~ SAVVY SIGNING ~

Real-Time Sign Language Gesture Recognition and Translation To Text System using CNN LSTMs.

SUPERVISOR	MR. STEPHEN OKETCH OBONYO
PRESENTED BY	KINYARI TRACIEBEL WAIRIMU
ADMISSION NO	146173

CONTENTS

Background Information

The world Federation of the Deaf reports approximately 72 million deaf people worldwide, with over 80% living in developing countries. There are more than 300 different sign languages used globally.

Sign language is divided into 2; natural gestures and formal cues Natural gestures are the unconscious physical expressions that accompany signs and add variation. These might include facial expressions indicating intensity, eyebrow movement to indicate emphasis, or head tilts for questioning.

In contrast, formal cues are deliberate/intentional and standardized when being developed, they have the same language as the spoken language of the community.

An example is the American Sign Language (ASL), which is the most widely used sign language in the world with the method of fingerspelling as a representation of the alphabets on cues (hand positions show each letter of the Latin alphabet).

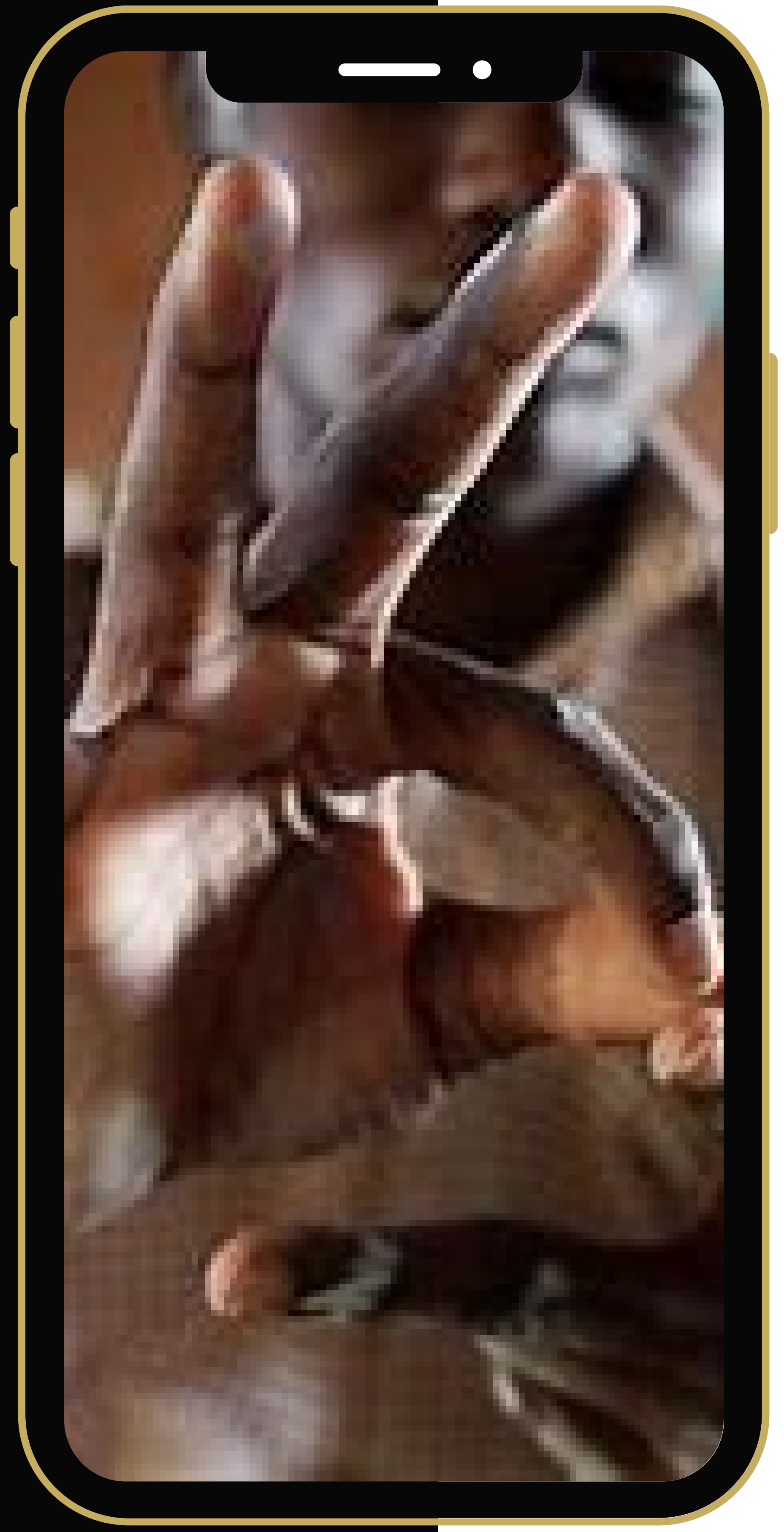
There has been advances in machine translation and speech translation which can be extended to sign language translation.

There are several projects that have been started to curb this problem but most of them do not look at certain aspects such as the problem, some models with the highest accuracy only allow static image input.

