German University in Cairo

Media Engineering and Technology Faculty

Boundary Detection For Continuous Sign Languge

Student

Zeina Walid Swelam

Supervisors

Dr. Nada Sharaf

Dr. Milad Ghantous

Bachelor Thesis - Spring 2025

OUTLINE

1. Introduction 
2. Problem Statment 
3. Background 
4 Related Work & Objective 
5. Proposed Design 
6. Implementation 
7. Results &Demo 
8. Future Work 

INTRODUCTION

- Communication shapes every aspect of our lives
- We use words, voices, and language to connect
- Communication isn’t limited to speaking

INTRODUCTION

- Sign language: A world without spoken words.
- Used by millions of deaf and hard-of-hearing people around the world.
- Helps engage in school, work, and social life.

INTRODUCTION

- Over 300 sign languages exist, each with unique cultural traits.
- Technology helps translate gestures into text or speech.
- Researchers still struggles with some great challenges.

PROBLEM STATMENT

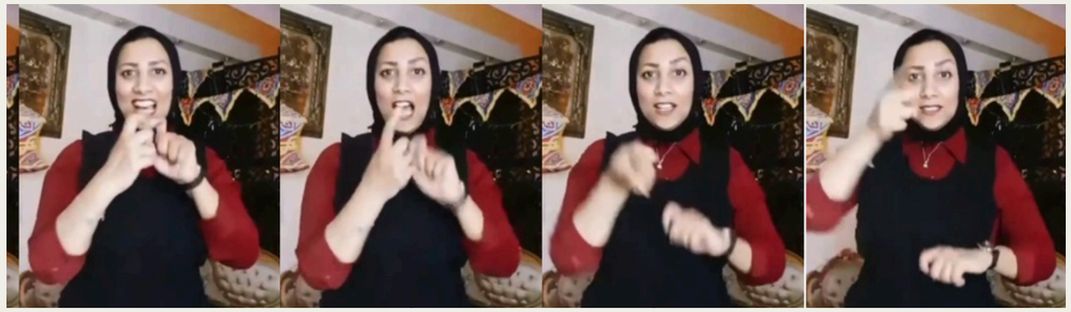
Howareyoudoingtoday?

How are you doing today?

PROBLEM STATMENT

- Segmenting continuous signs into individual ones is challenging.
  - No clear pauses make boundaries hard to detect.
  - Signing speed and style vary between individuals.
  - Coarticulation blends one sign into the next.
- Accurate segmentation enables translation of signs to text/speech.

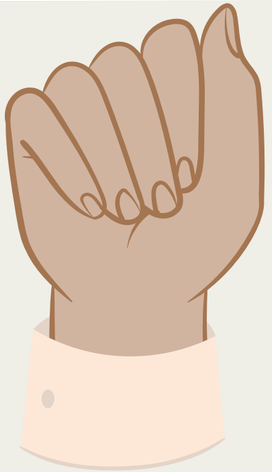
Sign 1 Sign 2



BACKGROUND-DEFENTIONS

Static Signs

- Signs that are made using a single hand posture with no movement.
- Example: Fingerspelling the letter "A"



Daynamic Signs

- Involve motion or movement of the hands or body.
- Example: Waving the hand to say “hello”.

BACKGROUND-DEFENTIONS

Isolated Signs

- Signs shown one at a time, with pauses in between.
- Example: Showing the word “name”, then stopping before the next sign.

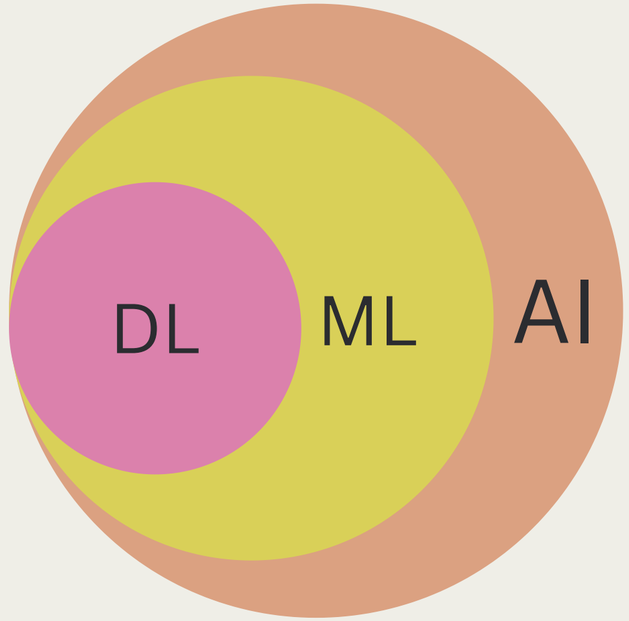
Continuous Signs

- Signs flow naturally and smoothly together in full sentences.
- Example: Signing “My name is John” without pauses

BACKGROUND-TECHNECIALS

Artificial Intelligence (AI)

- Field of computer science enabling machines to perform tasks requiring human-like intelligence, such as learning and decision-making.



Machine Learning (ML)

- Subset of AI where algorithms learn patterns from data to make predictions or classifications

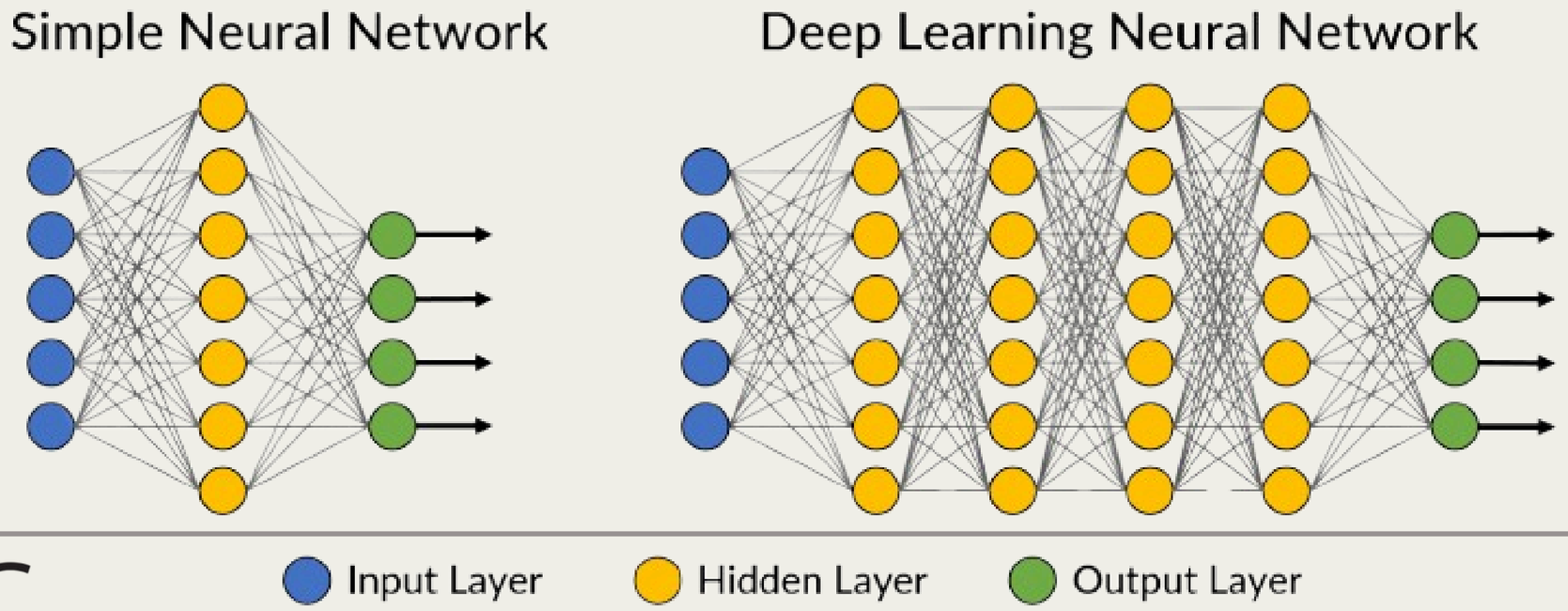
BACKGROUND-TECHNECIALS

Neural Networks

- Type of ML algorithm inspired by the brain.
- Used for tasks where traditional ML might struggle.

Deep Learning (DL)

- A subset of ML that uses deep neural networks meaning neural networks with many hidden layers



BACKGROUND-TECHNECIALS

Convolutional Neural Networks (CNNs)

- A type of neural network specially designed to work with images and visual data.
- CNNs can recognize patterns like edges, shapes, and textures, which helps in tasks like image classification, object detection, and facial recognition.

