation methodologies with improved generalization [**?**], [4].

The main objectives of this paper are to develop a key-point detection system with Long Short-Term Memory (LSTM) models for hearing-impaired people. Evaluating the model and exploring the potential applications [14].

Gesture recognition presents numerous challenges, with the primary obstacle being the intricate nature of expressions. Additionally, attaining high precision poses a significant difficulty. Interclass similarities among gestures further complicate accurate recognition. Lastly, challenges also arise from finger occlusion and variations in hand shapes. To overcome this we are proposing a multimodal gesture recognition system that is RGB and depth [13].

The primary objective is to enhance communication between students and teachers by translating hand gestures into meaningful symbols and helping the students in arithmetic operations. The project aims to enable the integration of sign language into digital communication platforms. The model is tested on Panamanian Sign Language hand shapes and American Sign Language (ASL) hand shapes. The model achieved 88 percent accuracy on validation data [17].

Existing systems use alphabet level signs to start the communication but this will not make the communication smooth. To introduce advanced level communication we introduce syllable level sign language recognition. For the experimental analysis KETI Korean language dataset is used and a syllable-level fingerspelling recognition framework that utilizes characteristics of Hangul is implemented. With this approach we have identified more improved communication amongst the community [8].

The research proposes a method for static Arabic sign language detection of 32 classes. The deep learning model AlexNet is used for classification loss function categorical cross-entropy. Through experimental validation on the ArSL2018 dataset, an F-measure of 97.38, indicating the method's precision in Arabic sign language recognition [6].

So far the existing systems use sign language detection systems but there is a lack of learning platforms for sign language. In this paper, we are proposing a SignQuiz for learning sign language. To implement this we have used deep learning architectures to teach. In the results analysis, we have found that learners can learn without any external help, and the performance is far better compared to the other mediums [9].

The paper proposes a network called CrossFeat which is a Multiscale Cross Feature Aggregation network. The proposed network is investigated on three benchmark datasets: ASL FingerSpelling, NUS-I, and NUS-II. This network CrossFit employs multi-scale filters: 1×1, 3×3, 5×5, and 7×7 allow the network to learn granular and coarse edges regions of the hand gestures. The experimental results and analysis show that the aggregation of multi-scale and cross features enhances the performance [3].

The paper presents the real-time recognition of 26 alphabet hand gestures in ASL. The system has several modules, including pre-processing, training, and testing, and achieved an accuracy of 95.8. For this project, we utilized deep learning, OpenCV, and TensorFlow as tools and technologies [2].

In this paper, we propose a novel approach to neuroevolution that optimizes the classification of low-dimensional datasets. A dataset of 1678 images. On the benchmark dataset three neuroevolution simulations are executed a mean 10-fold cross-validation is performed and achieved an accuracy of 97.44 The results show that the simulation finds a promising set of hyperparameters [5].

The proposed method uses the DRCAM model

incorporating an attention mechanism and cascading residual for real-time hand gesture recognition. The proposed architecture effectively captures features of hand gestures across various levels, from low to high. Dense connectivity within the network amplifies the propagation and reuse of these features. Experimental evaluations, conducted on both in-house and American Sign Language finger-spelling benchmark datasets, demonstrate superior performance [15].

This paper introduces a novel phonology-based approach to evaluate generated videos in sign language. Unlike traditional metrics like PSNR and SSIM, this method assesses linguistic information, focusing on hand movement and handshape channels. Objective scores are obtained by comparing class conditional probabilities extracted from source and generated videos using dynamic time warping. Experimental results show that this approach correlates better with subjective human ratings compared to conventional metrics like PSNR, SSIM, and MSE [16].

VI. Proposed Framework

A. Data collection and preparation

Data collection and preprocessing: The data is collected from the open-source repository Kaggle. The images are RGB with 3 channels and depth parameters. Collected data is resized and normalized according to the chosen model architecture.

Data is organized into color and depth images to feed the model.

