Department of Information Technology
Faculty of Computing & Information Technology (FCIT)
Lahore, Pakistan

VisioHear - Vision-based Pakistan Sign Language & Speech Interpretation for Differently-abled Individuals (BeyondTheWords.com)

Prof. Dr. Muhammad Kamran Malik

Department of Data Science
University of the Punjab, Lahore, Pakistan

mailto:kamran.malik@pucit.edu.pk

Rayyan Shabbir

Department of Information Technology Faculty of Computing & Information Technology (FCIT), Lahore, Pakistan

mailto:rayanshabir1@gmail.com,
https://orcid.org/0009-0002-4355-5615

Syeda Mujab Fatima

Department of Information Technology Faculty of Computing & Information Technology (FCIT), Lahore, Pakistan

mailto:mujabfatima@gmail.com,
https://orcid.org/0009-0009-1534-1418

Huma Saeed

Department of Information Technology Faculty of Computing & Information Technology (FCIT), Lahore, Pakistan

mailto:humasaeed275@gmail.com,

Hamaad Ghafar

Department of Information Technology Faculty of Computing & Information Technology (FCIT), Lahore, Pakistan

mailto:hamaadsaab@gmail.com,

ABSTRACT

VisioHear is a groundbreaking project showcased on **BeyondTheWords.com**, aimed at transforming communication for deaf and mute individuals fluent in Pakistan Sign Language (PSL). Leveraging cutting-edge technology, the system converts sign language gestures into text and audio, facilitating meaningful conversations and dismantling language barriers. The platform, hosted on Beyond The Words, serves as an inclusive space fostering effective communication and societal engagement.

In this project, we adopted a multifaceted approach, employing state-of-the-art machine learning models such as ***Long Short-Term Memory (LSTM), SSD MobileNet, EfficientDet-d0, Faster R-CNN, and YOLOv5*** to detect and translate sign language gestures accurately. Our dataset, drawn from the overlap between PSL and Indian Sign Language (ISL), comprises keypoints extracted from videos of essential daily routine words. The LSTM model emerged as the top performer, demonstrating superior accuracy in consistently detecting and translating sign language gestures into text. Conversely, SSD MobileNet exhibited poor performance, despite extensive training. EfficientDet-d0 showed moderate success, while Faster R-CNN and YOLOv5 faced challenges in accurate detection, compromising their suitability for real-time applications. Integration of the LSTM model into the VisioHear system ensures accurate and real-time translation of Urdu sign language gestures, enabling effective communication between deaf and non-deaf users. Furthermore, the project's exploration of global technologies and methodologies for translating and detecting various sign languages contributes to advancing inclusive communication technologies.

Keywords: Computer Vision Deaf, LSTM, Mute, Real-time Translation, Sign Language Recognition, Temporal Modeling

Figure 1. BeyondTheWords Logo.

1 INTRODUCTION

Scholars in the studying of sign language conduct a vast of scientific research that is directed toward removing the communication barriers for the compound and diverse group of the hearing impaired populace, which collectively in establishing of technologies, methodologies as well as educational resources through which the lives of the Deaf and mute populace would be enriched. These numerous papers aim at creating datasets alongside together with recognition of mainly different types of sign languages, as an example, the researchers work in the Pakistan Sign Language (PSL). The spoken and written language of PSL, owing to its syntactical and grammatical characteristics, poses many challenges that can only be solved by effective interventions, to enhance communication between Deaf and hearing impaired people and those in the normal range of hearing. For example, a research which proposes a dataset based on hand configurations concerning Urdu alphabets pure underlines the necessity of developing culturally-sensitive materials.

Technological possibilities appear in multiple studies, including deep learning ASL translation technology known as DeepASL that uses sign language to provide the translation for the American Sign Language. This unobtrusive translation takes place at both the lexical and text-linguistic levels. Similarly, a sign language recognition and classification system by using computer vision is proposed for gestural sign language used in Urdu language by the disabled people to have a convenient way of speaking. The research seeking to compare Arabic and Indian Sign Languages utilizes vision-based schemes. For instance, the study on Arabic sign language interpretation presents a deep learning architecture that comprises both the CNN and RNN for addressing the problems that are encountered by a sign recognition system. Another study present a sign language generation approach for ISL and natural language based generation system for translating natural language into ISL signs.

Apart from sign language translation applications, the compilation of works touches issues like image captioning and action recognition based on videos. The adaptively decided attention model employed in image captioning research enables the model to either attend to the image or rely solely on the language model at a certain point in time. It has used a Transformer model specifically to teach computers how to understand sign language from videos, and it emphasizes on body movement which shows the multifaceted nature of this particular area of research which connects computer science, linguistics and cognitive science.

Some of the research works present novel systems and even prototypes. For example, a system based on Python, YOLOv4, and Support Vector Machine (SVM), integrated with Mediapipe, recognizes objects with a high accuracy of 98. 8% and 98. It was further reported that implementation of the new algorithm increased accuracy by 62%, and real-time prediction by 62%, respectively. Another use is a combination of deep learning and CNNs to extract features from sign language videos; Another work completed is a software prototype using OpenCV and ANN for sign language recognition using computer vision techniques. Also, defining 26 hand gestures using MATLAB also works towards solving the issues of ISL. Other associated technologies such as object recognition systems and image processing systems are also investigated to scan the hand signs of the sign language alphabet and convert them into voice and text to assist the hearing impaired.

The social implication of these technological advancements is great, and holds opportunities for social productivity, education, and employment for the Deaf. For instance, the real-time recognition application for ASL gestures employing CNN results in an ever astounding accuracy of 96. , thus helping to ensure more advanced knowledge among various groups of people. Combined, these different subjects play an important role in creating technology-focused solutions for improving communication for the special needs community, which in turn begins the march towards creating a more open and accepting society.

1.1 Aims & Objectives

The main objective of this research paper is to critically review and summarize the existing studies related to vision-based PSL interpreter and speech systems to assist the disabled people. In this way, we will be able to describe the subject area, determine the current state of knowledge, and

define the possible trends in the development of the spheres and disciplines. Thus, the aim is to find ways to build better and more efficient equipment and devices that are helpful in the process of enhancing the quality of the life for the Deaf community in Pakistan.