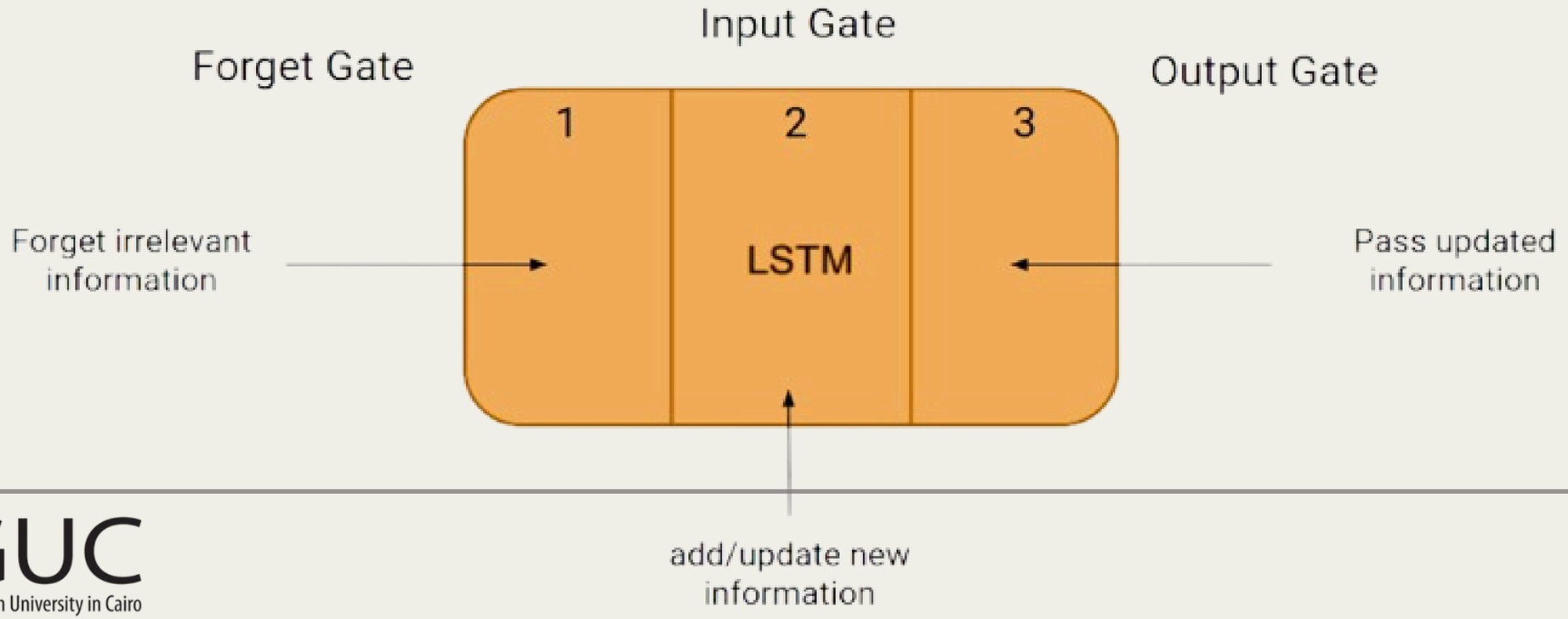
Recurrent Neural Networks (RNNs)

- A type of neural network designed for sequential data, like time series, text, or video frames.
- RNNs have a “memory”, they remember previous inputs, which helps them understand context.

BACKGROUND-TECHNECIALS

Long Short-Term Memory (LSTM):

- LSTM is a special kind of Recurrent Neural Network (RNN).
- remember important information for longer periods and forget irrelevant details.
- This solves the vanishing gradient problem in standard RNNs



BACKGROUND-TECHNECIALS

Computer Vision (CV):

- Field of(AI) enables machines to understand visual information just like humans do.

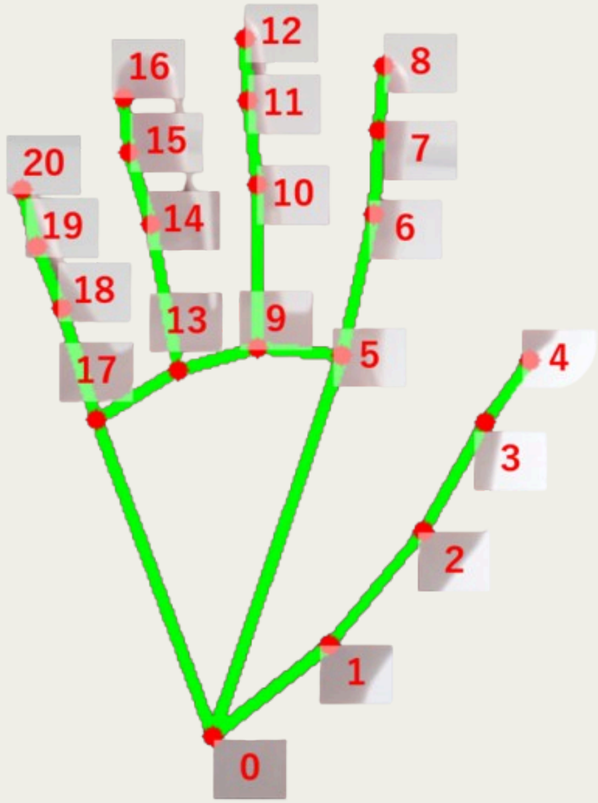
It's used in:

- Face recognition
- Gesture and sign language recognition
- Object detection
- Autonomous vehicles
- Medical imaging, and more

BACKGROUND-TECHNECIALS

MediaPipe:

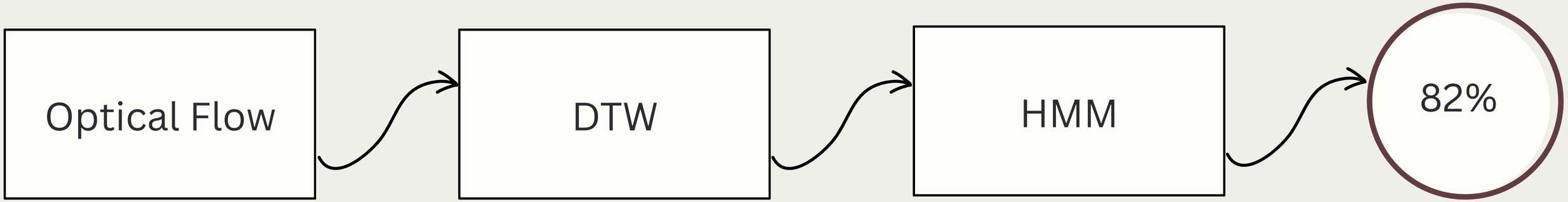
- Developed by Google.
- Used for real-time tracking of hands, faces, pose, and more.
- It offers pre-trained models and runs fast even on mobile devices.



RELATED WORK:

Approach 1: Han et al. (2009)

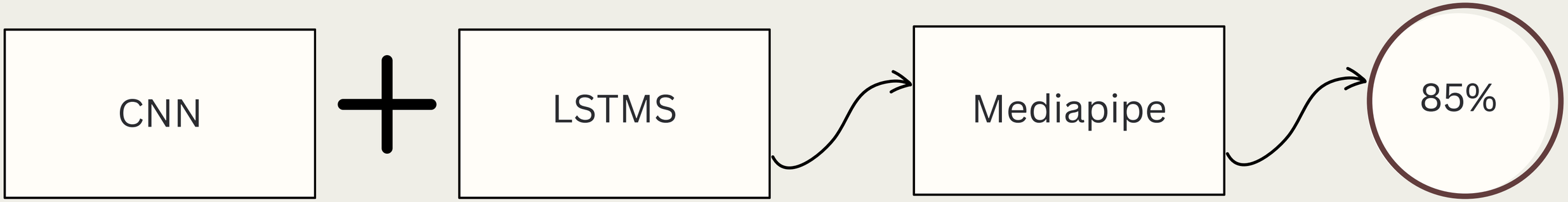
- Techniques: Optical Flow, DTW, HMM
- Results: 82% accuracy (Persian sign language)
- Limitations: Isolated signs only, sensitive to video changes, language-specific



RELATED WORK:

Approach 2: Srivastava et al.

