Real Time Translation of Sign Language Sentences to Bangla Text Using Deep Learning Models and Holistic Keypoints

1st Sakif Hussain Shachcha
Dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
190204033@aust.edu

2nd Ahnaf Samin
dept. of CSE
Ahsanullah University of Science and Tehcnology
Dhaka, Bangladesh
190204031@aust.edu

3rd Anik Paul
dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
190204047@aust.edu

4th Nowratun Oyshe
Dept. of CSE
Ahsanullah University of Science and Technology
Dhaka, Bangladesh
190204056@aust.edu

***Abstract*—This research explores the advancements in machine learning, particularly in the realm of sign language translation, focusing on Bangla, the national language of Bangladesh.. Traditionally reliant on sign language interpreters, this study delves into the potential of automated, real-time translation systems facilitated by deep learning. The methodology involves assembling a comprehensive dataset of Bangla sign language gestures, captured and annotated to create a robust training dataset. Different machine learning models are then trained and evaluated based on performance metrics, such as accuracy, speed, and robustness to real-world variations. The research delves into eight relevant papers, drawing insights from various models and datasets. Notably, it discusses the challenges and successes of using different models on datasets like WLASL, BdSL, and custom datasets like BD-Word. The proposed methodology integrates human skeleton detection using Mediapipe, focusing on hand gestures and lip reading expressions. This framework extracts keypoints for skeletal tracking, capturing intricate nuances of Bangla sign language. The detected gestures and lip movements are aligned with natural language meanings, forming the basis for training a specialized LSTM model. The resulting comprehensive model for Bangla sign language recognition captures both hand gestures and lip reading nuances, contributing to improved communication accessibility for individuals with hearing or verbal impairments.**

I. Introduction

Imagine a world where everyone can communicate effortlessly, regardless of hearing or speech abilities. Sadly, this isn't the case for everyone. Many individuals with these challenges rely on sign language, often with the help of interpreters. But what if technology could step in and make communication smoother? That's where our research comes in.

We're motivated to use the power of deep learning, a fancy term for smart computer systems, to create a system that can instantly translate sign language into spoken words. This could be a game-changer for people facing communication barriers. We want to make sure that everyone, regardless of their abilities, can communicate easily and feel included in everyday conversations. That's the driving force behind our project.

In recent years there has been dramatic progress in the field of machine learning. One such field is the field of language, specifically sign language. People with hearing and verbal disabilities use this language to communicate. This thesis aims to conduct a comparative analysis of machine learning models employed for translating sign language into Bangla text, the national language of Bangladesh. The primary focus of this thesis will be to explore the different machinelearning models for translating sign language into Bangla text. We will also try to evaluate the state-of-the-art models and analyze them to improve their accuracy. Our goal will be to identify the most effective model for translating sign gestures into meaningful Bangla text. Traditionally we have relied upon sign language interpreters to facilitate communication between the wider society and individuals with hearing or verbal impairment. With the help of machine learning, we can develop real time automated translation systems that can enable more individuals with hearing or verbal impairment to communicate more freely and effectively without needing the help from a third party. The research methodology employed in this thesis will involve gathering a comprehensive dataset of sign language gestures commonly used in Bangladesh. These gestures will be captured and annotated to create a robust training dataset. A comparative analysis will be conducted by training and evaluating different machine learning models using this dataset. The models will be assessed based on performance metrics such as accuracy, speed, and robustness to real-world variations. The findings will be statistically analyzed, and recommendations for the most effective model will be proposed.