1. Review and Synthesis of Existing Research: The first outcome of this research work is to synthesize a detailed and extensive literature review on the available vision-based PSL and speech interpretation systems. This involves exploring the different techniques and strategies employed in these systems as well as the different techniques or algorithm types, and technologies in it with emphasis on deep learning. As such, the present work is designed to compile the findings of over 1000 papers of which around 50 are considered to be its main sources, offering an overview of the state of research in this field.

2. Identification of Challenges and Limitations: Another research aim is to define specific weaknesses and constraints relevant to the current systems of PSL interpretation. This involves such factors as data limitation, intricacy of signs, differences in signs depicting the same word, and the ability of such systems to operate under different conditions. It will also be helpful in the conceptualization of future studies and interventions that aim at addressing such difficulties.

3. Evaluation of Data Sources and Datasets: A comparative assessment of the datasets and data sources for PSL analysis will enable identifying the gap in the available data and stressing the necessity for more kinds and volumes of data. This objective focuses on principal objective entails examining current annotated PSL datasets, their quality and suitability for deep learning model training and evaluation.

4. Assessment of Deep Learning Techniques: Therefore, the paper will compare CNNs, RNNs, and transformers which are commonly used in PSL interpretation systems. To achieve this goal, we compared and analyzed the methods' performance in terms of the speed, precision, and computational complexity.

5. Proposal of Future Research Directions: For the future research, we will suggest research questions and probable solutions to the challenges highlighted in this study. This includes suggestions for creation of a better PSL interpretation system that is more reliable, accurate and scalable and ways of enhancing the data capture and labeling methodology.

6. Highlighting the Societal Impact: Last but not least, the paper will conclude with a discussion of practical implications of designing effective PSL and speech interpretation systems for the broader society. It relates to their ability to improve communication, access, education, employment, and social participation for persons with disability hence equal opportunities for all.

1.2 Scope of Review

The aim of this review is to critically discuss the collected scientific efforts in the realization of vision-based Pakistan Sign Language (PSL) and speech interpreted system for ensuring the efficient bridge between the verbally impaired people or hearing impaired people and mainstream society. In the following review, a wide range of works is presented ranging from a methodological approach and technological applications to education and other resources to improve the experience of the Deaf and mute societies. This paper also entails a discussion regarding the datasets and the recognition systems applicable to PSL and where researchers discover the language complexity and recurring cultural context associated with this form of sign language. Despite its focus on the BSL, the review goes further in a discussion of technological solutions employed in other sign languages including ASL, Arabic Sign Language as well as ISL as the research findings can be transferable in contexts consequent to differences in language and culture. Moreover, the focus also includes papers that extend to other fields, including image captioning, action recognition based on videos, and some real-time gesture recognition applications are also included in the scope, as this area of research is highly interdisciplinary. Overall, the review aspires to specifically and systematically summarize and discuss the state of the current research area, highlight both existing and potential research limitations and difficulties, and guide future practice and research on Developing integrated and universally available communication technologies for the Deaf, which will ultimately contribute to the anti- discriminant and sociable society.

2 RELATED WORK

The presented literature spans a diverse array of innovative approaches towards sign language recognition and translation, collectively addressing the critical need for enhanced accessibility and communication for the deaf and hard-of-hearing communities.

In the field of sign language recognition and translation across different languages, including Pakistan Sign Language, Urdu Sign Language, Arabic Sign Language, and Indian Sign Language, several papers, such as those by [1] and [2], focus on creating datasets and using machine learning for recognizing hand configurations of the Urdu alphabet. Similarly, [3] propose a computer vision-based system to aid differently-abled individuals by recognizing and classifying Urdu sign language using artificial intelligence. [4] and [5] contribute to the development of e-learning systems and translation tools for the hearing-impaired community in Pakistan and the Arabic-speaking world, respectively.

Advancements in deep learning techniques are evident

in works like DeepASL by [6], which offers non-intrusive word and sentence-level translation, and [7], who implement a vision-based deep learning approach for Arabic sign language. Other notable contributions include [8] and [9], focusing on word-level recognition using multi-stream neural networks and video data, respectively. [10] explore adaptive attention mechanisms for image captioning, potentially enhancing sign language interpretation systems. Moreover, research by [11] and [12] delve into transformer-based models and convolutional neural networks, advancing the accuracy and efficiency of sign language recognition systems.

3 DISCUSSION

[1] contribute significantly by introducing a dataset of Pakistan Sign Language (PSL) and an Android application for real-time sign recognition, achieving an accuracy range of 80-90%. [6] propose DeepASL, a versatile translation system utilizing deep learning, compatible with various sensors and devices. [2] employs image processing and machine learning for Pakistan sign language alphabet recognition, attaining an impressive 95% accuracy for hand configurations. [3] present a computer vision-based system for Urdu Sign Language (USL) recognition with an accuracy of 97.5%, emphasizing its potential for enhancing communication for differently-abled individuals in Pakistan. [13] introduce a vision-based deep learning approach for Arabic sign language interpretation, achieving an outstanding accuracy of 98.5%, signifying its potential to augment accessibility in Arabic-speaking countries. [10] revolutionize image captioning through an adaptive attention model with a visual sentinel, surpassing existing methods in generating captions with enhanced efficiency and accuracy. In [11], it is proposed that a Sign Pose-based Transformer for Word-level Sign Language Recognition, attaining state-of-the-art results and emphasizing computational efficiency. [14] pioneer a system utilizing image processing and natural language processing for recognizing sign language, providing a bridge for communication between deaf and non-deaf individuals.