For example, in ASL some alphabets require hand movement, and therefore, need real-time, live translator.

There are Kenyan Sign Language dictionaries, or video resources that teach SL, but no known softwares that translate Kenyan Sign Language gestures into text.

Problem Statement



Communication barriers exist between hearing and the dead or hard-of-hearing community.

Learning SL can be **challenging** and the resources **not widely accessible**. **Current solutions** include **human interpreters** and **basic SL tutoring software**. **Interpreters** that teach SL are **expensive** and **limited**, while software **struggles** with **complex signs and body language.**

Mobile phones in this time and age are **widely accessible**, and with a **real-time SL translator application**, this will enable **effortless communication**, **promote inclusivity**, by **developing** such a **convenient** tool for **seamless conversations**.

Objectives

- Achieve **high word and sentence accuracy** through deep learning by using of a **comprehensive sign language dataset.**
- Acquiring available **Kenyan Sign language dataset** or performing **data augmentation on existing datasets** to increase its size for effective training of the model. In case of not enough data to train **create own dataset** by use of **known Kenya news broadcasting channels.**
- Introducing the aspect of **offline functionality.**
- Conducting **comprehensive testing** across various devices and scenarios and ensure **cross-browser compatibility.** Ensures that the application **functions consistently** across **different browsers** and **devices**
- Developing and testing a **user interface** that adheres to best practices for **accessibility**, ensuring a **smooth** and **intuitive experience** for **both deaf** and **hearing users.**

Related Works

Hand Talk

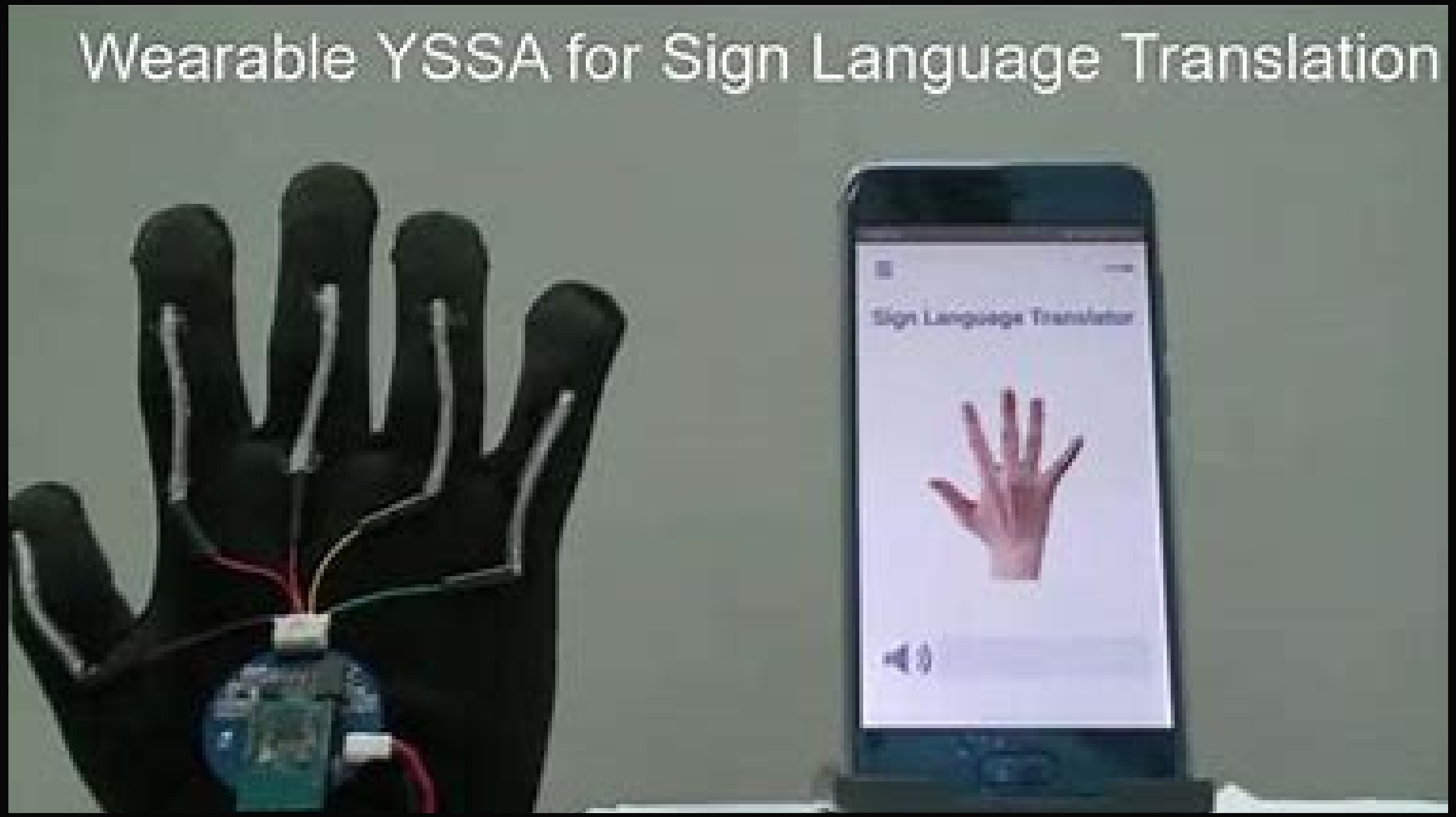
Created back in 2012. Works like a pocket translator, automatically translation text and voice into ASL (American Sign Language) and Libras (Brazilian SL).

