AMME5710 Computer Vision and Image Processing

Major Project Proposal

1. Team Member:

Yinuo Liu – 520211306 - yliu4077@uni.sydney.edu.au

Huiyu Ren – 520192902 - hren8555@uni.sydney.edu.au

Luoyi Li – 520376753 - luli0727@uni.sydney.edu.au

2. Project Content

We aim to design a real-time, high-robustness hand gesture and Auslan sign-language recognition system. The system integrates YOLO-based hand detection/segmentation, CNN-based classification, geometric and temporal modules, and depth-camera 3D sensing.

3. Methodology

(1) Data Acquisition

RGB and depth streams are captured synchronously using an Intel RealSense camera and aligned through intrinsic–extrinsic calibration. Depth data are pre-processed using bilateral filtering, hole filling, and normalization to produce a clean 2-D depth map representation that preserves geometric details while being invariant to lighting and skin color variations. Rosebag is used to get the feature.

Using different background/colour/angle/people's hand to capture the training data, testing data should be different from training data.

(2) YOLO Detection

For detection, a lightweight YOLO network is trained or fine-tuned to localize and segment hands within the RGB image. The detector outputs bounding boxes and pixel-level masks that are refined by Non-Max Suppression (NMS) to remove redundant overlaps while retaining multiple hands. The segmented hand regions are cropped and geometrically normalized for consistent scale and orientation before being passed to subsequent feature extractors.

(3) Feature Extraction

The recognition module consists of multiple parallel branches designed to extract complementary information from the hand region:

(1) a CNN-based static RGB branch (MobileNetV2 or EfficientNet) learns visual appearance features from the RGB ROI; (2) a CNN-based static depth branch processes the depth ROI to capture shape and contour information; (3) a geometric feature branch uses MediaPipe Hands to extract 21 hand landmarks and computes joint angles, bone-length ratios, and curvature descriptors, which are classified using an SVM or MLP; and (4) a temporal modeling branch (1D Temporal Convolutional Network or BiLSTM) analyzes sequential frames to recognize dynamic gestures involving hand motion.

These branches are trained separately and later combined during inference to improve recognition accuracy and robustness.

(4) Decision Fusion

If CNN confidence ≥ threshold → accept; else re-evaluate via geometric model.

Conflicting results trigger temporal verification.Weighted probability fusion: $P_{\text{final}} = \alpha P_{\text{CNN}} + \beta P_{\text{Geom}} + \gamma P_{\text{Temp}}$

(5) Evaluation and Benchmarking

The complete system will be evaluated using both quantitative and qualitative metrics, including Top-1/Top-5 accuracy, macro F1-score, and real-time inference speed (FPS). Comparative experiments will be conducted between RGB-only, depth-only, and RGB+Depth fusion settings, as well as between single-model and cascaded approaches. Robustness will be assessed under variable lighting, clothing, and background conditions, while cross-person validation will measure generalization to unseen users. These evaluations will demonstrate the effectiveness of the proposed cascaded and multimodal design for real-time Auslan gesture recognition.