[9] introduce a large-scale American Sign Language (ASL) dataset and a cutting-edge approach combining CNNs and GCNs for superior word-level sign recognition. [12] leverage a convolutional neural network (CNN) for sign language recognition, demonstrating high accuracy through advanced image preprocessing techniques. [8] revolutionize word-level sign language recognition by focusing on local regions, achieving a significant increase in accuracy through a multi-stream framework. [4] pioneer a web-based e-learning system for Pakistan Sign Language, empowering parents as educators and emphasizing the fundamental role of proper communication in learning. In [15], an intelligent translator for Arabic Sign Language (ArSL) has presented, leveraging hand tracking and facial expression recognition to classify dynamic gestures with enhanced accuracy and efficiency. [16] introduce a system for generating sign language based on Indian Sign Language (ISL) grammar, achieving a high BLEU score and demonstrating potential applications in diverse domains.

Recent advances in deep learning have significantly improved sign language recognition systems. [17] discuss a hybrid approach combining Long Short-Term Memory (LSTM) networks with MediaPipe holistic landmarks for dynamic signs and You Only Look Once (YOLO) for static signs. [18] present "Deepsign," which uses deep learning methods to enhance the detection and recognition of sign language. Their approach integrates convolutional neural networks (CNNs) and recurrent neural networks (RNNs), achieving high accuracy in recognizing various signs and demonstrating the potential of deep learning in facilitating communication for differently-abled individuals.

[19] focus on image-based recognition techniques for Pakistan Sign Language. They utilize convolutional neural networks (CNNs) to extract spatial features from images of sign language gestures, achieving high accuracy in recognizing static signs. This study highlights the potential of using CNNs for recognizing hand gestures and facial expressions in sign language. [5] present a deep learning approach for recognizing sign languages from video sequences. By combining CNNs with Long Short-Term Memory (LSTM) networks, they are able to capture both spatial and temporal features, significantly improving the recognition accuracy for continuous sign language gestures. [20] presented a system utilizing a capacitive touch sensor for recognizing hand gestures specifically designed for translating sign language. [21] investigate the barriers to healthcare access experienced by deaf individuals.

The paper by [22] discusses the application of convolutional neural networks (CNNs) for sign language recognition, emphasizing their effectiveness in accurately identifying hand gestures. [23] explore multi-modal sensor fusion for semantic segmentation in challenging snow driving scenarios, demonstrating how combining data from multiple sensors enhances the robustness and accuracy of the segmentation process. [24] provides a comprehensive review of machine learning techniques for sign language recognition, highlighting recent advancements and identifying future research directions. Together, these papers underscore the versatility and potential of machine learning and deep learning techniques in diverse applications, from enhancing accessibility for the hearing-impaired to improving safety in autonomous driving.

The paper by [25] presents a vision-based hand gesture recognition system using deep learning to interpret sign language, highlighting significant advancements in accuracy and real-time performance. [26] discuss an efficient

technique for human activity recognition leveraging deep learning, demonstrating improvements in recognizing complex activities from video data. [27] focus on face detection in video summaries, using an enhancement-based fusion strategy to improve detection accuracy, which is essential for effective video summarization. [28] explore the views, knowledge, and beliefs about genetics and genetic counseling among deaf people, emphasizing the unique perspectives and informational needs within the deaf community.

[29] provide a comprehensive review of sign language recognition systems over the past decade, highlighting the evolution of techniques and technologies used to improve the accuracy and efficiency of these systems. Their systematic literature review covers various methodologies, including machine learning and deep learning approaches, and discusses the challenges and future directions in the field. [30] propose a novel method for hand extraction in Arabic Sign Language Recognition (ArSLR) systems, which enhances the accuracy of recognizing hand gestures crucial for interpreting Arabic sign language. In a follow-up study, [31] present the design of an automatic translation system from Arabic sign language to Arabic text, showcasing the potential for technology to bridge communication gaps for the Arabic-speaking deaf community.

[32] focus on the medical field, specifically the use of dermoscopy with and without visual inspection for diagnosing melanoma in adults. Their systematic review in the Cochrane Database of Systematic Reviews evaluates the diagnostic accuracy of dermoscopy, underscoring its importance in early melanoma detection and treatment planning. [33] developed a real-time system for translating Arabic sign language into Arabic text and sound, providing a valuable tool for improving communication for the hearing-impaired Arabic-speaking community. [34] evaluated various texture features for content-based image retrieval, contributing to advancements in image processing and retrieval systems by identifying the most effective features for accurate image categorization. [35] designed an automatic Arabic Sign Language Recognition System (ArSLRS), which enhances the accuracy and efficiency of sign language interpretation through the use of advanced algorithms.

[36] investigate key frame extraction for continuous Indian sign language gesture recognition and sentence formation in Kannada, comparing the effectiveness of various classifiers in improving recognition accuracy. [37] explore human action recognition using a transfer learning approach, demonstrating enhanced performance by leveraging pre-trained models. [38] discusses the effectiveness of deep learning using Rectified Linear Units (ReLU), highlighting their impact on improving model convergence and performance. Al- [39]present a 3D Convolutional Neural Network (3DCNN) for hand gesture recognition in sign language, showing significant improvements in accurately interpreting complex gestures.

[40] provide a systematic literature review on multi-objective feature selection approaches, highlighting various techniques and their applications in optimizing feature subsets for improved model performance. [41] offer a foundational understanding of convolutional neural networks (CNNs), detailing their architecture and the principles behind their effectiveness in image recognition tasks. [42] propose an optimized transfer learning-based approach for the automatic diagnosis of COVID-19 from chest X-ray images, showcasing significant improvements in diagnostic accuracy. [43] conduct a comprehensive review on the automatic recognition of handwritten Arabic characters, summarizing various methodologies and advancements in this field.

[44] present a hybrid framework for COVID-19 segmentation and recognition (HMB-HCF) using deep learning and genetic algorithms, demonstrating enhanced diagnostic accuracy. In a subsequent study, [7] develop a comprehensive framework for the accurate recognition and prognosis of COVID-19 patients, leveraging deep transfer learning and feature classification to improve patient outcomes. Additionally, [45] propose a hybrid deep learning and genetic algorithms approach (HMB-DLGAHA) for early ultrasound diagnoses of breast cancer, highlighting its effectiveness in early detection.