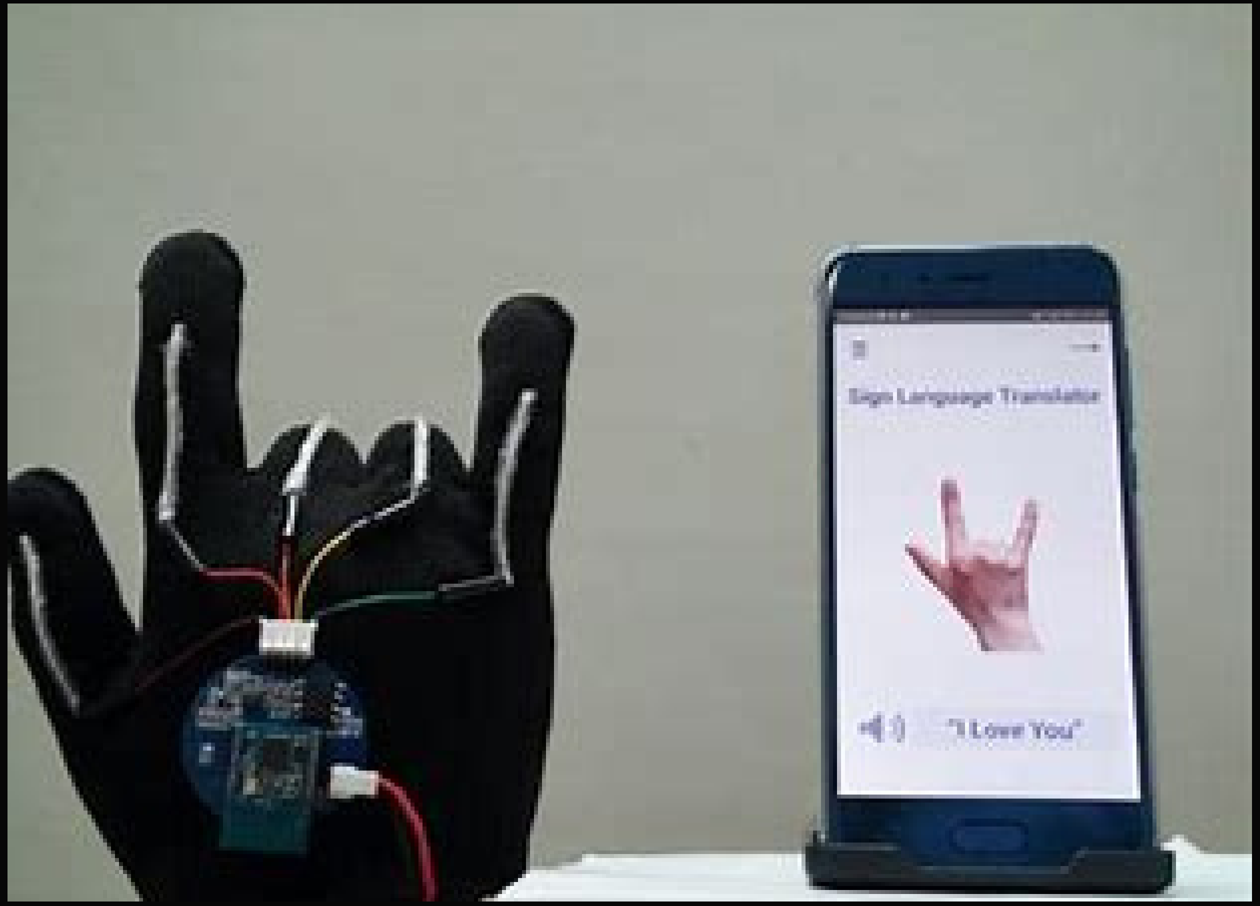
To perform these translations, it **relies** on **AI** and virtual translators.

You are able to save translations - favourite and most used translations

Customizing your virtual translators as you wish in terms of season, their attire.

Related Works





Wearable Glove - UCLA

The system **includes a pair of gloves** with **thin, stretchable sensors** that **run the length** of **each** of the **five fingers.**
The sensors, made from **electrically conducting yarns, pick up hand motions** and **finger placements** that **stand for individual letters, numbers, words and phrases.**

The device then **turns the finger movements** into **electrical signals**, which are **sent to a dollar-coin-sized circuit board worn on the wrist.**

The **board then transmits those signals wirelessly to a smartphone** that **translates them into spoken words** at the **rate** of about **one word per second.**

A **custom machine learning algorithm** turns the **gestures into letters, numbers and words they represent.**

The **system** was able to **recognize 600 signs**, including **each letter of the alphabet and numbers 0 through 9**

Related Works

SignAll

SignAll is an automated sign language translation solutions.