B. Model architecture

Figure.2. represents MMCNN architecture takes multiple modalities of input data. In the context of ASL detection, this typically includes RGB images and depth images. The architecture typically consists of separate branches for each input modality. In the case of ASL detection, there would be one branch for processing RGB images and another for processing depth images. After processing each modality independently, the model incorporates shared layers or parameters to combine the information learned from both branches.

Following the fusion of multi-modal features, the architecture typically includes classification layers for making predictions. These layers may consist of fully connected (dense) layers followed by softmax activation for multi-class classification tasks like ASL detection. The model learns to map the combined features to the corresponding ASL classes, thus classifying the sign language gestures

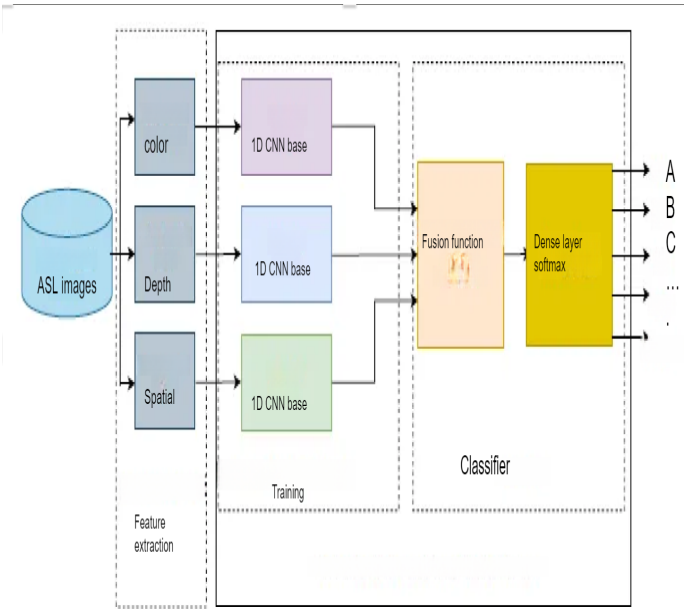

Figure 2. Muti Modal CNN architecture

C. Training

Data is split into training and testing in an 80/20 ratio. The sequential model is built using TensorFlow. After building the model it is trained for 10 epochs and the model achieved 75 percent accuracy which is promising. To show the results visually we have visualized the model accuracy and loss plot refer to Figure 3, Figure 4.

D. Testing

The model was tested on the validation dataset and achieved 67 percent accuracy which is promising. The model is saved using Keras.

VII. Results summary

A. Model performance

Model performance is evaluated using the confusion matrix shown in Figure 5 and the classification report shown in Figure 6. The confusion matrix gives the number of misclassifications and the classification report gives the qualitative report of predictions