[46] developed a real-time system for translating sign language into text and speech using convolutional neural networks (CNNs), significantly enhancing communication for the hearing-impaired. [47] similarly created a sign language translator utilizing machine learning techniques to improve the accuracy and usability of sign language recognition systems. [48] focused on real-time American Sign Language (ASL) recognition using CNNs, demonstrating the effectiveness of deep learning models in accurately interpreting ASL gestures.

[49] developed an interpreter for Indian Sign Language aimed at assisting deaf and mute individuals, utilizing image processing and machine learning techniques to translate gestures into text and speech. [50] focused on interpreting Swedish Sign Language using convolutional neural networks (CNNs) and transfer learning, demonstrating improved accuracy and efficiency in gesture recognition. [51] presented a sign language interpreter that employs image processing and machine learning to convert sign language gestures into readable text, enhancing communication for the hearing-impaired.

3.1 Dataset

Diverse approaches have been employed in the creation of datasets, each tailored to specific linguistic and cultural

contexts.

The creation process for the Urdu alphabet dataset involved capturing 40 images of each letter, resulting in a dataset of static hand configurations in Pakistan Sign Language with 37 classes. This dataset was generated by recording a native signer's hand configurations using a high-resolution camera.

The evaluation of the adaptive attention model for image captioning used two datasets: the COCO image captioning 2015 challenge dataset with over 330,000 images and the Flickr30K dataset with 31,783 images. These datasets, widely employed in image captioning research, were divided into training, validation, and test sets. For word-level sign recognition, a new large-scale ASL dataset comprising over 2000 words performed by 100 signers was introduced. It includes diverse backgrounds, illumination conditions, and signer appearances, with 1750 static images for each sign used in training.

The Arabic Sign Language (ArSL) dataset consists of 1000 samples of gestures, collected, labeled, preprocessed, and augmented for improved model accuracy. It is divided into training and testing sets, considering variations in speed, movement, orientation, background, and the age of gesture operators.

The Indian Sign Language (ISL) corpus, based on HamNoSys notation, comprises 2,950 commonly used English words categorized into various domains. This dataset was utilized to develop a sign language generation system, including a parser for processing input sentences and generating ISL animations, with the document discussing the system's results and evaluation.

Dataset Name	Description	Size	Usage
Urdu Alphabet	Captured 40 images of each letter with various hand configurations.	37 classes	Pakistan Sign Language recognition
COCO Image Captioning 2015 Challenge	Over 330,000 images with 5 captions each, used for image captioning research.	Divided into sets	Evaluation of adaptive attention model
Flickr30K	31,783 images with 5 captions each, commonly used for image captioning research.	Divided into sets	Evaluation of adaptive attention model
Word-level Sign Recognition (ASL)	Over 2000 words performed by over 100 signers, the largest public ASL dataset for research.	1750 images per sign	Word-level sign recognition research
Arabic Sign Language (ArSL)	1000 samples of gestures with variations in speed, movement, orientation, background, age.	Divided into training and testing sets	Gesture recognition in Arabic Sign Language
Indian Sign Language (ISL) Corpus	2,950 commonly used English words categorized into various domains.	-	Sign language generation system for ISL

Figure 2. Details of datasets used in research.

The data preprocessing typically involves background subtraction to isolate the hand gestures from the rest of the scene. Noise reduction techniques, such as Gaussian blurring and median filtering, are applied to smoothen the images. Keypoint detection algorithms are used to identify and track critical points on the hands, such as fingertips and joints. Additionally, normalization methods are used to standardize the input data, ensuring consistency across the dataset. Some studies also employ segmentation algorithms to delineate hand regions more precisely, facilitating more accurate feature extraction and subsequent classification.

4 DATASET FOR VISIOHEAR

4.1 Dataset Collection and Preparation

For our project, we've assembled an exceptional collection of **"110"** distinct words or classes in Pakistan Sign Language.

Types of Classes	No. of Classes
Common Nouns	40
Prepositions	4
Adverbs	9
Verbs	25
Pronouns	6
Question Words	6
Adjectives	9
Seasons	2
Time-related Words	7
Interjections	2

Figure 3. Details of selected classes/words

These selected words play a crucial role in facilitating basic communication within Pakistan Sign Language. While there may be additional words that serve this purpose, we have deliberately chosen to focus our scope on the aforementioned set. Each word has meticulously captured in approx. **"55"** images, ensuring we encompass all the various angles and nuances of its signing.

While a relatively smaller dataset, with 110 words, this set of 55 images per word lays a solid foundation. It's important to note that this number represents just a starting point. Through the application of data augmentation techniques, we'll dynamically generate even more diverse samples, further enhancing the robustness of our dataset. This technique allows us to create variations in the existing images, effectively expanding our dataset beyond its initial size. Altogether, this comprehensive approach resulted in a substantial dataset, totaling an impressive **"6034"** images. After collection of dataset, we labelled our dataset using **Roboflow**, which provides the ease of downloading labelled dataset in various formats, and also divide it into train, test and valid sections. Out of **6034 images**, we have **4268 images** in **training** set, **1214 images** in **validation** set and **552 images** in **testing** set.

4.2 Comparative Dataset Inclusion

To ensure diversity of our dataset at initial stage, we collected from **"12-14" signers** including both male and female, so in this way, our model would be able to generalize signing styles are contexts. This dataset not only provided a robust basis for training our model but also ensured its ability to recognize a wide range of sign language gestures, even in varying contexts and conditions.

To overcome the unavailability of word based dataset of Pakistan Sign Language, we researched that which other sign language is comparable to the PSL, and it eventually came out that "Indian Sign Language (ISL)" is one of the most resembling sign languages, due to some common cultural factors, and language. Approximately, 40% words that we are including in our dataset, have same signs in both PSL and ISL. We have also taken a few images for our dataset from [52] and [53].

5 ALGORITHMS/MODELS USED

To achieve effective sign language recognition, we employed a variety of machine learning and deep learning models:

1. LSTM (Long Short-Term Memory Networks): We used LSTM networks to handle sequences of keypoints extracted from hand gestures. LSTM's ability to capture temporal dependencies made it suitable for recognizing dynamic signing patterns. The model was implemented using TensorFlow and Keras.

