American Sign Language Recognition Using Dual-Input CNNs

Mohamed Aiman Alkhozendar
mohamed.alkhozendar@bahcesehir.edu.tr
Bahçeşehir University

Introduction

American Sign Language (ASL) is a vital communication method for the Deaf community. Traditional recognition systems rely on static images and struggle in real-time due to lighting, background, and pose variation. This project presents a robust real-time ASL recognition system using a dual-input CNN architecture combining grayscale images and 3D hand landmarks.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi ultricies eget libero ac ullamcorper. Integer et euismod ante. Aenean vestibulum lobortis augue, ut lobortis turpis rhoncus sed. Proin feugiat nibh a lacinia dignissim. Proin scelerisque, risus eget tempor fermentum, ex turpis condimentum urna, quis malesuada sapien arcu eu purus.

Motivation and Challenges

Image-only models fail under real-world variability. Webcam input introduces lighting, background clutter, and hand jitter. A dual-input system addresses these with visual and spatial awareness.

- **Visual Similarity Between Signs** Many ASL letters (M,N,S,T) have similar hand shapes , making it hard for the machine to observe
- **Live Video Distortions** Real-time environments introduce lighting variation, motion blur, background clutter, and partial occlusions — conditions that static training data doesn't represent well.
- **Heavy Models Struggle in Real Time** Pretrained models like ResNet50, although accurate in validation, are too slow and inconsistent for real-time feedback under noisy conditions.

My Contributions

Dual-Input CNN Architecture: We designed a novel convolutional neural network that processes two input streams—grayscale hand images and 3D landmark coordinates extracted using MediaPipe. This fusion improves spatial awareness and robustness.

Landmark-Guided Input Stream: I introduced landmark-based representation into the classification pipeline, allowing the model to focus on hand structure rather than just appearance.

High Accuracy Under Real-World Conditions: Our model maintains strong performance across various real-time webcam conditions including lighting changes, background clutter, and occlusion.

Real-Time Interactive Feedback: We implemented gesture stability logic and speech synthesis to provide real-time translation from signs to spoken language.

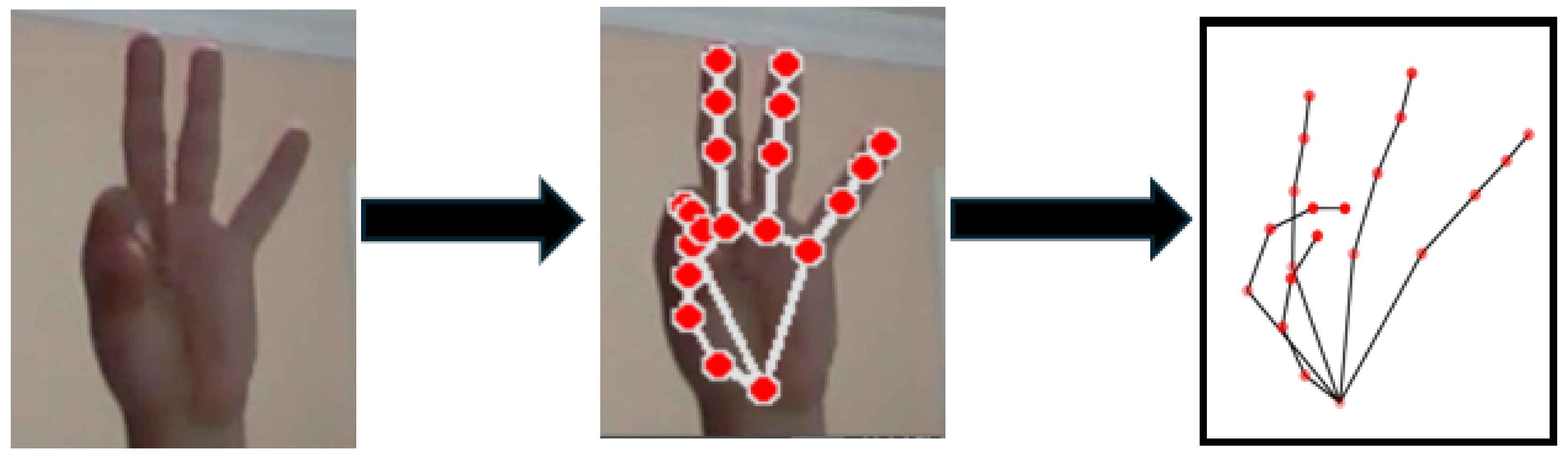


Figure 1. Landmarks extraction

Methodology

I have developed a dual-branch Convolutional Neural Network (CNN) that simultaneously processes grayscale hand images and 3D hand landmark arrays. This multimodal approach enhances robustness in real-world conditions. **Data Pipeline:**