It has a **lab concept** that helps in **learning, practicing and giving quizzes** to **students who want to learn ASL.**

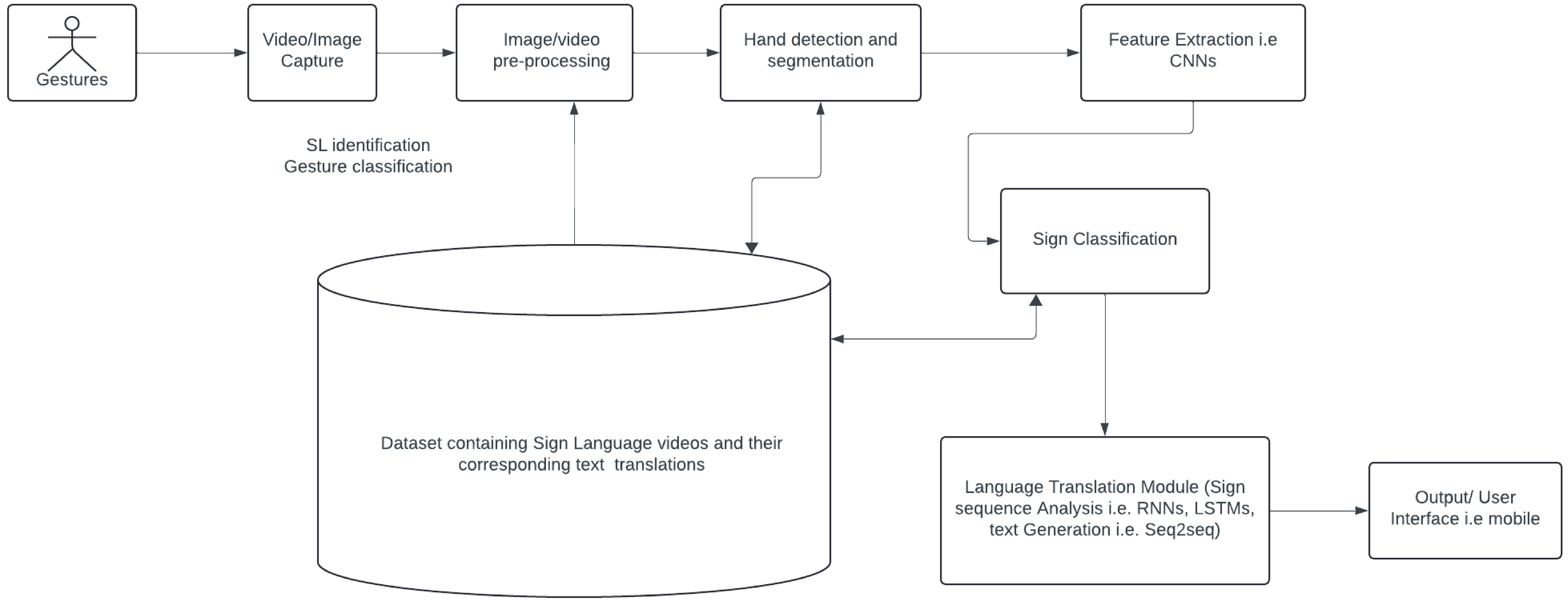
It **only registers signs** that **follow the 5 parameters of ASL** and you have to be specific and accurate for the translator to translate from SL to text.

Mostly deals with **finger spelling,** and **one must wear some specific gloves that come with the system for you to be able to sign and the translator** able to recognize the signs and gestures made by the user.

There is **no comprehensive 3D representation** of all signs.