2. SSD MobileNet: We merged the Single Shot MultiBox Detector (SSD) with MobileNet for efficient object detection. This combination was chosen for its balance between speed and accuracy, leveraging TensorFlow's Object Detection API and extensive training epochs.

3. YoloV5 (You Only Look Once Version 5): Known for its rapid detection capabilities and high accuracy, YoloV5 was employed for real-time sign language detection. This model's architecture facilitates fast processing and precise localization of hand gestures.

4. Faster RCNN (Region Convolutional Neural Network): We implemented a pre-trained Faster RCNN model with a MobileNet backbone using PyTorch. This model was chosen for its robust object detection capabilities and accuracy in identifying complex gestures.

5. EfficientDet-d0: Utilizing TensorFlow's Object Detection API, EfficientDet-d0 was trained on our custom dataset formatted in COCO. This model was selected for its efficient and scalable detection performance, particularly useful for mobile and embedded applications.

6 METHODOLOGY

6.1 Long Short-Term Memory (LSTM):

The LSTM model was implemented using TensorFlow and Keras. We utilized a pretrained LSTM model, which was fine-tuned on our dataset. The architecture of the LSTM network included multiple layers to effectively capture the temporal dependencies in the sequences of keypoints. The model was trained using the training set, with hyperparameter tuning and cross-validation conducted to optimize its performance.

To train our LSTM model, we assembled a dataset of 10 distinct Urdu words. For each word, we recorded 30 videos, each containing 30 frames. These frames captured various angles and nuances of the signing gestures. The keypoints of the hand gestures were extracted from each frame to create a sequence of keypoints representing each video. This data was then divided into training, validation, and testing sets to ensure robust model evaluation.

6.2 SSD MobileNet:

SSD MobileNet for Object Detection We integrated the Single Shot MultiBox Detector (SSD) with MobileNet to achieve efficient object detection. This combination was selected due to its optimal balance between speed and accuracy, leveraging TensorFlow's Object Detection API and extensive training periods. For this project, we utilized a dataset of 110 Urdu words, with each word represented by 55 images labeled in Pascal VOC format. We applied a pretrained MobileNet SSD model from TensorFlow's Object Detection API to our dataset.

The SSD MobileNet model was implemented using TensorFlow's Object Detection API. This model synergizes the SSD architecture, which excels at efficient object detection, with the MobileNet backbone, renowned for its lightweight and efficient structure. We fine-tuned the pretrained model on our dataset over **250,000 training epochs**. This extensive training regimen enabled the model to effectively learn and recognize the nuances of Urdu sign language gestures.

6.3 You Only Look Once version 5 (YoloV5):

For the real-time sign language detection task, we employed the YOLOv5 model, renowned for its rapid detection capabilities and high accuracy. YOLOv5's architecture is specifically designed for fast processing and precise localization of objects, making it an ideal choice for recognizing hand gestures in sign language.

Our dataset comprised 110 Urdu words, with 55 images representing each word. The images were labeled using the PASCAL VOC format, which provides a standardized way to annotate object detection datasets with bounding boxes and class labels.

To develop our custom sign language detection model, we cloned a pre-trained YOLOv5 model from a repository on GitHub. This pre-trained model served as the base, allowing us to leverage transfer learning to improve training

efficiency and performance. We utilized TensorFlow and PyTorch libraries to train the model on our custom dataset. The training process was conducted over **500 epochs**, a number chosen to ensure the model could adequately learn from the data while balancing the risk of overfitting.

We utilized Google Colab as our platform for training the model. Google Colab provides a powerful and accessible environment with free GPU resources, which significantly accelerates the training process. After training, the model weights were generated and saved in the ONNX format.

6.4 Faster R-CNN:

For our object detection task, we implemented a pre-trained Faster R-CNN model with a MobileNet backbone, leveraging the capabilities of **PyTorch** for training and model management. Faster R-CNN is a popular and robust object detection model known for its accuracy, and MobileNet provides a lightweight backbone suitable for real-time applications without compromising performance.

Our dataset was labeled in the **COCO (Common Objects in Context)** format, a widely used format for object detection datasets that provides detailed annotations including bounding boxes, object categories, and more. This standardized format ensured compatibility and facilitated seamless training of the model.

To train the model, we used a dataset with extensive annotations and conducted the training process over 400 epochs. The choice of **400 epochs** was aimed at ensuring the model had ample opportunity to learn the features of the dataset thoroughly, striking a balance between comprehensive learning and avoiding overfitting.

Google Colab was chosen as the training platform due to its accessibility and provision of free GPU resources, which significantly enhanced the training speed and efficiency. Utilizing Google Colab allowed us to leverage powerful computational resources without the need for extensive local hardware.

Upon completion of the training process, the model weights were generated and saved in the ONNX (Open Neural Network Exchange) format.

6.5 EfficientDet-d0:

We utilized TensorFlow's Object Detection API to train an EfficientDet-d0 model on our custom dataset. EfficientDet-d0 was selected due to its reputation for efficient and scalable detection performance, which is particularly advantageous for deployment in mobile and embedded applications where computational resources are often limited.

Our dataset was formatted in the COCO (Common Objects in Context) format, which is a widely adopted standard for object detection tasks. The COCO format provides comprehensive annotations including bounding boxes, object categories, and segmentation masks, enabling detailed and accurate model training.

We employed TensorFlow's Object Detection API for training the model. This API simplifies the process of building, training, and deploying object detection models by providing pre-configured pipelines and a wide range of tools for data preprocessing, model evaluation, and fine-tuning. The training process was executed over a substantial number of epochs to ensure the model thoroughly learned from the dataset, optimizing its detection capabilities.

The choice of TensorFlow's Object Detection API, combined with the efficient architecture of EfficientDet-d0, facilitated the development of a robust and scalable object detection model.