- Webcam captures hand image.
- MediaPipe extracts 21 hand landmarks (x, y, z).
- Grayscale image and landmark array are passed to the CNN model.
- Model outputs a letter prediction if gesture is stable for 1.5 seconds.
- Letters form a sentence, which is spoken aloud with text-to-speech.

Model Architecture:

- **Image Branch:** 3 Conv2D layers + MaxPooling + Dropout.
- **Landmark Branch:** 2 Conv2D layers + Dropout.
- **Fusion:** Flattened outputs are concatenated and passed to dense classification layers.

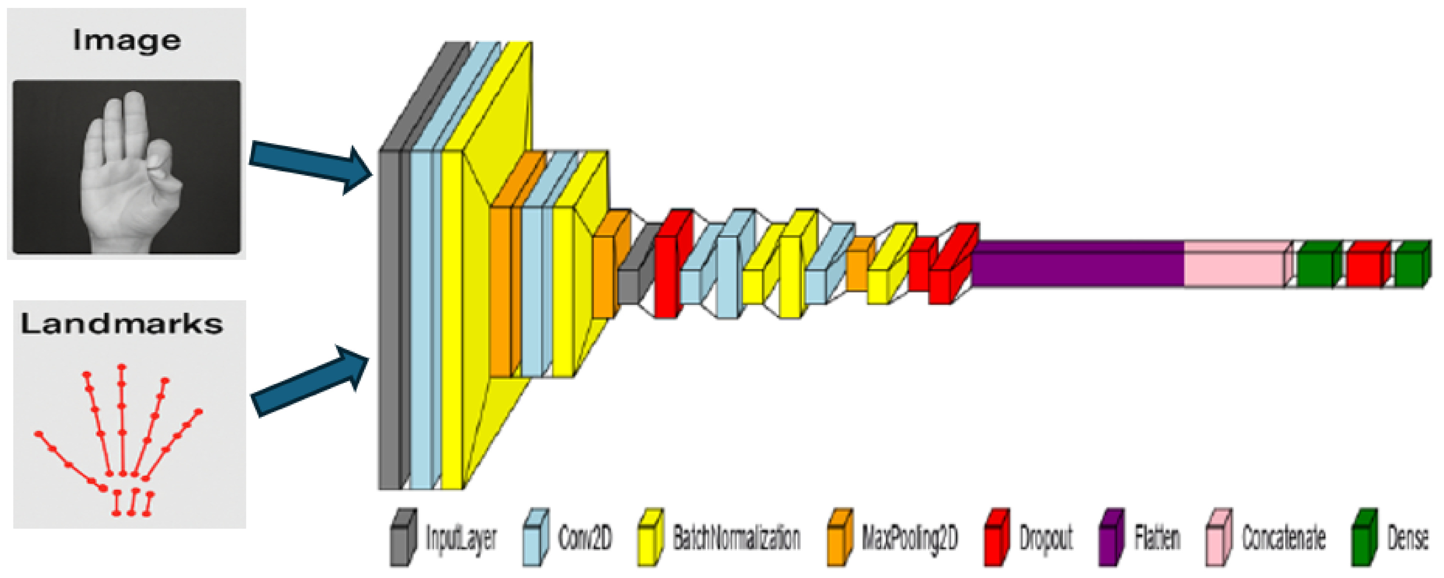


Figure 2. Dual-inputCNN

Experiments and Results

Baseline Comparison:

- **Image-only CNN:** average validation accuracy but poor generalization in real-time.
- **ResNet50:** Accurate but latency-prone and horrible in live usage.

