A comparison of common classifier models for hand pose estimation for the New Zealand Sign Language alphabet

Ryan B right1\*

1 Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand

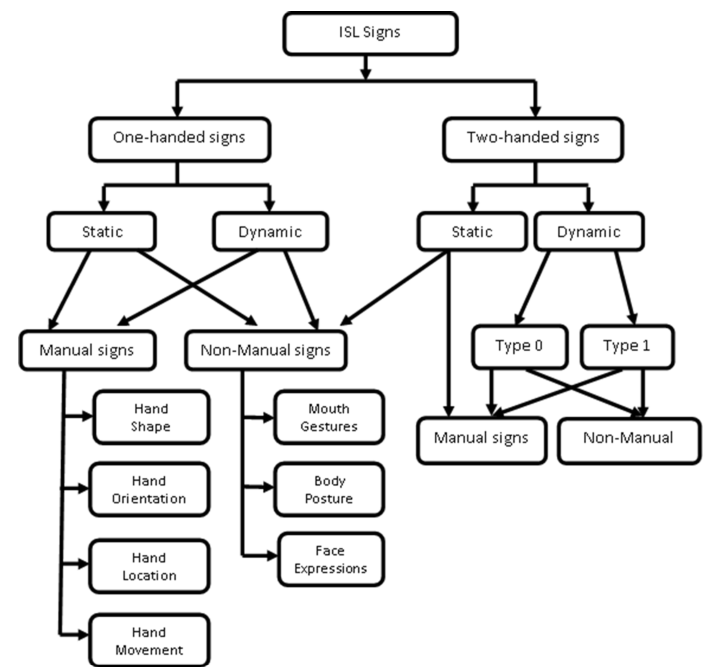
\* Corresponding Author Email: rbr115@uclive.ac.nz

|  |
| --- |
| *ABSTRACT:*  *A concise summary covering:*   * *The problem you address (e.g., NZSL gesture recognition).* * *Your approach (e.g., using MediaPipe and a neural network classifier).* * *Key results (accuracy, inference behavior).* * *Your main contribution (e.g., a live system for NZSL letter classification).* |

KEYWORDS: New Zealand Sign Language; MediaPipe Hands; Hand Pose Estimation

# Introduction

New Zealand Sign Language (NZSL) is one of New Zealand’s official languages, yet digital tools that provide accessibility and ease of social integration remain limited. There is a significant language barrier between people who rely on NZSL and those who don’t. Unlike spoken language, sign languages pose a unique set of challenges for translation tools. As shown in Figure 1, NZSL relies on precise hand positioning, movement, facial shapes, and body poses to communicate [1]. Additionally, dictionaries are challenging to use and develop due to NZSL not being a written language.



NZSL

Figure 1. Classification of sign language word forms. Adapted from [1]

This makes consistent, real-time classification a non-trivial task.

Previous efforts for other sign languages like American Sign Language (ASL) and Chinese Sign Language (CSL) show a wide range of approaches and methodologies. For example, the use of Hough transformations and CNNs [2], the use of 3D hand tracking gloves [3], and the use of various neural network models [4]. Notably, camera based methods in papers reviewed in [1], all focus on single hand signs. This poses a challenge for NZSL which has a high amount of two handed signs.

Presented is a complete pipeline including data gathering, training and evaluation with additional scripts for evaluating training data balance and form. The current system detects one or two handed static manual signs using MediaPipe to generate hand landmarks, and then a multilayer perceptron (MLP) network to classify the hand pose. The system is evaluated on validation data and real time webcam input, showing high accuracy in controlled environments and highlighting some key limitations.

# Methodology

Explain your pipeline:

* **Data Collection:** How you captured and labeled hand poses.
* **Preprocessing:** Landmark normalization, handling one/missing hands.
* **Augmentation:** Shifting, noise injection.
* **Model Architecture:** Layers, activation functions, loss function.

**Training Setup:** Dataset split, batch size, epochs, optimizer.

## Inference

The inference script uses MediaPipe and a trained MLP network to take a webcam feed and return a detected hand pose. Key steps of this pipeline are outlined in Figure 2.

Figure 2. Flow of presented inference pipeline

To get webcam video feed, the python library OpenCV-python is used

To extract spatial features of hand gestures, Google's MediaPipe Hands [5] library was selected as a starting point. It provided a lightweight, real-time hand-tracking solution capable of detecting 21 landmarks per hand, Figure 3, with relatively high accuracy. Compared to training a custom model from scratch, MediaPipe reduces the problem from the domain of camera pixels to that of hand landmark coordinates.

A hand with dots and lines

AI-generated content may be incorrect.

Figure 3. Indexed Landmarks from MediaPipe Hand’s model.

Then a MLP neural network is employed to take the landmarks and classify the hand pose.

## Data collection

Due to the lack of open-source data sets for NZSL, a data collection tool was developed.

This reduces the

# Results

Present:

* Training/validation accuracy and loss plots.
* Confusion matrices and per-class performance.
* Comparison between different models (if evaluated).
* Live inference performance and failure modes.

# Discussion

Interpret your findings:

* Accuracy vs. generalization issues.
* Effectiveness of augmentation and normalization.
* Model overfitting or underfitting signs.
* Observations about specific signs (e.g., M vs. N).

# Conclusion

Summarize:

* What your system achieves.
* Limitations (e.g., low variability in training data).
* Possible extensions (e.g., full word recognition, deployment on mobile).

# References

[1] Wadhawan A, Kumar P (2021) Sign Language Recognition Systems: A Decade Systematic Literature Review. *Arch Computat Methods Eng* **28**, 785–813.

[2] Munib Q, Habeeb M, Takruri B, Al-Malik HA (2007) American sign language (ASL) recognition based on Hough transform and neural networks. Expert Syst Appl 32(1):24–37

[3] Oz C, Leu MC (2007) Linguistic properties based on American Sign Language isolated word recognition with artificial neural networks using a sensory glove and motion tracker. Neurocomputing 70(16):2891–2901

[4] Wu J, Tian Z, Sun L, Estevez L, Jafari R (2015) Real-time American Sign Language recognition using wrist-worn motion and surface EMG sensors. In: IEEE 12th international conference on wearable and implantable body sensor networks (BSN), pp 1–6

[5] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). MediaPipe Hands: On-device real-time hand tracking. In *arXiv [cs.CV]*. http://arxiv.org/abs/2006.10214

WEBCAM FLUDITIAL MARKERS TO CORRECT FISH LENS EFFECT

Limitations and next steps:

* Overlapping hands and strong occlusion (media pipe issues)
* The rest of HOLM, movement or body position/ inclusion of pose, and facial expression
* I think that the rapid drop in val loss, and increase in training accuracy is due to there being redundant data
* A screenshot of a computer

  AI-generated content may be incorrect.

Context to cover