7 RESULTS & DISCUSSION

7.1 Results

The performance of various models for sign language detection and recognition was evaluated based on their accuracy and efficiency. The results obtained from each model are summarized as follows:

Long Short-Term Memory (LSTM):

- Achieved the highest accuracy in detecting and converting sign language gestures into text.
- Consistently identified and translated the 10 distinct Urdu words accurately across various video samples.

SSD MobileNet:

- Produced the poorest results despite extensive training over 250,000 epochs.
- Accurately identified only 2 out of the 110 Urdu words in the dataset.

EfficientDet-d0:

- Showed moderate performance, successfully detecting some words but not all.
- Several words were either misclassified or not detected, affecting overall reliability.

Faster R-CNN:

- Notably slow to train, requiring substantial computational resources and time.
- After 400 epochs, managed to detect only 2 words accurately.
- Inefficient for real-time applications due to slow training process and limited detection capability.

YOLOv5:

- Demonstrated good detection speed, suitable for real-time applications.
- Accuracy was not satisfactory, with frequent incorrect detections of signs.

7.2 Discussion

The **LSTM** model emerged as the most effective solution for sign language detection and recognition. Its ability to capture temporal dependencies in sequences allowed for high accuracy in detecting and converting sign language gestures into text. The model was consistently reliable across various samples, making it the optimal choice for integration into our **"BeyondTheWords"** platform. The robust performance of the LSTM model highlights the importance of temporal modeling in understanding sign language gestures.

In contrast, the **SSD MobileNet** model performed poorly despite extensive training. Its failure to detect more than 2 words accurately underscores the limitations of this architecture for complex gesture recognition tasks. The model's inability to handle the intricacies of sign language gestures indicates that a balance between speed and accuracy is crucial, but not at the expense of detection capability.

EfficientDet-d0 showed promise with its efficient and scalable detection capabilities, but its inconsistent accuracy prevented it from being the top choice. The model's partial success in word detection suggests that further optimization and fine-tuning might be necessary to improve its performance for sign language recognition.

Faster R-CNN, while known for its accuracy in object detection, was hampered by its slow training process and limited success in detecting Urdu words. The extensive time and resources required for training, combined with its inability to detect more than 2 words, highlight the challenges of using this model for real-time applications. Its inefficiency in our specific context emphasizes the need for models that balance accuracy and computational efficiency.

YOLOv5, with its rapid detection capabilities, seemed promising for real-time applications. However, the trade-off in accuracy, with frequent incorrect detections, made it less suitable for our sign language detection task. The need for precise localization and accurate detection of gestures is paramount, and YOLOv5's performance did not meet these requirements satisfactorily.

8 CONCLUSION

The VisioHear project represents a significant step forward in the realm of inclusive communication technologies, particularly for the deaf and mute community fluent in Pakistan Sign Language (PSL). By leveraging cutting-edge machine learning models and a comprehensive dataset, we have developed a system capable of accurately detecting and translating sign language gestures into text and audio in real-time.

Through our experimentation with various models, we have identified the Long Short-Term Memory (LSTM) model as the most effective solution for sign language recognition, demonstrating superior accuracy and reliability compared to other architectures. This finding underscores the importance of temporal modeling in understanding and interpreting sign language gestures accurately.

While our project has achieved promising results, there are still areas for improvement and further exploration. Future work could focus on refining the existing models to enhance accuracy and efficiency, particularly in real-world scenarios with diverse lighting conditions and signing styles. Additionally, expanding the dataset to include a wider range of sign language gestures and dialects would improve the system's robustness and adaptability.

9 FUTURE WORK

9.1 Model Optimization

We plan to further optimize the LSTM model to improve its performance in real-world settings, including noisy environments and varying signing styles. This optimization may involve fine-tuning hyperparameters, exploring novel architectures, and implementing advanced techniques for temporal modeling.

9.2 Dataset Expansion

Expanding the dataset to include a more extensive range of sign language gestures and dialects, beyond the overlap between PSL and Indian Sign Language (ISL), will enhance the system's versatility and adaptability. This expansion may involve collecting data from additional signers and incorporating gestures from diverse cultural contexts.

9.3 Real-World Deployment

We aim to deploy the VisioHear system in real-world settings, such as educational institutions, community centers, and workplaces, to evaluate its effectiveness and usability in practical scenarios. This deployment will involve collaboration with stakeholders to gather feedback and iterate on the system's design and functionality.

9.4 Accessibility Features

In addition to sign language translation, we plan to incorporate accessibility features into the VisioHear platform, such as voice recognition for non-signing users and text-to-speech functionality for deaf and mute individuals. These features will further enhance communication and accessibility for users with diverse needs.

9.5 Integration with Assistive Technologies

We intend to explore opportunities for integrating VisioHear with existing assistive technologies, such as augmented reality (AR) glasses and wearable devices, to provide seamless and immersive communication experiences. This integration will enable users to access the VisioHear system hands-free and in various environments.

REFERENCES

[1] Imran, A., Razzaq, A., Baig, I. A., Ali, M. & Khan, M. A. Dataset of pakistan sign language and automatic recognition of hand configuration of urdu alphabet through machine learning. *Data Brief* **36**, 107021 (2021). URL https://www.sciencedirect.com/science/article/pii/S235234092100305X.

[2] Khan, N. S. A vision-based approach for pakistan sign language alphabets recognition. *Pensee J.* **76**, 274 (2014). URL https://doi.org/10.1038/s41598-022-15864-6.

[3] Zahid, H. *et al.* A computer vision-based system for recognition and classification of urdu sign language dataset for differently abled people using artificial intelligence. *Mob. Inf. Syst.* 1–14 (2021). URL https://doi.org/10.7717/peerj-cs.1174.

[4] Dewani, A. *et al.* Sign language e-learning system for hearing-impaired community of pakistan. *Int. J. Inf. Technol.* **10**, 225–232 (2018). URL https://doi.org/10.1007/s41870-018-0105-4.

[5] Shahid, A. S., A. & Shah, J. Automatic recognition of sign languages using deep learning. *IEEE Xplore* (2019). URL https://ieeexplore.ieee.org/document/8785559.