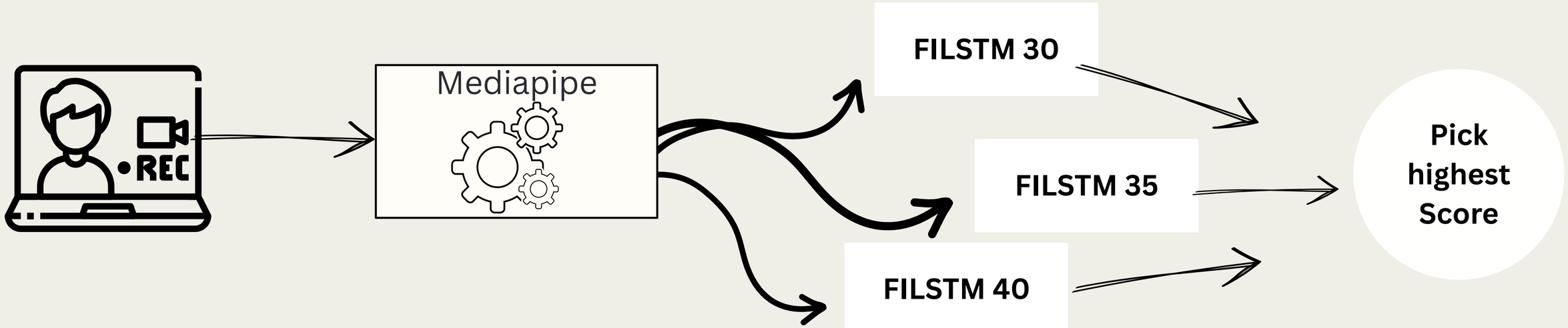
- Uses CNN-LSTM with MediaPipe Holistic for keypoint extraction
- Focuses on real-time recognition, 85% accuracy on small dataset
- Limitations: lacks explicit boundary detection, sensitive to dataset size



RELATED WORK: MOTION ANALYSIS WITHOUT DEEP LEARNING

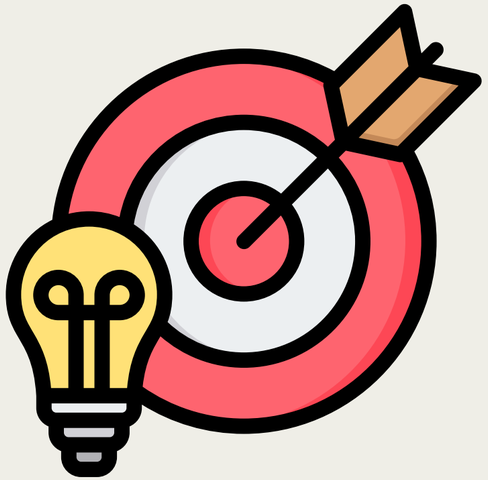
Approach 3: Albasmy(2024)

- Uses multiple Fixed Input-Length LSTMs (30, 35, 40 frames)
- Keypoints extracted via MediaPipe
- Ensemble boundary detection: selects model with highest confidence
- Limitations: small dataset, sensitive to lighting/background



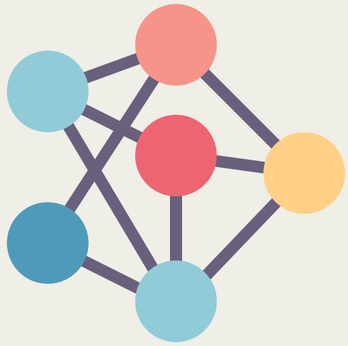
OBJECTIVES

- Support the deaf community
- Develop an innovative approach for continuous sign language recognition
- Apply machine learning models and computer vision techniques
- Detecting boundaries between signs in continuous signing
- Accurate segmentation of signing sequences into individual signs



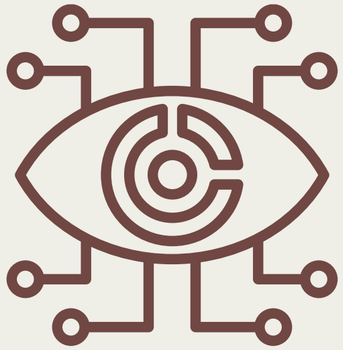
PROPOSED DEISGNS

LSTM-Based Approach



- Uses LSTM models to analyze sequential keypoint data for temporal patterns.

Computer Vision Approach



- Analyzes motion and posture directly from video frames using MediaPipe.

Combined Approach

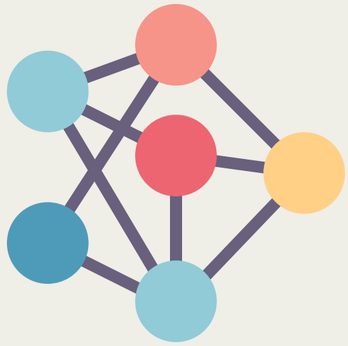
- Integrates computer vision for segmentation and BLSTM for better detection.

IMPLEMENTATION-TOOLS

- Google Colab: Used for preprocessing (MediaPipe) and training (GPU acceleration).
- Jupyter Notebooks: For prototyping and testing submodels.

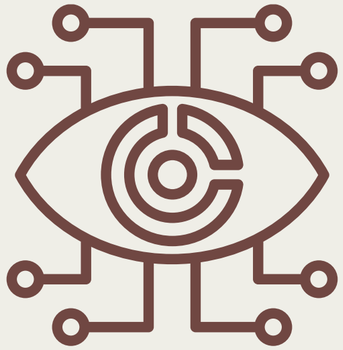
PROPOSED DEISGNS

LSTM-Based Approach



- Uses LSTM models to analyze sequential keypoint data for temporal patterns.

Computer Vision Approach

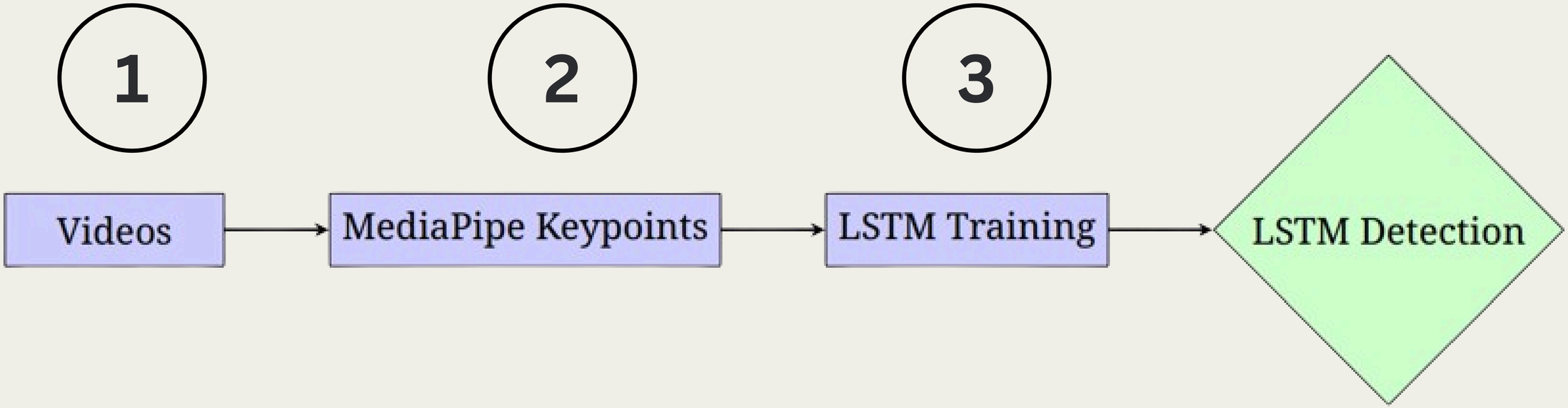


- Analyzes motion and posture directly from video frames using MediaPipe.

Combined Approach

- Integrates computer vision for segmentation and BLSTM for better detection.

PROPOSED DEISGN - LSTM APPROACH



IMPLEMENTATION - LSTM APPROACH (DATASET-EXTRACTION)

- (ArSL) 20-word dataset by Saafan(2023).
- 8,467 videos, 72 participants.
- Includes static and dynamic signs.
- Preprocessed with MediaPipe.