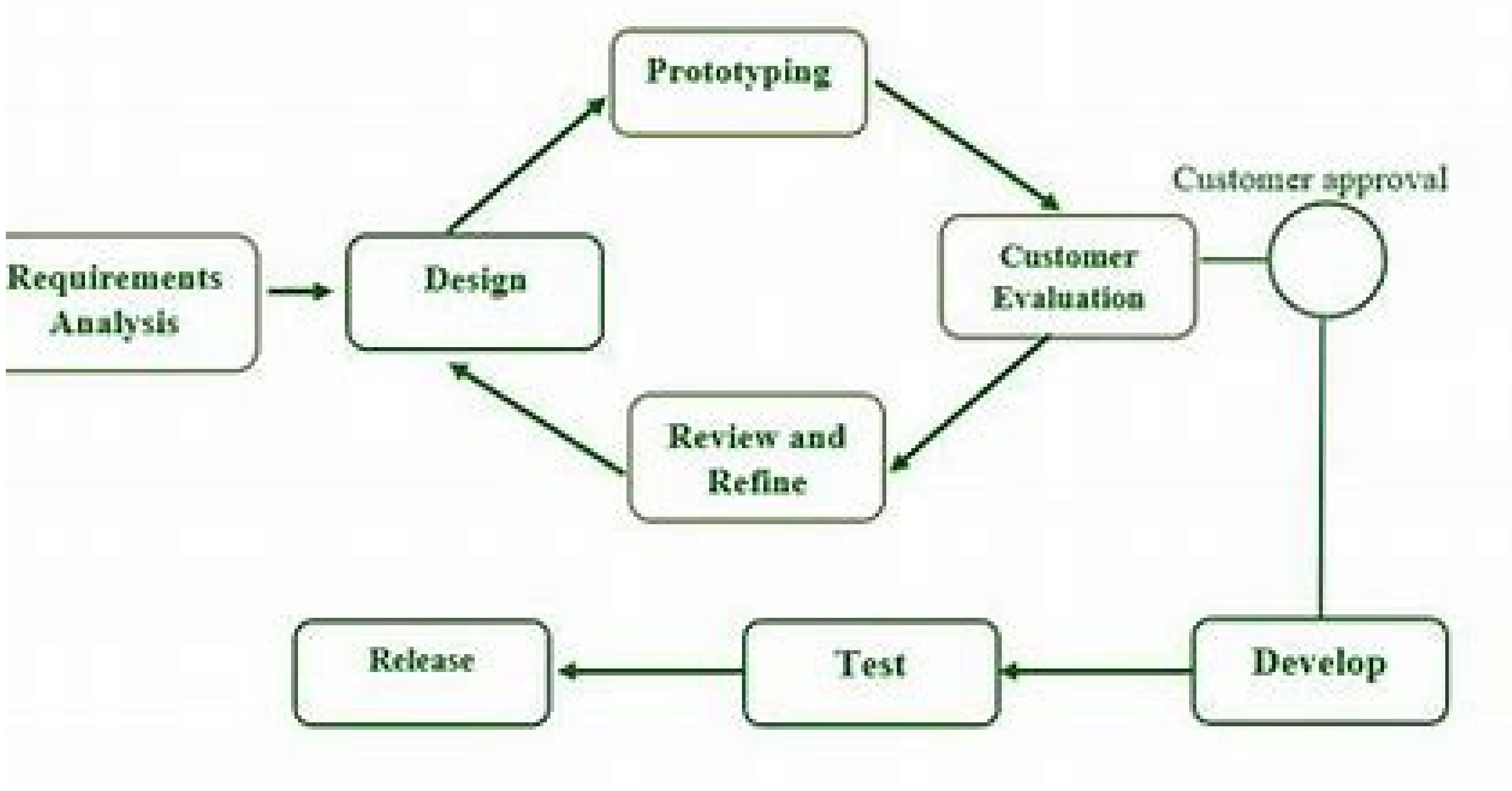
They also have a Software development KIt that allows developers to incorporate sign language input into their applications.