[6] Fang, Y., Xu, C., Zhang, Y. & Stankovic, J. 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)*, 1–14 (ACM, 2017). URL https://doi.org/10.1145/3131672.3131693.

[7] Balaha, H. M., El-Gendy, E. M. & Saafan, M. M. A complete framework for accurate recognition and prognosis of covid-19 patients based on deep transfer learning and feature classification approach. *Artif Intell Rev* 1–46 (2022). URL https://doi.org/10.1007/s10462-021-10127-8.

[8] Maruyama, M. *et al.* Word-level sign language recognition with multi-stream neural networks focusing on local regions. *arXiv preprint arXiv:2102.07167* (2021). URL https://api.semanticscholar.org/CorpusID:235683225.

[9] Li, Y., Zhang, J. & Nevatia, R. Word-level deep sign language recognition from video. *arXiv preprint arXiv:2011.09206* (2020). URL https://api.semanticscholar.org/CorpusID:204851909.

[10] Lu, J., Xiong, C., Parikh, D. & Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning (2017). URL https://doi.org/10.48550/arXiv.1612.01887.

[11] Boháček, M. & Hrúz, M. Sign pose-based transformer for word-level sign language recognition. *arXiv preprint arXiv:2105.05097* (2021). URL https://ieeexplore.ieee.org/document/9707552.

[12] Nandhini, A. S. *et al.* Sign language recognition using convolutional neural network. *J. Physics: Conf. Ser.* **1916**, 012091 (2021). URL https://iopscience.iop.org/article/10.1088/1742-6596/1916/1/012091.

[13] Balaha, M. M. *et al.* A vision-based deep learning approach for independent-users arabic sign language interpretation. *Multimed. Tools Appl.* (2022). URL https://doi.org/10.1007/s11042-022-13423-9.

[14] Wazalwar, S. S. & Shrawankar, U. Sign language recognition using image processing and natural language processing techniques. *Int. J. Comput. Appl.* **177** (2017). URL https://api.semanticscholar.org/CorpusID:67145491.

[15] Ahmed, A. M. *et al.* Arabic sign language translator. *J. Comput. Sci.* **15**, 1522–1537 (2019). URL https://doi.org/10.3844/jcssp.2019.1522.1537.

[16] Sugandhi, P., Kumar, P. & Kaur, S. Sign language generation system based on indian sign language grammar. *ACM Transactions on Asian Lang. Inf. Process.* **19**, Article 54 (2020). URL https://dl.acm.org/doi/10.1145/3384202.

[17] Buttar, A. M. *et al.* Deep learning in sign language recognition: A hybrid approach for the recognition of static and dynamic signs. *Math.* **11**, 3729 (2023). URL https://doi.org/10.3390/math11173729.

[18] Kothadiya, D. *et al.* Deepsign: Sign language detection and recognition using deep learning. *Electron.* **11**, 1780 (2022). URL https://doi.org/10.3390/electronics11111780.

[19] Raees, M., Ullah, S., ur Rahman, S. & Rabbi, I. Image based recognition of pakistan sign language. *J. Eng. Res.* **4**, 1–21 (2016). URL https://api.semanticscholar.org/CorpusID:55465527.

[20] Abhishek, K. S., Qubeley, L. C. F. & Ho, D. Glove-based hand gesture recognition sign language translator using capacitive touch sensor. In *Electron Devices and Solid-State Circuits (EDSSC), 2016 IEEE International Conference on*, 334–337 (IEEE, 2016). URL https://ieeexplore.ieee.org/document/7785276.

[21] Kuenburg, A., Fellinger, P. & Fellinger, J. Health care access among deaf people. *J. deaf studies*

deaf education **21** **1**, 1–10 (2016). URL https://api.semanticscholar.org/CorpusID:204991690.

[22] Kumar, A., Kumar, S., Singh, S. & Jha, V. Sign language recognition using convolutional neural network. In *ICT Analysis and Applications*, 915–922 (Springer, 2022). URL https://doi.org/10.1007/978-981-16-5655-2_87.

[23] Vachmanus, S., Ravankar, A. A., Emaru, T. & Kobayashi, Y. Multi-modal sensor fusion-based semantic segmentation for snow driving scenarios. *IEEE Sensors J.* **21**, 16839–16851 (2021). URL https://api.semanticscholar.org/CorpusID:235842063.

[24] Elakkiya, R. Machine learning based sign language recognition: a review and its research frontier. *J. Ambient Intell. Humaniz. Comput.* **12**, 7205–7224 (2021). URL https://doi.org/10.1007/s12652-020-02396-y.

[25] Sharma, S. & Singh, S. Vision-based hand gesture recognition using deep learning for the interpretation of sign language. *Expert. Syst. with Appl.* **182**, 115657 (2021). URL https://doi.org/10.1016/j.eswa.2021.115657.

[26] Khelalef, A., Ababsa, F. & Benoudjit, N. An efficient human activity recognition technique based on deep learning. *Pattern Recognit. Image Analysis* **29**, 702–715 (2019). URL https://doi.org/10.1134/S1054661819040084.

[27] Mishra, R. & Subban, R. Face detection for video summary using enhancement based fusion strategy. *Int. J. Renew. Energy Technol.* **3**, 69–74 (2014). URL https://ieeexplore.ieee.org/document/7043634.

[28] Middleton, A., Emery, S. D. & Turner, G. H. Views, knowledge, and beliefs about genetics and genetic counseling among deaf people. *Sign Lang. Stud.* **10**, 170–196 (2010). URL https://doi.org/10.1353/sls.0.0038.

[29] Wadhawan, A. & Kumar, P. Sign language recognition systems: A decade systematic literature review. *Arch. Comput. Methods Eng.* **28**, 785 – 813 (2019). URL https://api.semanticscholar.org/CorpusID:213429407.

[30] Ahmed, A. M., Alez, R. A., Taha, M. & Tharwat, G. Propose a new method for extracting hand using in the arabic sign language recognition (arslr) system. *Int. journal engineering research technology* **4** (2015). URL https://api.semanticscholar.org/CorpusID:55247326.