English	Arabic	Videos	English	Arabic	Videos
1. baby	طفل	430	11. happy	سعيد	445
2. eat	يأكل	440	12. hear	يسمع	420
3. father	أب	452	13. house	منزل	430
4. finish	نيهي	440	14. important	مهم	446
5. good	جيد	414	15. love	حب	435
6. mall	مركز تجاري	427	16. me	أنا	430
7. mosque	مسجد	427	17. mother	أم	406
8. normal	عادي	410	18. sad	حزين	420
9. thinking	يفكر	366	20. worry	قلق	349

MPLEMENTATION DETAILS - LSTM APPROACH

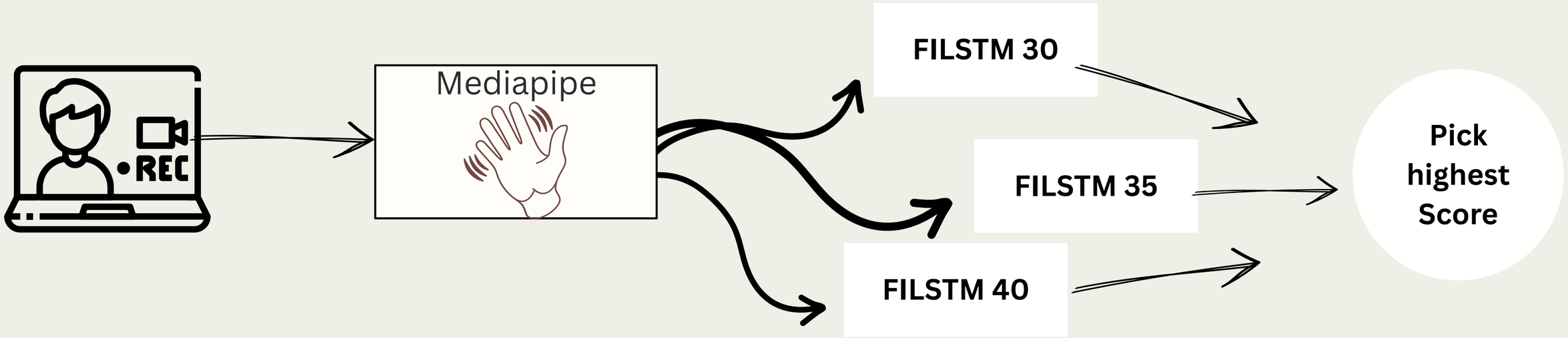
Baseline Model (30+35+40):

- Sequential processing, same as related work.
- Limitations: Small dataset (5,888 samples, 14 words), high computational cost.

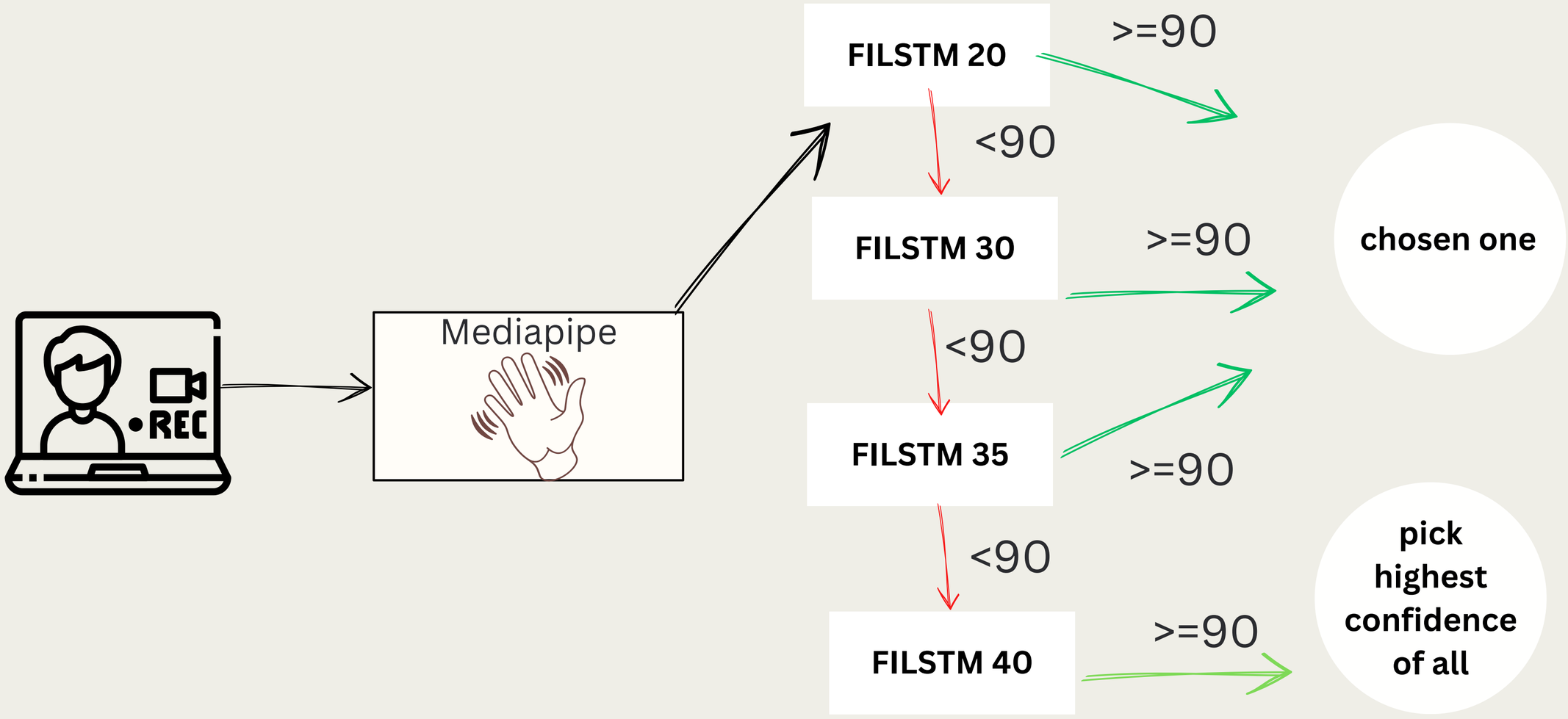
Improved Approach:

- Dataset: 8,467 samples, 20 words for better generalization.
- Parallel processing reduces computational cost.
- Architecture: 3 LSTM layers (64, 128, 64 units), ReLU, batch normalization.

PROPOSED DESIGN



PROPOSED DESIGN

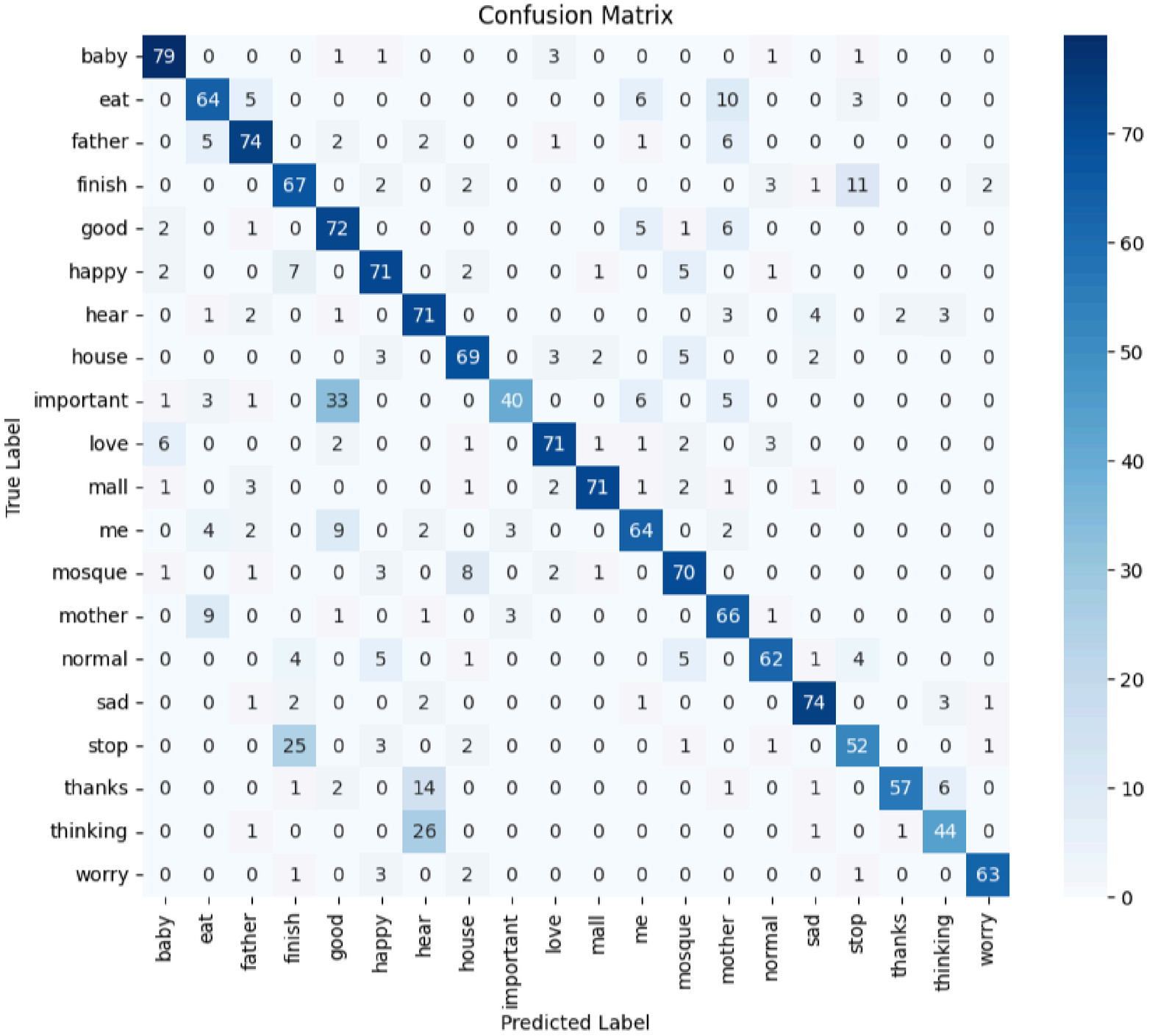


TRAINING DETAILS: LSTM APPROACHES

- Submodels (20, 30, 35, 40):
  - Trained on keypoint sequences.
  - Adam optimizer, ReLU activation.
- Confusion Matrices:
  - Model 20: 65.75%.
  - Model 30: 77.07%.
  - Model 35: 71.21%.
  - Model 40: 68.25%.