Conceptual Framework

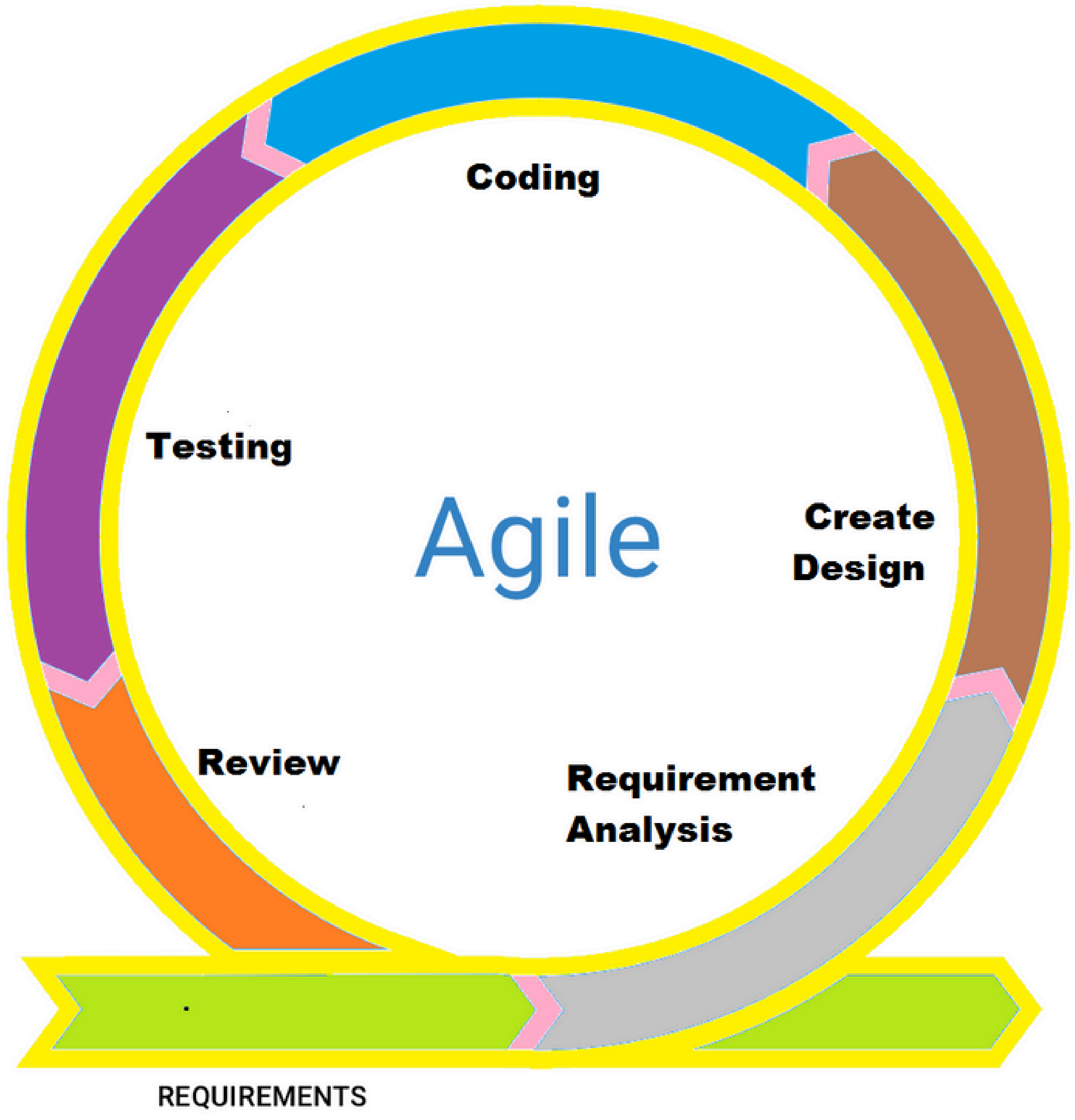


Research Methodology

- This chapter includes a structured approach to the project outlining the steps taken from the initial requirements gathering to the final deployment.



+



Research Methodology

- **Prototyping** allows one to **build functional versions** of the system, while **Agile's sprints** enable **regular updates** and **feature improvements.**

- The use of Agile software development **ensures** that the **development process** is **divided into manageable sprints**, with **each iteration** involving **design**, **development, testing, and evaluation of specific features**.

- The **synergy** between **Prototyping** and **Agile software development** allows for **continuous feedback from users**, allowing the **developer** to r**efine the application**, **improve model accuracy**, and **enhance the user experience throughout** the **development lifecycle.**

Design and Development Tools

Programming languages

- Python
- TensorFlow, handle the **computational backend**, **efficiently managing model training**, processing **large datasets**, and enabling **GPU acceleration.**
- Keras for CNN & LSTMS
- Front-end development ; JavaScript (React) or Flutter

Development Tools

- Jupyter Notebooks: Model training and experimentation
- Visual Studio Code (VSCode)

Framework

- Django : Back-end development
- WebRTC: Real-time video stream handling
- Firebase: **Offline** functionality
- TensorFlow/Keras: Deep learning model implementation

Version Control

- GitHHub: Source control

Firebase - **Realtime Database** and **Firestore cache data locally** when a **device goes offline.** Once the device **reconnects** to the internet, Firebase **automatically syncs any local changes to the server.**

Methods

MS-ASL Dataset

- The dataset contains 87000 images for training and testing, 29 labels; A - Z, space, delete, and nothing
- Label encoding; A = 0, B = 1, nothing = 28 for training. 25,000 annotated videos for real-life sign language.
- Videos must be downloaded separately, and they are mostly **hosted on platforms** like **YouTube or cloud repositories**. To work with images, you will need to download the videos from the platforms, **extract frames from these videos** to **create image datasets** for training **the CNN-LSTM model.**
- The files **include information about the video**, such as the **gesture label (class)**, the **path to the video**, and **metadata** like the **start and end time of the gesture within the video.**