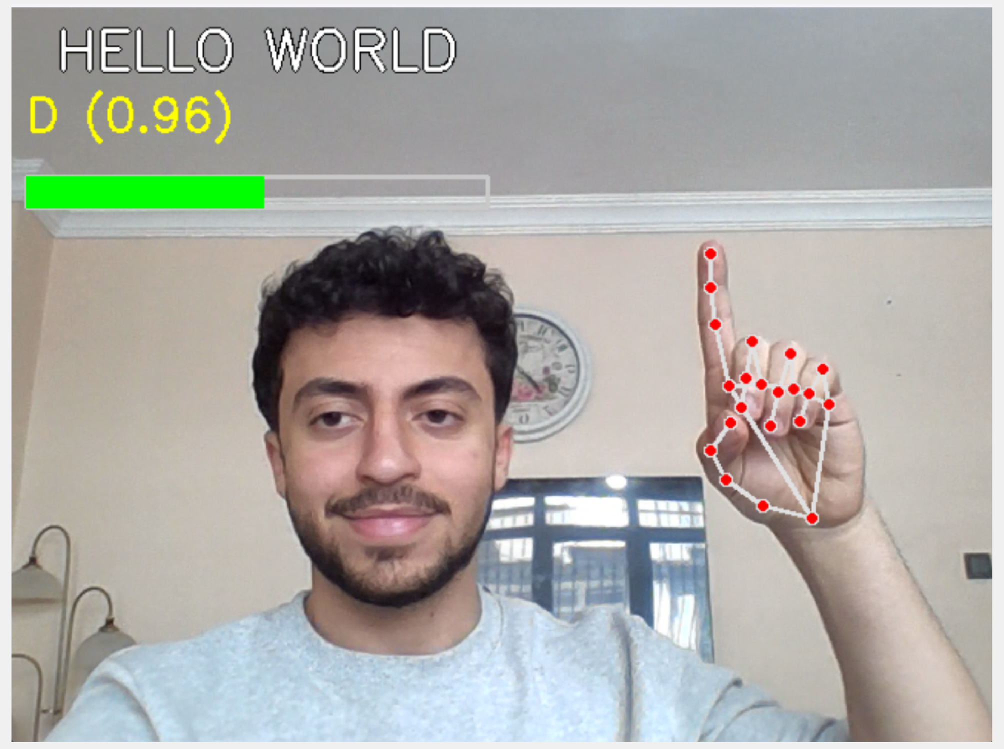
Proposed Dual-Input CNN:

- Combines grayscale image and hand landmark data.
- Achieved stable, accurate predictions in real webcam settings.

Model Performance Summary:

Metric	Score
Accuracy	0.96
Precision	0.96
Recall	0.94
F1-Score	0.94

Model Live Deployment



The system successfully tracks hand gestures in real-time using a laptop webcam.
Once a gesture is held steadily for 1.5 seconds, the corresponding letter is added to the sentence. If the hand is removed, a space is inserted and the word is spoken aloud using a text-to-speech engine.

Conclusion

I developed a real-time ASL recognition system using a dual-input CNN architecture that processes both grayscale hand images and 3D landmarks.

Compared to image-only and transfer learning baselines, our model achieved higher accuracy, better generalization, and stable real-time performance.

The integration of gesture stabilization and text-to-speech feedback makes it practical for real-world accessibility applications

Future Work

- Extend the system to recognize dynamic ASL signs and phrases using video sequences.
- Integrate Recurrent Neural Networks (RNNs) or Transformers to capture temporal patterns.
- Expand the dataset to include more users, lighting conditions, and background diversity.

References

1. P. C. Torres and C. A. Olvera, "American Sign Language Recognition with Deep Learning," *Computación y Sistemas*, vol. 24, no. 3, pp. 1211–1220, 2020.
2. O. K. Oyedotun and A. Khashman, "Deep learning in vision-based static hand gesture recognition," *Expert Systems with Applications*, vol. 122, pp. 112–134, 2018.
3. R. Mittal and G. Balakrishnan, "Hand Gesture Recognition using MediaPipe and Deep Learning," *Journal of Artificial Intelligence and Data Science*, vol. 2, no. 1, pp. 35–42, 2021.
4. F. Zhang et al., "MediaPipe Hands: On-device Real-time Hand Tracking," Google Research, 2020.
5. K. He et al., "Deep Residual Learning for Image Recognition," in *CVPR*, 2016.
6. pyttsx3 Text-to-Speech Library, https://pypi.org/project/pyttsx3