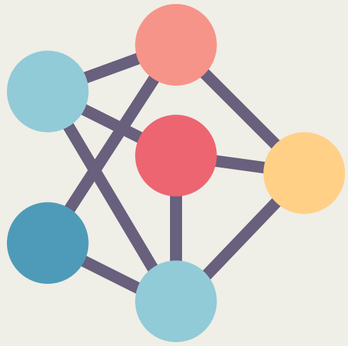
Confusion matrix used to evaluate training for all models

Model 30:



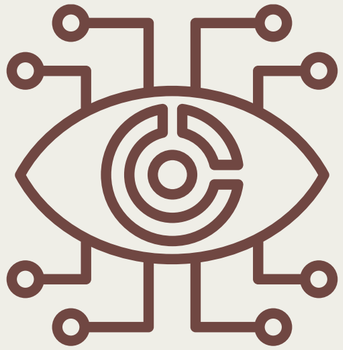
PROPOSED DEISGNS

LSTM-Based Approach



- Uses LSTM models to analyze sequential keypoint data for temporal patterns.

Computer Vision Approach

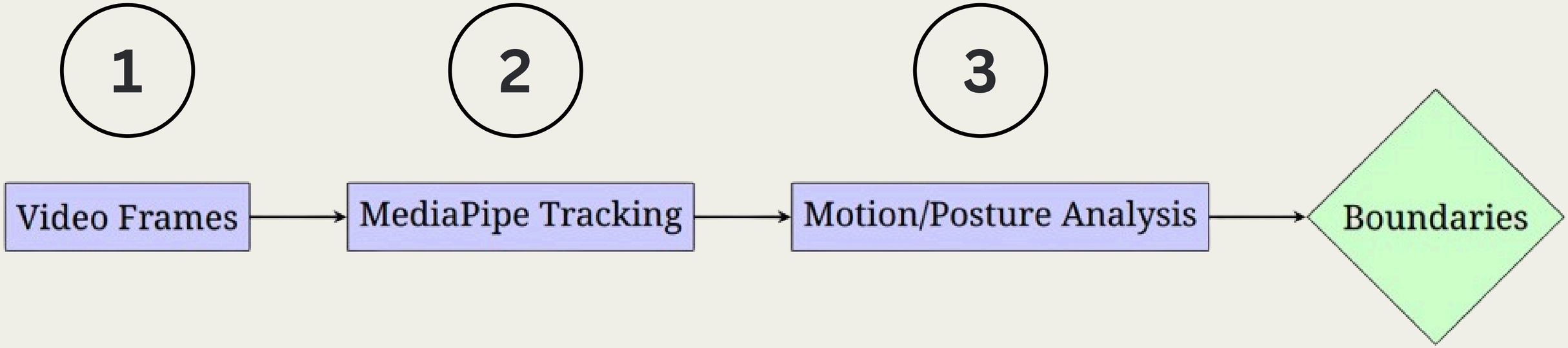


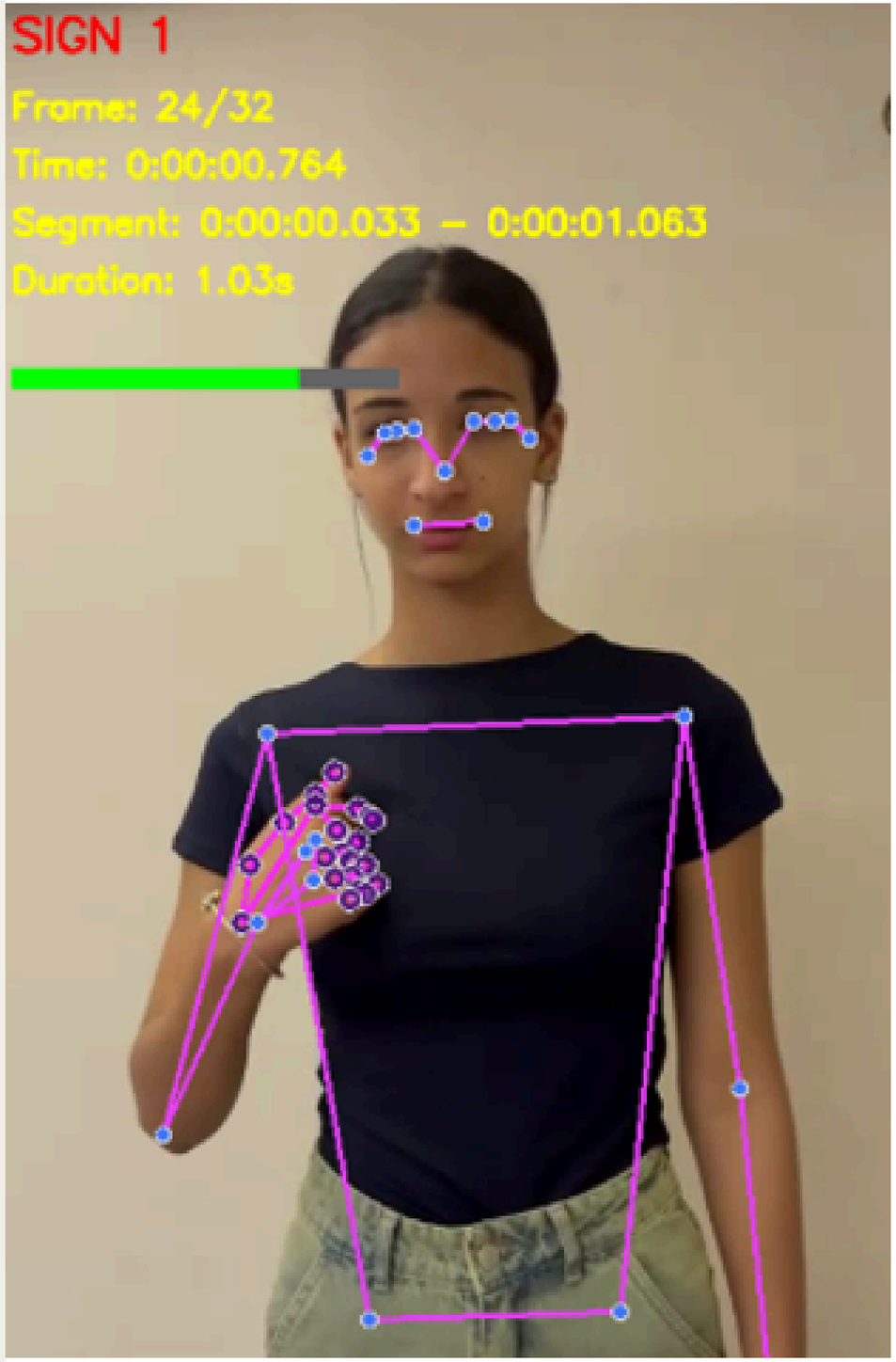
- Analyzes motion and posture directly from video frames using MediaPipe.