DATASET

MSASL_train.json

MSASL_train.json


```
[{"org_text": "match [light-a-MATCH]", "clean_text": "match", "start_time": 0.0, "signer_id": 0, "signer": 0, "start": 0, "end": 83, "file": "match light-a-MATCH", "lab
{"org_text": "FAIL", "clean_text": "fail", "start_time": 0.0, "signer_id": 0, "signer": -1, "start": 0, "end": 74, "file": "FAIL", "label": 542, "height": 360.0, "fps":
{"org_text": "laugh", "clean_text": "laugh", "start_time": 0.0, "signer_id": 4, "signer": 26, "start": 0, "end": 31, "file": "SignSchool Laugh with Legs 2", "label": 31
{"org_text": "BOOK", "clean_text": "book", "start_time": 0.0, "signer_id": 0, "signer": -1, "start": 0, "end": 66, "file": "BOOK(3)", "label": 38, "height": 360.0, "fps
{"org_text": "sign-language", "clean_text": "sign language", "start_time": 0.0, "signer_id": 0, "signer": -1, "start": 0, "end": 75, "file": "SIGN-LANGUAGE-S-CLAW-F", "
{"org_text": "school", "clean_text": "school", "start_time": 1.101, "signer_id": 1, "signer": 44, "start": 33, "end": 110, "file": "ASL Vocabulary school", "label": 10,
{"org_text": "school", "clean_text": "school", "start_time": 4.671, "signer_id": 1, "signer": 44, "start": 140, "end": 206, "file": "ASL Vocabulary school", "label": 10
{"org_text": "easter", "clean_text": "easter", "start_time": 0.0, "signer_id": 2, "signer": 58, "start": 0, "end": 116, "file": "Easter", "label": 794, "height": 360.0,
{"org_text": "Boring ", "clean_text": "boring", "start_time": 0.0, "signer_id": 13, "signer": -1, "start": 0, "end": 71, "file": "ASL Boring ", "label": 46, "height": 3
{"org_text": "PAST", "clean_text": "past", "start_time": 0.0, "signer_id": 191, "signer": 13, "start": 0, "end": 32, "file": "PAST", "label": 510, "height": 720.0, "fps
{"org_text": "telephone", "clean_text": "phone", "start_time": 0.0, "signer_id": 2, "signer": 81, "start": 0, "end": 56, "file": "Telephone", "label": 120, "height": 36
{"org_text": "telephone", "clean_text": "phone", "start_time": 2.462, "signer_id": 2, "signer": 81, "start": 73, "end": 123, "file": "Telephone", "label": 120, "height"
{"org_text": "telephone", "clean_text": "phone", "start_time": 4.991, "signer_id": 2, "signer": 81, "start": 148, "end": 196, "file": "Telephone", "label": 120, "height
{"org_text": "LIBRARY", "clean_text": "library", "start_time": 0.0, "signer_id": 0, "signer": -1, "start": 0, "end": 73, "file": "LIBRARY(2)", "label": 168, "height": 3
{"org_text": "germany", "clean_text": "germany", "start_time": 0.0, "signer_id": 4, "signer": 46, "start": 0, "end": 36, "file": "SignSchool Germany 3", "label": 193, "
{"org_text": "like", "clean_text": "like", "start_time": 0.0, "signer_id": 269, "signer": 53, "start": 0, "end": 52, "file": "SignSchool really like", "label": 6, "heig
{"org_text": "COCHLEAR IMPLANT-[bent V version]", "clean_text": "cochlear implant", "start_time": 0.0, "signer_id": 0, "signer": 0, "start": 0, "end": 112, "file": "COC
{"org_text": "rainbow ", "clean_text": "rainbow", "start_time": 1.804, "signer_id": 144, "signer": -1, "start": 54, "end": 165, "file": "rainbow - ASL sign for rainbow"
{"org_text": "rainbow ", "clean_text": "rainbow", "start_time": 5.813, "signer_id": 144, "signer": -1, "start": 174, "end": 269, "file": "rainbow - ASL sign for rainbow
{"org_text": "rainbow ", "clean_text": "rainbow", "start_time": 9.354, "signer_id": 144, "signer": -1, "start": 280, "end": 435, "file": "rainbow - ASL sign for rainbow
{"org_text": "LETTER", "clean_text": "letter", "start_time": 0.0, "signer_id": 0, "signer": -1, "start": 0, "end": 73, "file": "LETTER(3)", "label": 579, "height": 360.
{"org_text": "FROM", "clean_text": "from", "start_time": 0.0, "signer_id": 0, "signer": -1, "start": 0, "end": 68, "file": "FROM(1)", "label": 298, "height": 240.0, "fp
{"org_text": "its/his/her", "clean_text": "his", "start_time": 0.0, "signer_id": 62, "signer": -1, "start": 0, "end": 108, "file": "Sign ITSHISHER", "label": 771, "heig
{"org_text": "HONG KONG-[fingerspelled-version]", "clean_text": "hong kong", "start_time": 0.0, "signer_id": 0, "signer": 0, "start": 0, "end": 72, "file": "HONG KONG-f
```

Model Training

- A pre-trained model, **performance-wise**, **transfer learning models beat** traditional deep learning models because the TL models include data (features, weights, etc.) from previously trained models, **possessing a comprehensive grasp of the features.**
- 87000 images in the MS-ASL dataset will be divided into training and testing data. **8000 images** for **training data** and **7000 images** for **testing.**
- **Merging datasets**, **normalizing** and **standardizing labeling** across the datasets being used. **Rename classes for uniformity** and **unify labels** for overlapping gestures. Create labels for unique, non-overlapping classes.
- **Resize images** to a **consistent size** (e.g., 64x64 or 224x224 pixels).
- Apply **data augmentation** (rotation, flipping, zooming).
- Normalize pixel values (e.g., scale between 0 and 1).
- **Concatenate image data** and **labels** using libraries like pandas, numpy, or TensorFlow.
- **Shuffle dataset** to **prevent model bias** towards **one dataset.**

