Detecting Object Manipulations in an Assembly Task

Jacob Rosensköld

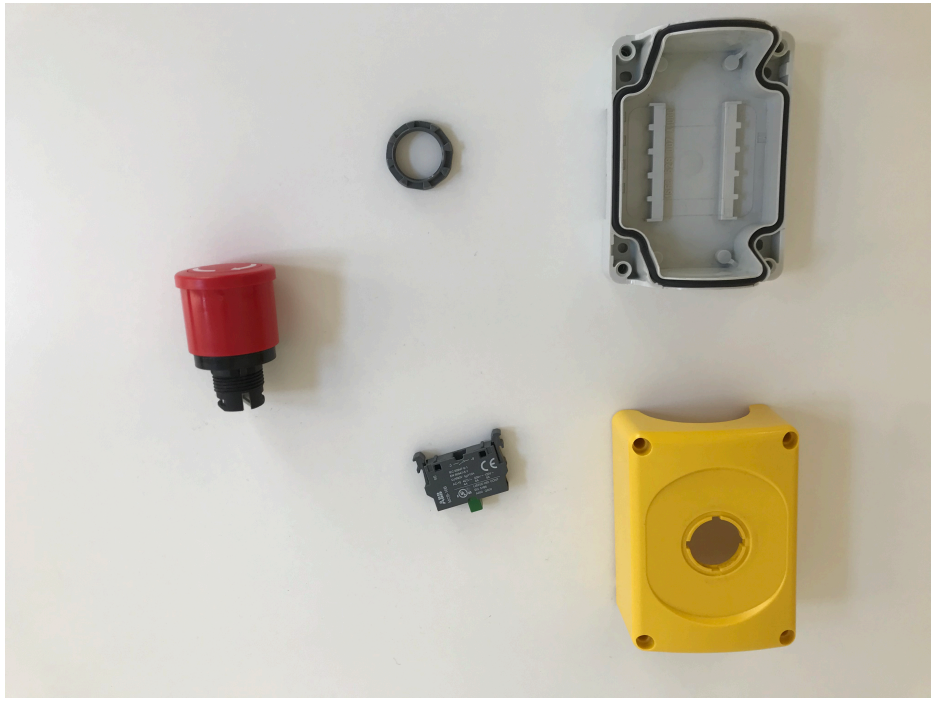
mas15jro@student.lu.se

Agenda

- Introduction
- Methodology
- YOLO – Object detection
- OpenPose – Pose estimation
- Hand activity recognition
- Determine which object is being manipulated
- Discussion
- Future work

Introduction

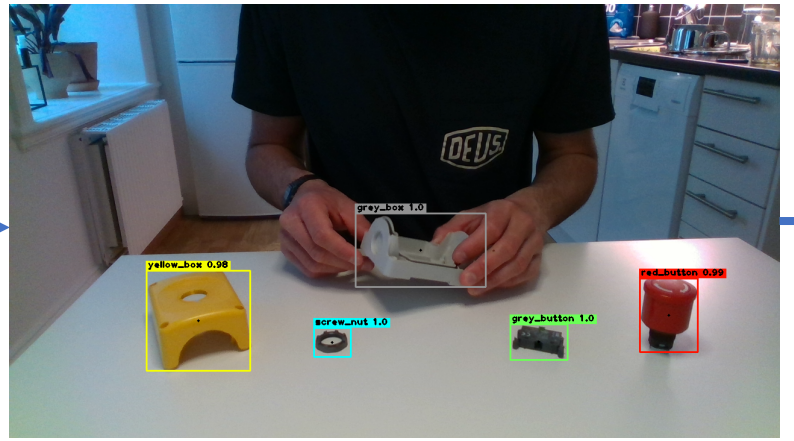
- Detecting object manipulations
- Extract information from an assembly video