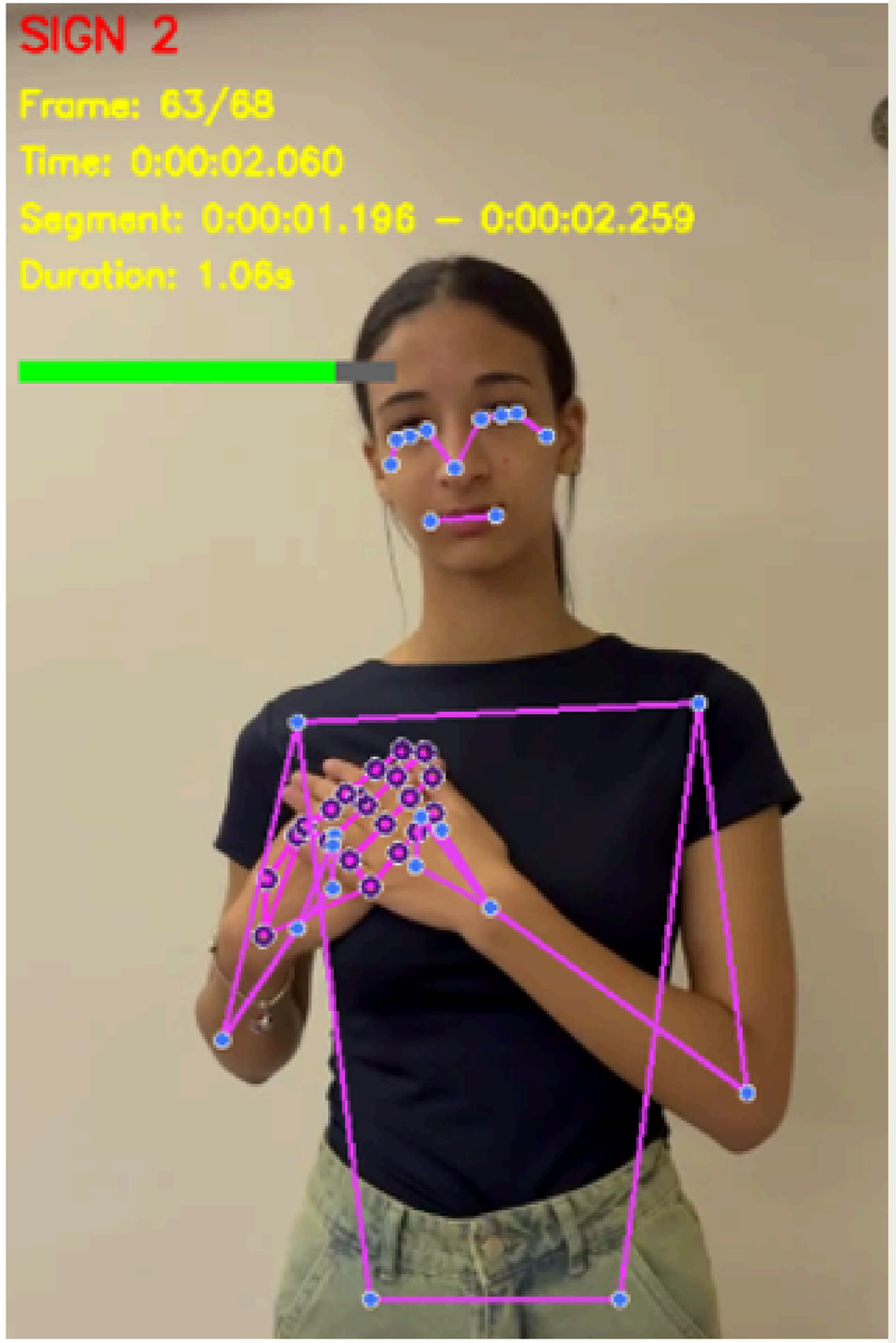
Combined Approach

Integrates computer vision for segmentation and BLSTM for better detection.

PROPOSED DEISGN - COMPUTER VISION




SIGN 1
Frame: 24/32
Time: 0:00:00.764
Segment: 0:00:00.033 – 0:00:01.063
Duration: 1.03s

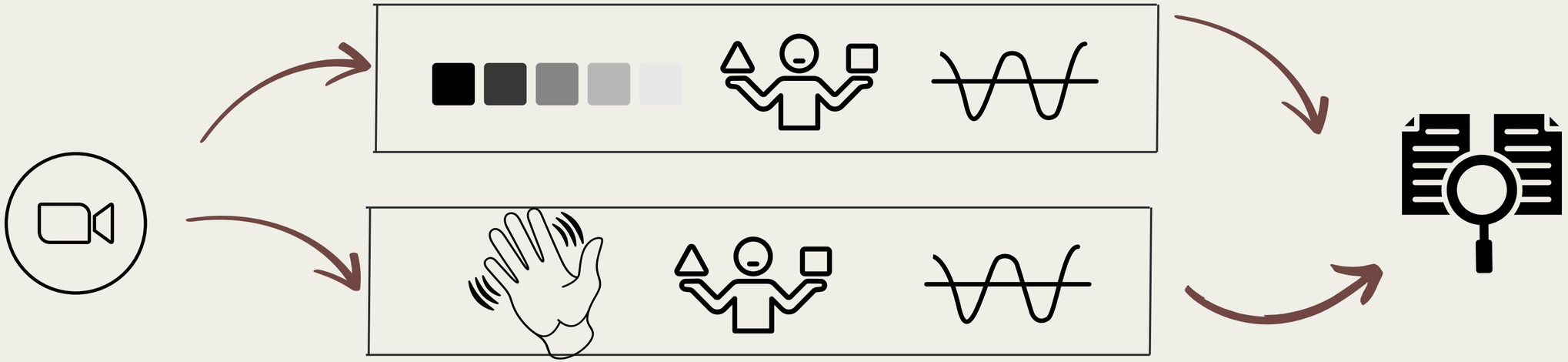

SIGN 2
Frame: 63/68
Time: 0:00:02.060
Segment: 0:00:01.196 – 0:00:02.259
Duration: 1.06s

IMPLEMENTATION - COMPUTER VISION

1. Global+Hand Motion: Frame differencing (motion detection) + MediaPipe hand tracking to isolate hand movements.
2. Full-Body+Hand Motion: MediaPipe Holistic extracts body and hand landmarks for comprehensive analysis.
3. Posture-Based Gesture Analysis: Posture change detection + stability checks to differentiate static vs. dynamic signs.

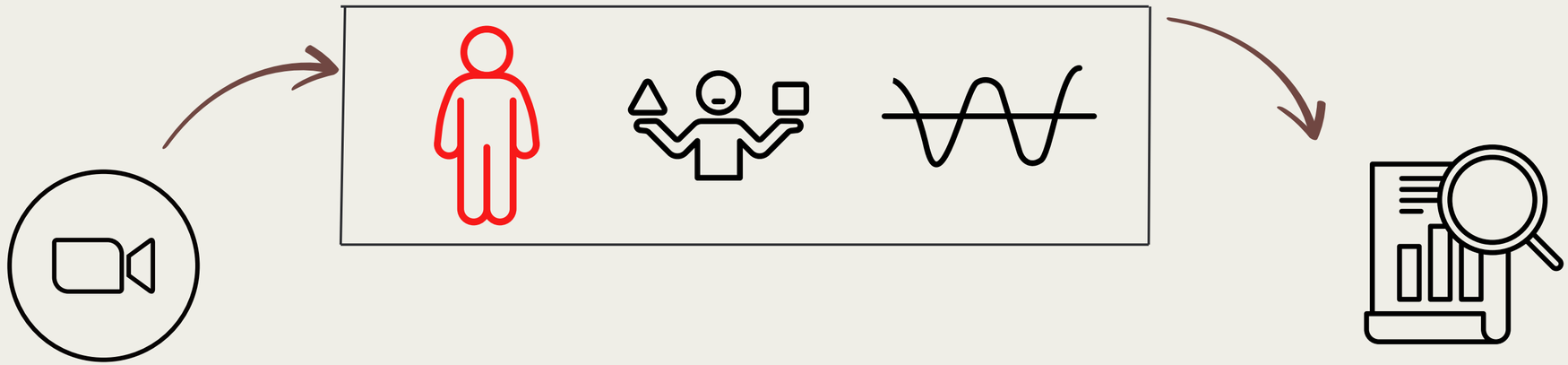
IMPLEMENTATION - GLOBAL+HAND MOTION METHOD

- **How It Works:** Uses two motion signals—one from whole frames, one from hands.
- **Techniques:** Grayscale conversion, frame differencing, normalization, convolution filter, sliding window, adaptive thresholds, Euclidean distance.