[31] Ahmed, A. M. *et al.* Towards the design of automatic translation system from arabic sign language to arabic text. *2017 Int. Conf. on Inventive Comput. Informatics (ICICI)* 325–330 (2017). URL https://api.semanticscholar.org/CorpusID:44153913.

[32] Dinnes, J., Deeks, J. J., Chuchu, N., Ferrante, D. R. L. & Matin, R. N. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. *Cochrane Database Syst. Rev.* **12**, CD011902 (2018). URL https://doi.org/10.1002/14651858.CD011902.pub2.

[33] ElAlfi, A. E. E., El-Gamal, A. F. & El-Adly, R. A. Real time arabic sign language to arabic text & sound translation system. *Int. journal engineering research technology* **3** (2014). URL https://api.semanticscholar.org/CorpusID:124474718.

[34] Howarth, P. & Rüger, S. M. Evaluation of texture features for content-based image retrieval. In *ACM International Conference on Image and Video Retrieval* (2004). URL https://api.semanticscholar.org/CorpusID:14196440.

[35] Ibrahim, N. B., Selim, M. M. & Zayed, H. H. An automatic arabic sign language recognition system (arslrs). *J. King Saud Univ. - Comput. Inf. Sci.* **30**, 470–477 (2018). URL https://doi.org/10.1016/j.jksuci.2017.09.007.

[36] Kagalkar, R. M. & Gumaste, S. Gradient based key frame extraction for continuous indian sign language gesture recognition and sentence formation in kannada language: A comparative study of classifiers. *Int. J. Comput. Sci. Eng.* **4**, 1–11 (2016). URL https://www.ijcseonline.org/pdf_paper_view.php?paper_id=1047&1-IJCSE-01798.pdf.

[37] Abdulazeem, Y., Balaha, H. M., Bahgat, W. M. & Badawy, M. Human action recognition based on transfer learning approach. *IEEE Access* **9**, 82058–82069 (2021). URL https://doi.org/10.1109/ACCESS.2021.3086668.

[38] Agarap, A. F. Deep learning using rectified linear units (relu). *ArXiv* **abs/1803.08375** (2018). URL https://api.semanticscholar.org/CorpusID:4090379.

[39] Al-hammadi, M. *et al.* Hand gesture recognition for sign language using 3dcnn. *IEEE Access* **8**, 79491–79509 (2020). URL https://api.semanticscholar.org/CorpusID:218597379.

[40] Al-Tashi, Q., Abdulkadir, S. J., Rais, H. M., Mirjalili, S. M. & Alhussian, H. S. A. Approaches to multi-objective feature selection: A systematic literature review. *IEEE Access* **8**, 125076–125096 (2020).

URL https://api.semanticscholar.org/CorpusID:220668135.

[41] Albawi, S., Mohammed, T. A. & Al-Zawi, S. Understanding of a convolutional neural network. *2017 Int. Conf. on Eng. Technol. (ICET)* 1–6 (2017). URL https://api.semanticscholar.org/CorpusID:3819513.

[42] Bahgat, W. M., Balaha, H. M., AbdulAzeem, Y. & Badawy, M. M. An optimized transfer learning-based approach for automatic diagnosis of covid-19 from chest x-ray images. *PeerJ Comput. Sci* **7**, e555 (2021). URL https://doi.org/10.7717/peerj-cs.555.

[43] Balaha, H. M., Ali, H. A. & Badawy, M. Automatic recognition of handwritten arabic characters: a comprehensive review. *Neural Comput. Appl.* **33**, 3011 – 3034 (2020). URL https://api.semanticscholar.org/CorpusID:220612934.

[44] Balaha, H. M., Ali, H. A., Saraya, M. & Badawy, M. Hybrid covid-19 segmentation and recognition framework (hmb-hcf) using deep learning and genetic algorithms. *Artif Intell Med* **119**, 102156 (2021). URL https://doi.org/10.1016/j.artmed.2021.102156.

[45] Balaha, H. M., Saif, M., Tamer, A. & Abdelhay, E. H. Hybrid deep learning and genetic algorithms approach (hmb-dlgaha) for the early ultrasound diagnoses of breast cancer. *Neural Comput. Applic* 1–25 (2022). URL https://doi.org/10.1007/s00521-021-06851-5.

[46] Ojha, A., Pandey, A., Maurya, S., Thakur, A. & Dayananda, P. Sign language to text and speech translation in real time using convolutional neural network. *Int. journal engineering research technology* **8** (2020). URL https://api.semanticscholar.org/CorpusID:222284650.

[47] Gowda, H., Chandra, V., Vishwas, S. *et al.* Sign language translator using machine learning. *Int. J. Appl. Eng. Res.* **13** (2018). URL https://www.ripublication.com/ijaerspl2018/ijaerv13n4spl_01.pdf.

[48] Gracia, B. & Viesca, S. A. Real-time american sign language recognition with convolutional neural networks. (2016). URL http://cs231n.stanford.edu/reports/2016/pdfs/214_Report.pdf.

[49] Chavan, A., Bhat, A., Mishra, S. *et al.* Indian sign language interpreter for deaf and mute people. *IJCRT* **9** (2021). URL https://www.ijcrt.org/papers/IJCRT2103178.pdf.

[50] Halvardsson, G., Peterson, J., Soto-Valero, C. & Baudry, B. Interpretation of swedish sign language using convolutional neural networks and transfer learning. *SN Comput. Sci.* **2** (2020). URL https://api.semanticscholar.org/CorpusID:222377868.

[51] Vedak, O., Zavre, P., Todkar, A. *et al.* Sign language interpreter using image processing and machine learning. *IRJET* **06** (2019). URL https://www.irjet.net/archives/V6/i4/IRJET-V6I4413.pdf.

[52] Ayushirajput1. Indian sign language (2021). URL https://www.kaggle.com/datasets/ayushirajput1/indian-sign-language.

[53] Chouhan, K. Indian sign language animated videos (2021). URL https://www.kaggle.com/datasets/koushikchouhan/indian-sign-language-animated-videos.