II. Literature Review

We have reviewed a total of 8 research papers and Donxu Li's paper [4] dived deep into the whole idea of translating sign language and the paper gave us insights into the models that can be used and the type of dataset. They used the WLASL dataset containing more than 2000 datasets in video format and ran 3D-CNN on the dataset having achieved word level accuracy of 67.83% but it achieved 87.99% on the same dataset with less data. This hints us that the accuracy of 3D-CNN might fall with increasing datasets. They also tried using cross-domain knowledge transfer techniques [3] with RCNN with 2D CNN such as LSTM and GRU with was able to achieve an accuracy of 68% and 89.92% respectively. Then we can see in Nasima Begum's paper [2] which worked with the Bangla hand sign language dataset known as the BdSL dataset. They used the YOLOV4 model and achieved a stunning word level accuracy of 99.7% but it was not in real time and it was mainly done with images and not sequential frames from videos. Finally, in Akash's paper [1], we can see the implementation of real-time Bangla hand sign language translation. They have used YOLOV4 and RCNN to detect the letters. In this approach, the Blazepose algorithm comprehensively extracts body key points. No convolutional network was used in this method as it would increase the complexity. As a result, their strategy was to concentrate on the sequence of sign actions. LSTM is used to extract a model from the action sequence. They have achieved 93% accuracy with the BD-Word dataset which is a custom dataset having 300 videos with 30 FPS each. But in terms of real time capture,the accuracy fell to 73%.

III. Dataset

We are planning to take videos from different sources also create some of our own and merge them with the already existing BD-word dataset to create a larger dataset with more words.

IV. Background Study

A. Convolutional Neural Network for Detection

CNN are a class of neural network that are highly useful in solving computer vision problems. They found inspiration from the actual perception of vision that takes place in the visual cortex of our brain. They make use of a filter/kernel to scan through the entire pixel values of the image and make computations by setting appropriate weights to enable detection of a specific feature.

The CNN is equipped with layers like convolution layer, max pooling layer, flatten layer, dense layer, dropout layer and a fully connected neural network layer. These layers together make a very powerful tool that can identify features in an image. The starting layers detect low level features that gradually begin to detect more complex higher-level features.

B. Long Short-Term Memory

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to overcome the limitations of traditional RNNs in capturing and learning long-range dependencies in sequential data. LSTMs have since become a fundamental building block in various deep learning applications, particularly in the field of natural language processing, speech recognition, and time series analysis.

The key strength of LSTMs lies in their ability to effectively capture and store information over extended sequences while mitigating the vanishing gradient problem that often plagues traditional RNNs. This is achieved through a more complex internal structure, including a memory cell with three gating mechanisms: the input gate, forget gate, and output gate.

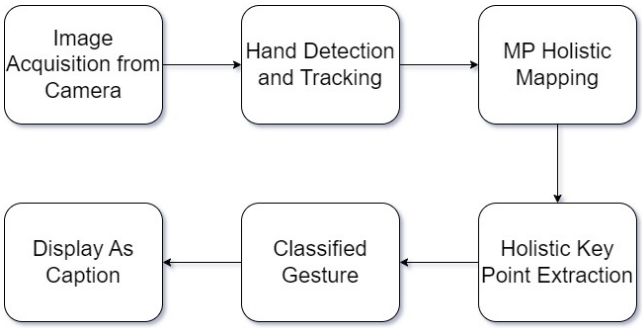
V. Methodology

The research methodology seamlessly integrates two complementary approaches, addressing distinct challenges in human interaction analysis. Focused on Bangla sign language recognition, particularly hand gestures, the methodology leverages Mediapipe for human skeleton detection, capturing sign language nuances through meticulous tracking of key skeletal points. This detailed skeletal tracking, combined with linguistic alignment, forms the basis for training a dedicated LSTM model. Optimized through paired key points and Bangla language representations, the LSTM adeptly captures the temporal dynamics of sign language expressions. Simultaneously, the methodology extends to human pose estimation in videos, employing OpenCV and 2D CNN models. The systematic approach encompasses video preprocessing, holistic keypoint extraction using the MediaPipe model, and integration of a 2D CNN for refinement. Post-processing techniques handle noise and outliers, and rigorous evaluation metrics ensure the validity and efficiency of the methodology. The discussion delves into the strengths, limitations, and potential enhancements of both approaches, providing a holistic overview of the research findings. Ethical considerations, particularly in video data usage, are duly acknowledged. This comprehensive approach combines advanced technologies for sign language recognition with the harnessing of Bangla-BERT's capabilities in a unified methodology, contributing to both gesture-based and natural language processing applications in the Bangla language.