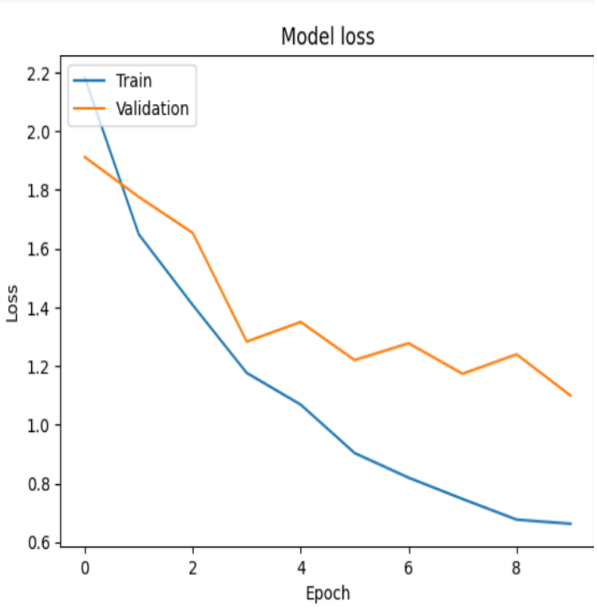

Figure 3. Model loss plot

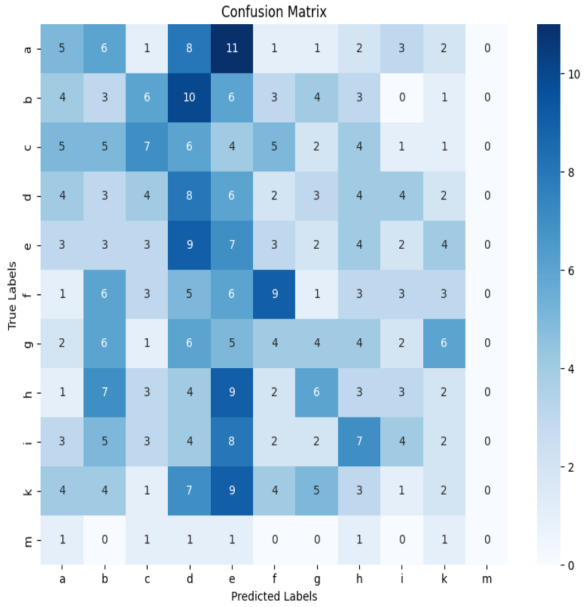

Figure 5. Classification report-Ada boosting

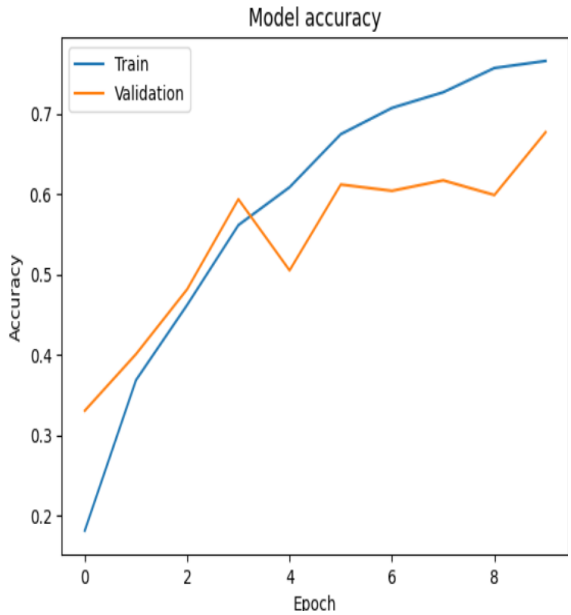

Figure 4. Model accuracy plot

Classification Report:

	precision	recall	f1-score	support
a	0.15	0.12	0.14	40
b	0.06	0.07	0.07	40
c	0.21	0.17	0.19	40
d	0.12	0.20	0.15	40
e	0.10	0.17	0.12	40
f	0.26	0.23	0.24	40
g	0.13	0.10	0.11	40
h	0.08	0.07	0.08	40
i	0.17	0.10	0.13	40
k	0.08	0.05	0.06	40
m	0.00	0.00	0.00	6
accuracy			0.13	406
macro avg	0.12	0.12	0.12	406
weighted avg	0.13	0.13	0.13	406