IMPLEMENTATION - FULL-BODY+HAND MOTION METHOD

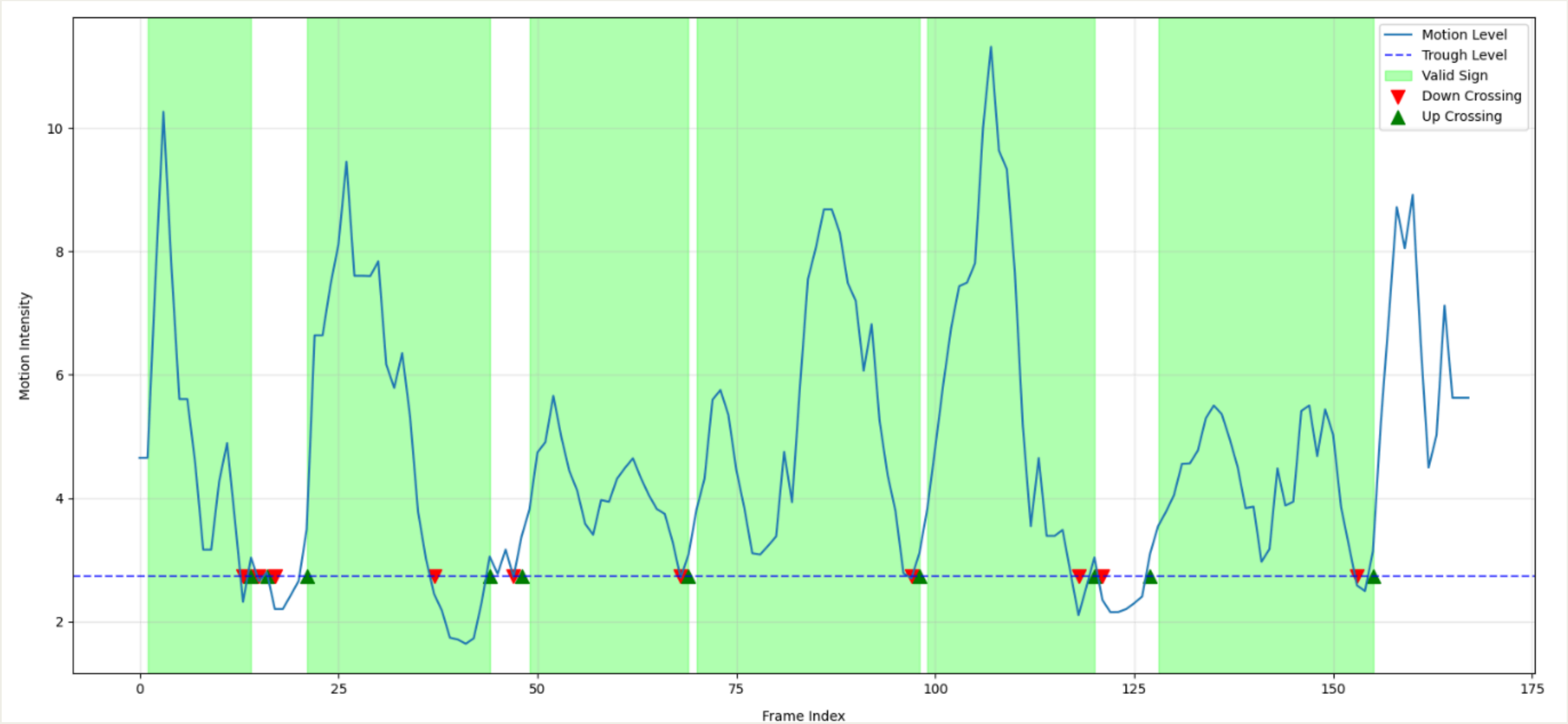
- **How It Works:** Watches 258 body and hand dots to spot motion patterns.
- **Techniques:** Euclidean distance, averaging, normalization, convolution filter, sliding window, dynamic thresholds, crossing pattern.



IMPLEMENTATION - POSTURE-BASED METHOD

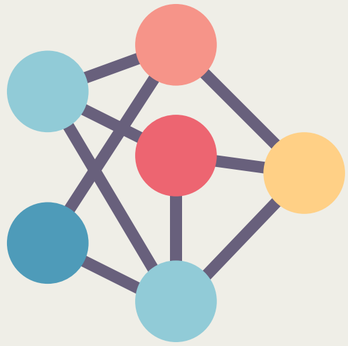
- **How It Works:** Tracks body motion and pose changes to find boundaries.
- **Techniques:** Euclidean distance, normalization, convolution filter, sliding window, sensitive adaptive thresholds, posture queue, Euclidean distance.

Graph-Based Analysis for Posture Motion Segmentation



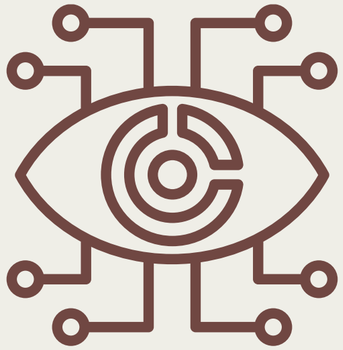
PROPOSED DEISGNS

LSTM-Based Approach



- Uses LSTM models to analyze sequential keypoint data for temporal patterns.

Computer Vision Approach

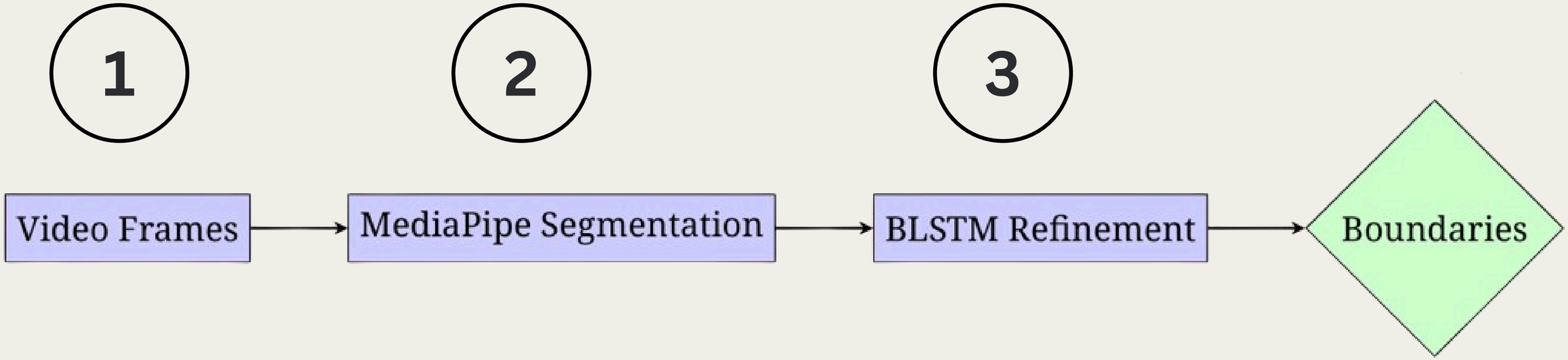


- Analyzes motion and posture directly from video frames using MediaPipe.

Combined Approach

- Integrates computer vision for segmentation and BLSTM for better detection.

PROPOSED DEISGN - COMBINED



IMPLEMENTATION DETAILS - COMBINED APPROACH

- BLSTM model achieved 95.4% accuracy on isolated sign recognition.
- Combined it with CV to refine segmentation and recognition.
- Integration failed due to oversegmentation and unreliable boundaries.

EVALUATION METRICS

- TP: Correctly predicted signs.
- FN: Detected but misclassified signs.
- FP: Missed ground truth signs.
- TN: Correctly identified non-signs.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

RESULTS- LSTM APPROACHES

Baseline (30+35+40)

- Video 1: 25% F1-score.
- Video 2: 75% F1-score.
- Video 3: 88% F1-score.
- Average sentence-level F1-score: 25+75+88 3 =62.6%

Sequential (20+30+35+40)

- .Video 1: 50% F1-score.
- Video 2: 85% F1-score.
- Video 3: 67%F1-score.
- Average sentence-level F1-score: 67.3% F1-score.

BLSTM

- BLSTM (Standalone):
  - 95% word-level F1-score.
- BLSTM + CV Pipeline:
  - 21.4% word-level F1-score.

RESULTS- COMPUTER VISION APPROACHES

Global+Hand Motion:

- Video 1: 67% F1-score.
- Video 2: 67% F1-score.
- Video 3: 89% F1-score.
- Average sentence-level F1-score:74.3%

Full-Body+Hand Motion

- .Video 1: 50% F1-score.
- Video 2: 57% F1-score.
- Video 3: 30%F1-score.
- Average sentence-level F1-score:45.7%.

Posture-Based

- .Video 1: 89% F1-score.
- Video 2: 100% F1-score.
- Video 3: 89%F1-score.
- Average sentence-level F1-score:92.7%% .

COMPARISON OF RESULTS

Sequential (20+30+35+40)

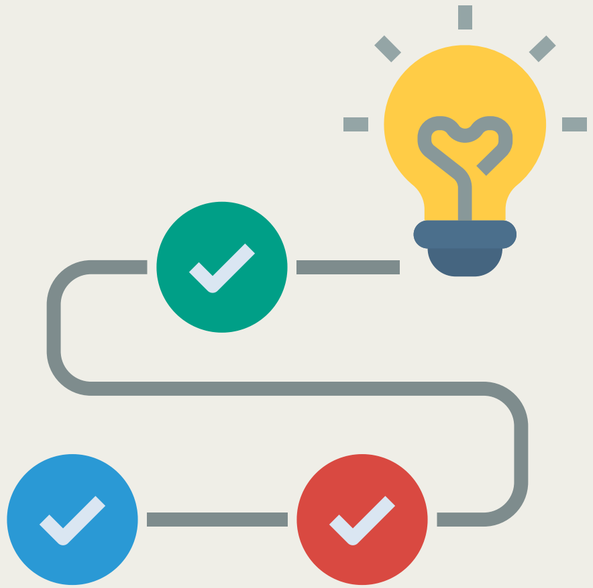
- .Video 1: 50% F1-score.
- Video 2: 85% F1-score.
- Video 3: 67%F1-score.
- 67.3% F1-score.