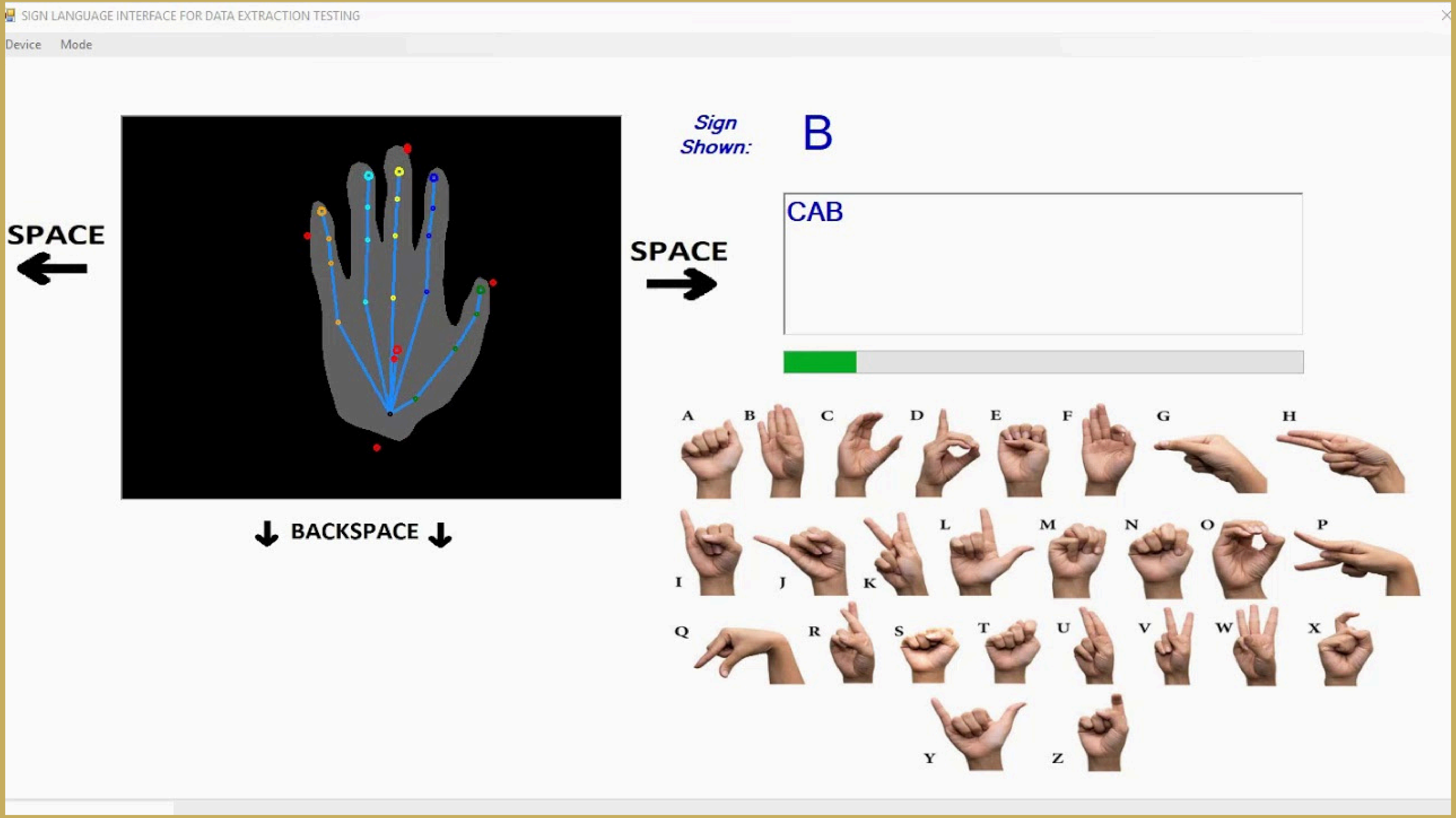
Training, Sampling and Testing

- **Ensure** that **both datasets** use the **same number of frames** per sequence.
- **Pad Sequences:** If the number of frames per video differs across datasets, you may need to pad or truncate sequences to a fixed length before feeding them into the LSTM.
- The LSTM network **complements** CNN by capturing temporal dependencies, which are **crucial** for **recognizing the sequence of gestures over time.**
- Together, CNN **extracts the visual features,** and the LSTM **processes the temporal flow of these gestures**, ensuring accurate recognition of SL phrases.
- **Unit testing** will be incorporated whereby **each individual module** of the system, including the **gesture recognition engine**, **text translation logic**, and **user interaction components**, is **tested** separately to **ensure that they work as expected.**
- Testing **across different devices** and **web browsers** to **confirm compatibility** and **minimal latency,** ensuring a **smooth user experience regardless** of the **platform**.

Validation Experiment

- **Combine and Split**: After merging, split the combined dataset into **training, validation, and testing sets**. You can use tools like **train_test_split** from **scikit-learn** to **maintain class balance across splits.**
- Split into 70% training, 15% validation, and 15% testing.
- **Cross-Validation:** Using **k-fold cross-validation** to **evaluate the model** on **different subsets** of the data **for better generalization.**
- The **sample size** will be from the **direct end users** who are the **deaf and hard-of-hearing community**. A **diverse representation** in terms of **age, proficiency in sign language**, etc.
- Participants **interact with the app in real-time,** and **gestures are translated to text.** Metric used to test performance will include **recognition accuracy**, **response time, user feedback.**
- Some **controlled variables** include **same lighting conditions**, **frame rate**, and **input quality during testing**.

Vision


SIGN LANGUAGE INTERFACE FOR DATA EXTRACTION TESTING
Device
Mode
SPACE
SPACE
BACKSPACE
Sign Shown:
B
CAB
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z

THANK YOU