Figure 6. Confusion matrix Ada Boosting

B. *Sample predictions*

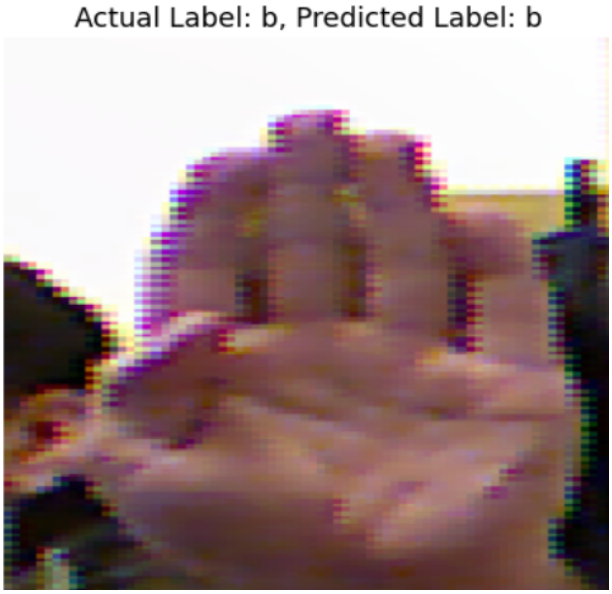

Figure 7. Classification report-Ada boosting

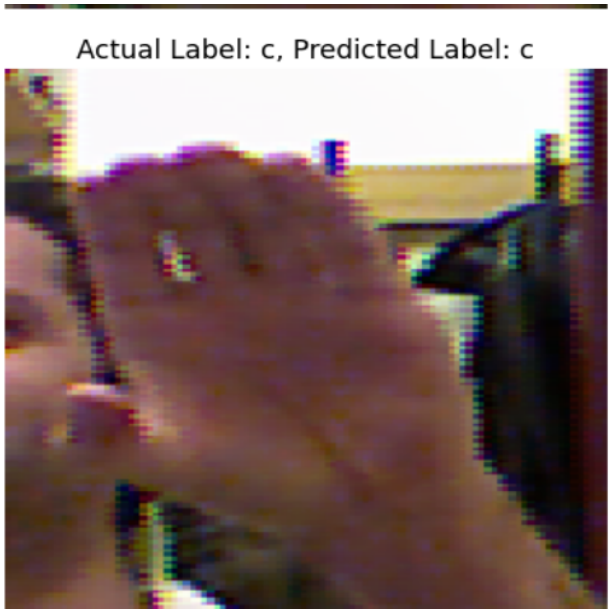

Figure 8. Confusion matrix Ada Boosting

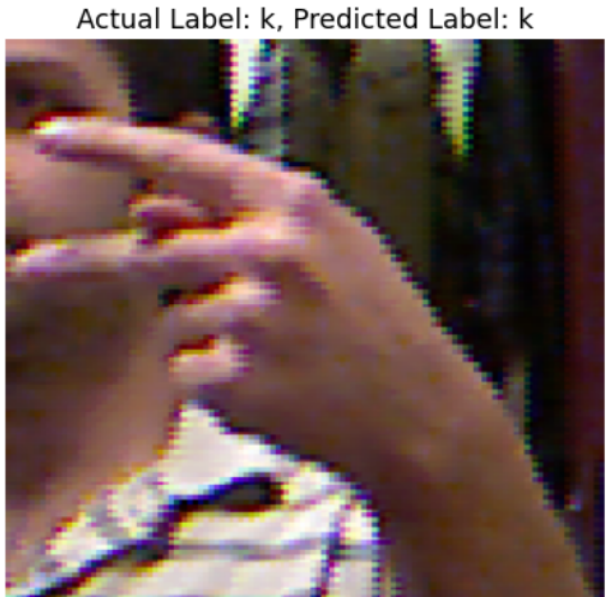

Figure 9. Classification report-Ada boosting

Figure 10. Confusion matrix Ada Boosting

REFERENCES

[1] Soumya Ashwath and Ashwin Shenoy M. Neural network-based real-time recognition of american sign language finger-spelled gestures: Bridging communication gaps. In *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, pages 170–174, 2023.

[2] Anagha Bhardwaj, Akshita Singhal, Prakhar Mamgain, Utkarsh Joshi, and Siddhant Thapliyal. A real time conversion model for hand gestures to textual content. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–7, 2023.

[3] Gopa Bhaumik, Monu Verma, Mahesh Chandra Govil, and Santosh Kumar Vipparthi. Crossfeat: Multi-scale cross feature aggregation network for hand gesture recognition. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 274–279, 2020.

[4] A. Bhavana, K Shalini Reddy, Madhu, and D Praveen Kumar. Deep neural network based sign language detection. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 1474–1479, 2022.