মা ফুল গাছ ভালোবাসা	মা ফুল গাছ ভালোবাসে
বাবা ফুল গাছ ভালোবাসা	বাবা ফুল গাছ ভালোবাসে
ফুল গাছ ভালোবাসি	আমি ফুল গাছ ভালোবাসি
তুমি খারাপ বন্ধু	তুমি খারাপ বন্ধু

A. Flow Diagram

The flowchart explains the steps occuring to accocmplish the objectives of the project. These steps have been explained in a detail below:



1) Image Capture: Utilizing the web camera, the signing gestures are captured in real-time. Employing the OpenCV video stream, the entire signing sequence is recorded

2) Holistic Mapping: We adopted a holistic mapping approach during image capture. We focused on capturing the entire signing sequence using OpenCV video stream processing. This comprehensive method ensures a thorough representation of dynamic sign language gestures.

3) Hand Detection and Tracking: Captured images undergo preprocessing to scan for hand gestures, enhancing segments for improved model predictions. This refinement significantly boosts prediction accuracy.

4) Holistic key points extraction: Holistic key points extraction in our research captures interconnected hand gestures and facial expressions, encompassing the entire signing sequence. This approach ensures a comprehensive representation for accurate deep learning model training.

VI. CONCLUSIONS

The thesis investigates two neural network architectures: a 2D CNN achieves an impressive 96.97% overall accuracy and 83.33% confusion accuracy at epoch 75, showcasing its proficiency in image-based classification. In contrast, the LSTM model achieves a categorical accuracy of 40%, indicating challenges in sequence-based categorization. This comparison highlights the strengths of the 2D CNN in image tasks and identifies areas for LSTM optimization. The findings contribute to a nuanced understanding of neural network applications in classification tasks, guiding future research.

REFERENCES

[1] Shartaz Khan Akash, Debobrata Chakraborty, Mehedi Mahmud Kaushik, Barsan Saha Babu, and Md. Saniat Rahman Zishan on "Action Recognition Based Real-time Bangla Sign Language Detection and Sentence Formation"

[2] Nasima Begum, Rashik Rahman, Nusrat Jahan, Saqib Sizan Khan,Tanjina Helaly, Ashraful Haque, and and Nipa Khatun on "BornoNet: A Real-Time Bengali Sign-Character Detection and Sentence Generation System Using uantized Yolov4-Tiny and LSTMs"

[3] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, Hongdong Li from The Australian National University, Australian Centre for Robotic Vision (ACRV), University of Technology Sydney, DATA61-CSIRO on "Transferring Cross-domain Knowledge for Video Sign Language Recognition"

[4] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, Hongdong Li from The Australian National University, Australian Centre for Robotic Vision (ACRV), University of Technology Sydney, DATA61-CSIRO on "Wordlevel Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison"

[5] Shahjalal Ahmed, Md. Rafiqul Islam, Jahid Hassan, Minhaz Uddin Ahmed, Bilkis Jamal Ferdosi, Sanjay Sahak, Md. Shopon from Department of Computer Science and Engineering, University of Asia Pacific on "Hand Sign to Bangla Speech: A Deep Learning in Visionbased system for Recognizing Hand Sign Digits and Generating Bangla Speech"

[6] Ms.Ayesha Khatun, Mohammad Sajid Shahriar, Md. Hasibul Hasan,Krishna Das, Sabbir Ahmed, Md. Sakibul Islam from IUBAT on "A Systematic Review on the Chronological Development of Bangla Sign Language Recognition Systems" in 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition(icIVPR)

[7] F. M. Javed Mehedi Shamrat, Sovon Chakraborty, Md. Masum Billah, Moumita Kabir, Nazmus Shakib Shadin, Silvia Sanjana on "Bangla numerical sign language recognition using convolutional neural networks"

[8] Kanchon Kanti Podder, Muhammad E. H. Chowdhury, Anas M. Tahir,Zaid Bin Mahbub, Amith Khandakar, Md Shafayet Hossain, and Muhammad Abdul Kadir on "Bangla Sign Language (BdSL) Alphabets and Numerals Classification Using a Deep Learning Model"