MONASH University

MONASH
ENGINEERING

Gesture Recognition for Initiating Human-to-Robot Handovers

Jun Kwan, Chinkye Tan and Akansel Cosgun

GROUP OF EIGHT AUSTRALIA

Problem Statement

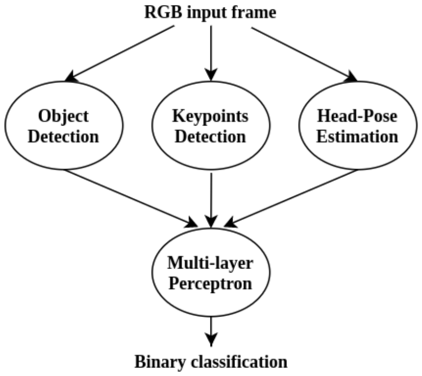
MONASH University

- Object handovers can happen in two directions
- Robot-to-Human
  - The robot delivers a requested object to a human
- Human-to-Robot
  - The robot acquires an object from a human
- Develop a system to recognize the act of the human handing an object over to the robot

Overview

MONASH University

- Extract features relevant to the task
  - Object detection
  - Keypoints detection
  - Head pose estimation

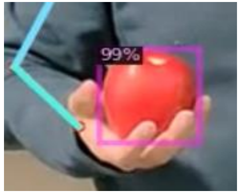


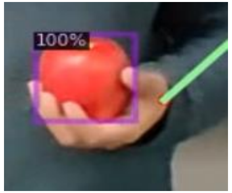
- Train a classifier
  - Detect the existence of a handover gesture based on the extracted features

Object Detection Module

- Faster R-CNN with Detectron2 engine
  - Proposed by Ren et al. [1]

- Presence and location of object extracted
  - Bounding box around the object returned if object is detected
  - x, y, width and height






[1] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99

Body Keypoints Detection Module

- Keypoints R-CNN with Detectron2 engine
  - Proposed by He et al. [2]

- Coordinates of various joints in the body extracted
  - 11 keypoints are selected (upper body)

- Centralized around the detected object
  - Object centric frame

[2] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969

Head Pose Estimation Module

- Multi-loss Resnet50 architecture
  - Proposed by Ruiz et al. [3]

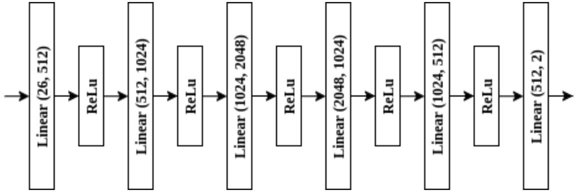
- Multiple losses are designated for different Euler angles

- Multi-task cascaded convolution networks (MTCNN) [4]
  - Face detection

[3] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” in IEEE conference on computer vision and pattern recognition workshops, 2018
[4] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016

Multi-layer Perceptron

- An input layer, four hidden layers, and an output layer



- Feature vector
  - The **1st** parameter
    - The presence of an object in the scene
  - The **2nd** to **23rd** parameters
    - Pixel coordinates of upper body keypoints
  - The **24th** to **26th** parameters
    - The yaw, pitch and roll of the head orientation

Custom Dataset

MONASH University

- We designed a custom dataset to train the multi-layer perceptron

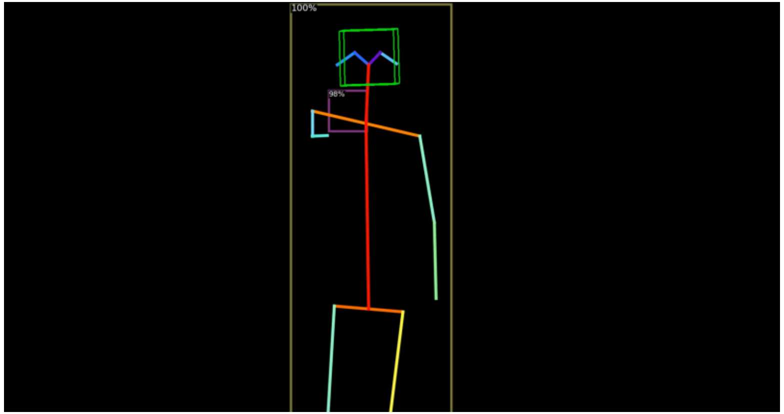
- A total of 25 videos were recorded in various environments
  - Containing a total of 2506 images

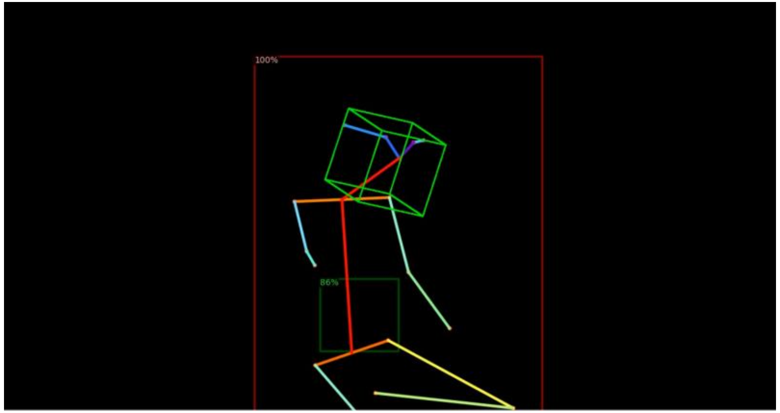
- Each image was labelled ‘1’ denoting a handover scenario or ‘0’ otherwise

Custom Dataset

Skeleton Images

- ResNet50 used instead of Multi-layer Perceptron
- Features are placed on a black image and fed into a CNN





End-to-End Method

- Standard CNN used as a baseline
- Alexnet [5] and ResNet50 [6] used
- Raw RGB images used as input to the CNN

[5] I. S. Alex Krizhevsky and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in International Conference on Neural Information Processing Systems, 2012
[6] S. R. Kaiming He, Xiangyu Zhang and J. Sun, “Deep residual learning for image recognition,” arXiv preprint arXiv:1512.03385, 2015

Results

Methods	**Accuracy (%)**
End-to-end (Alexnet)	50.0
End-to-end (Resnet50)	89.4
CNN on skeleton images	83.3
MLP (absolute pixels)	90.1
MLP (relative to object)	**90.6**

Video Demonstration

Discussion

- The system with object centric frame is more robust
  - Absolute position of human no longer taken into account

- MLP system outperforms skeleton image CNN system
  - MLP receives features directly
  - CNN has to decipher features from the skeleton images

Future Work

MONASH University

- Temporal information to be included
- Use of other communication cues, e.g. verbal and gaze
- Dataset to be more robust
- Ablation study of each module

MONASH University

MONASH
ENGINEERING

Thank You

GROUP OF EIGHT AUSTRALIA