[5] Jordan J. Bird, Isibor Kennedy Ihianle, Pedro Machado, David J. Brown, and Ahmad Lotfi. A neuroevolution approach to keypoint-based sign language fingerspelling classification. In *2023 15th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, pages 215–220, 2023.

[6] I. Hmida and N. B. Romdhane. Arabic sign language recognition algorithm based on deep learning for smart cities. In *The 3rd International Conference on Distributed Sensing and Intelligent Systems (ICDSIS 2022)*, volume 2022, pages 119–127, 2022.

[7] Antara Howal, Atharva Golapkar, Yunus Khan, Siddhantha Bokade, Satishkumar Varma, and Madhura Vikram Vyawahare. Sign language finger-spelling recognition system using deep convolutional neural network. In *2023 5th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)*, pages 1–6, 2023.

[8] Yoonyoung Jeong and Han-Mu Park. Syllable-level korean fingerspelling recognition from a video. In *2021 21st International Conference on Control, Automation and Systems (ICCAS)*, pages 2206–2210, 2021.

[9] Jestin Joy, Kannan Balakrishnan, and M. Sreeraj. Signquiz: A quiz based tool for learning fingerspelled signs in indian sign language using aslr. *IEEE Access*, 7:28363–28371, 2019.

[10] Baby Syla L. and Salim A. Sign language recognition of malayalam alphabets using transfer learning. In *2023 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pages 1–4, 2023.

[11] Adam Ahmed Qaid Mohammed, Jiancheng Lv, and Md Sajjatul Islam. Small deep learning models for hand gesture recognition. In *2019 IEEE Intl Conf on Parallel Distributed Processing with Applications, Big Data Cloud Computing, Sustainable Computing Communications, Social Computing Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 1429–1435, 2019.

[12] Chakravartula Raghavachari and GA Shanmugha Sundaram. Deep learning framework for fingerspelling system using cnn. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 469–473, 2020.

[13] Rajesh George Rajan and P. Selvi Rajendran. Gesture recognition of rgb-d and rgb static images using ensemble-based cnn architecture. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1579–1584, 2021.

[14] Ankith S, Darshan G Naidu, Shreesha R Bhat, J N Sai Vamshi, and Vanishree M L. Dactylology interpretation using key-points detection and lstm. In *2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA)*, pages 7–11, 2023.

[15] Jaya Prakash Sahoo, Suraj Prakash Sahoo, Samit Ari, and Sarat Kumar Patra. Hand gesture recognition using densely connected deep residual network and channel attention module for mobile robot control. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11, 2023.

[16] Neha Tarigopula, Preyas Garg, Skanda Muralidhar, Sandrine Tornay, Dinesh Babu Jayagopi, and Mathew Magimai.-Doss. Content-based objective evaluation of artificially generated sign language videos. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3815–3819, 2024.

[17] Alvaro Teran-Quezada, Victor Lopez-Cabrera, Jose Carlos Rangel, and Javier E. Sanchez-Galan. Hand gesture recognition with convnets for school-aged children to learn basic arithmetic operations. In *2022 IEEE 40th Central America and Panama Convention (CONCAPAN)*, pages 1–6, 2022.

[18] Wuttichai Vijitkunsawat, Teeradaj Racharak, and Minh Le Nguyen. Deep multimodal-based number fingerspelling recognizer for thai sign language. In *2023 22nd International Symposium on Communications and Information Technologies (ISCIT)*, pages 99–104, 2023.

[19] Samiya Kabir Youme, Towsif Alam Chowdhury, Hossain Ahamed, Md. Sayeed Abid, Labib Chowdhury, and Nabeel Mohammed. Generalization of bangla sign language recognition using angular loss functions. *IEEE Access*, 9:165351–165365, 2021.

[20] Juan Zamora-Mora and Mario Chacón-Rivas. Costa rican sign language recognition using mediapipe. In *2022 International Conference on Inclusive Technologies and Education (CONTIE)*, pages 1–6, 2022.