Posture-Based

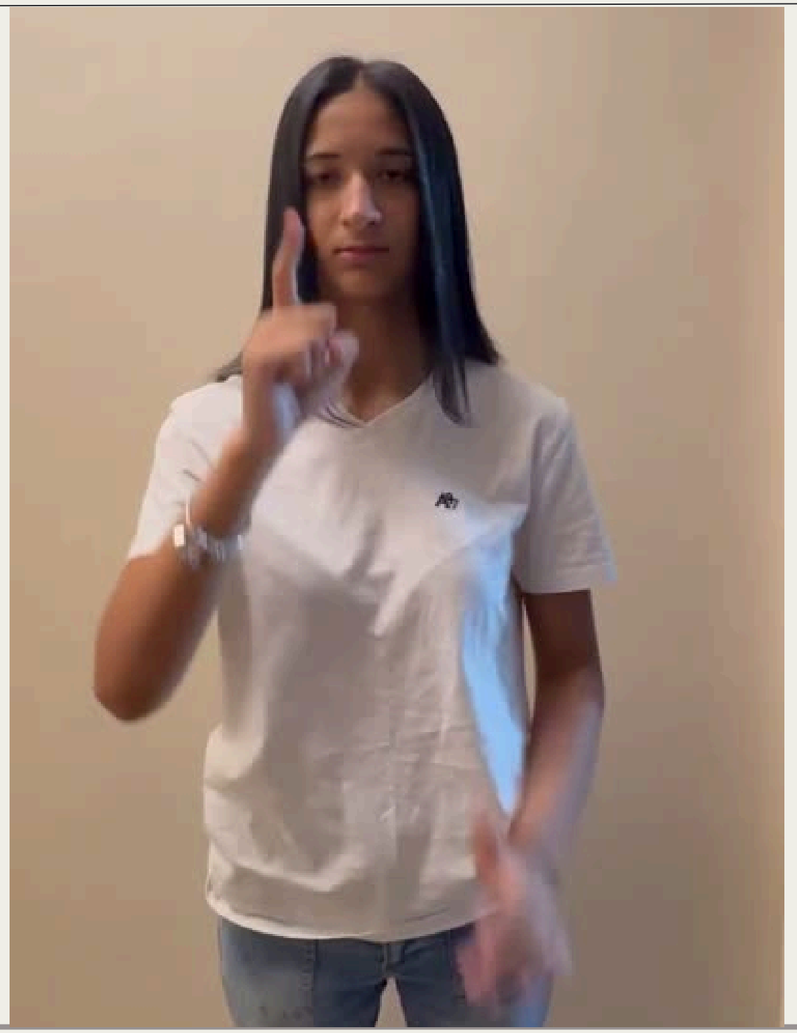
- .Video 1: 89% F1-score.
- Video 2: 100% F1-score.
- Video 3: 89%F1-score.
- 92.7% F1-score.

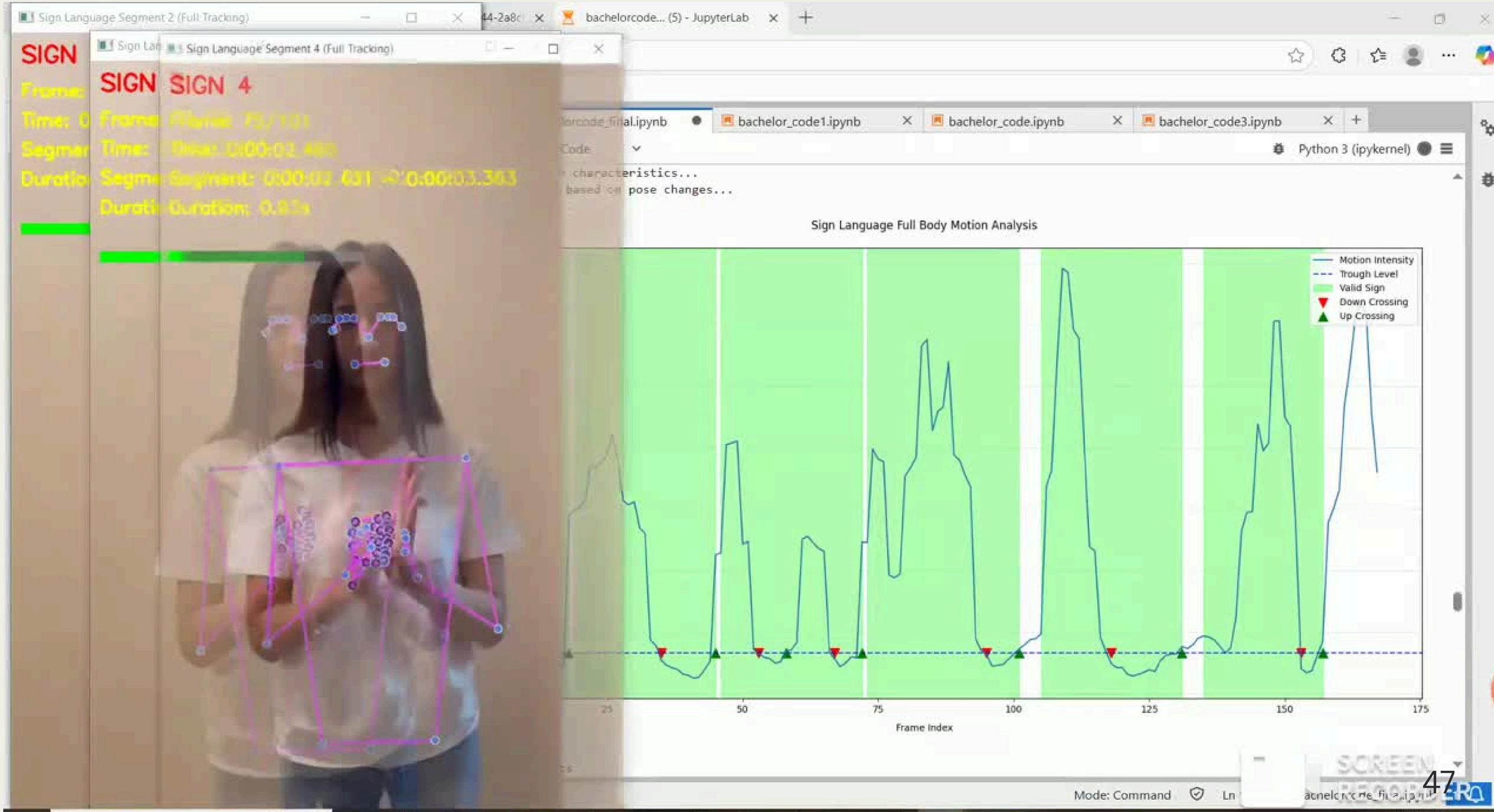
CONCLUSION

- key findings on boundary detection.
- Posture-Based Gesture Analysis achieved a 92.7% F1-score.
- Unsupervised CV methods show strong potential for ArSL recognition despite dataset and computational challenges.
- Future work will build on these insights to enhance system performance.



DEMO




Sign Language Segment 2 (Full Tracking)
Sign Language Segment 4 (Full Tracking)
SIGN
SIGN 4
bachelorcode... (5) - JupyterLab
bachelor_code1.ipynb
bachelor_code.ipynb
bachelor_code3.ipynb
Python 3 (ipykernel)
characteristics...
pose changes...
Sign Language Full Body Motion Analysis
Motion Intensity
Trough Level
Valid Sign
Down Crossing
Up Crossing
Frame Index
Mode: Command
Type here to search
ENG
12:26 PM
6/12/2025

FUTURE WORK

- Expanded Dataset
- Developing hybrid CV-LSTM approaches
- Enhancing CV methods

REFERENCES

1. Nour Albasmy and Milad Ghantous. Real-time boundary detection for continuous arabic sign language translation. In 2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), pages 21–26, 2024.
2. Nigar Alishzade and Jamaladdin Hasanov. Azsld: Azerbaijani sign language dataset for fingerspelling, word, and sentence translation with baseline software. Data in Brief, 58:111230, 2025.
3. Mahmoud Msaafan. Arabic sign language dataset, 2023. Accessed: May 28, 2025.
4. World Health Organization. Deafness and hearing loss, 2019. Accessed: Mar. 18,2025.
5. Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. Word separation in contin-uous sign language using isolated signs and post-processing. Expert Systems with Applications, 249:123695, 2024.
6. Jay Suthar. Sign language recognition for static and dynamic gestures, 2021. Accessed: May 5, 2025.

Thank you!

Any Questions?


TRUSTED


FAST