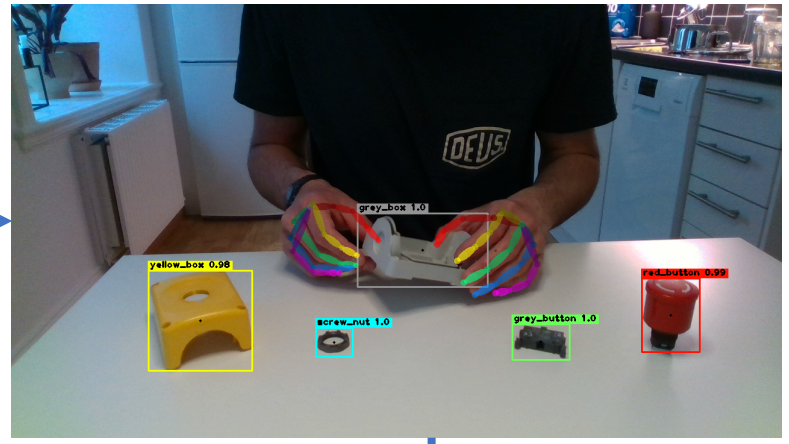
Methodology

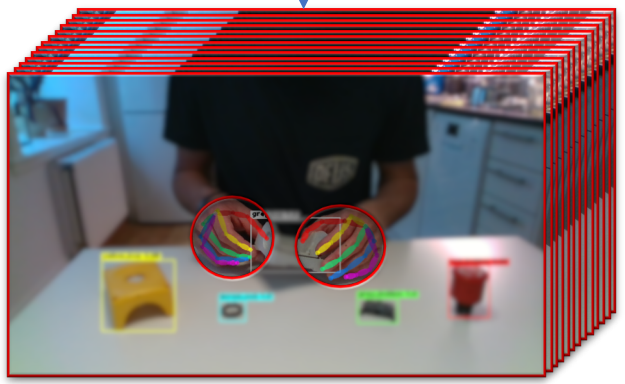
Video frame

1. Object detection



2. Pose estimation





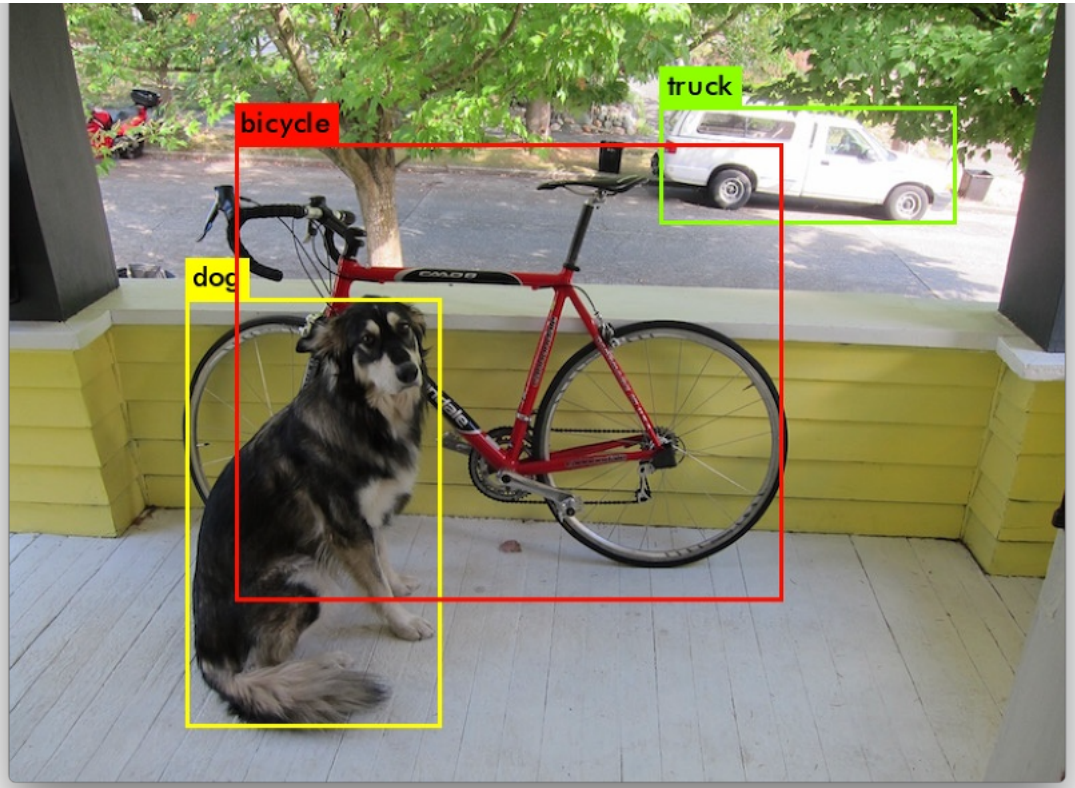
Sequence/sliding window of hand key points from 32 latest frames

3. Hand activity detection

4. Using depth camera to determine which object is being manipulated

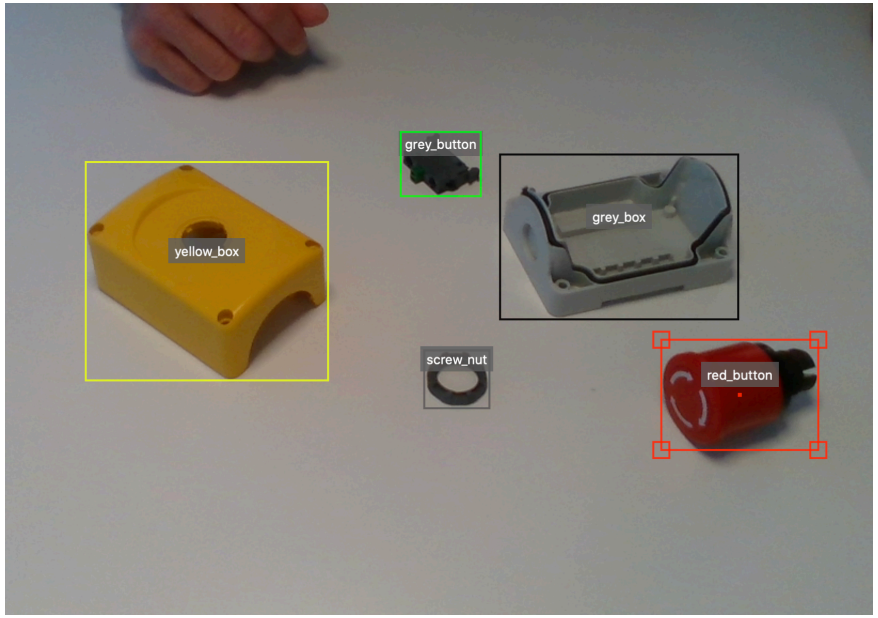
1. YOLO – Object detection

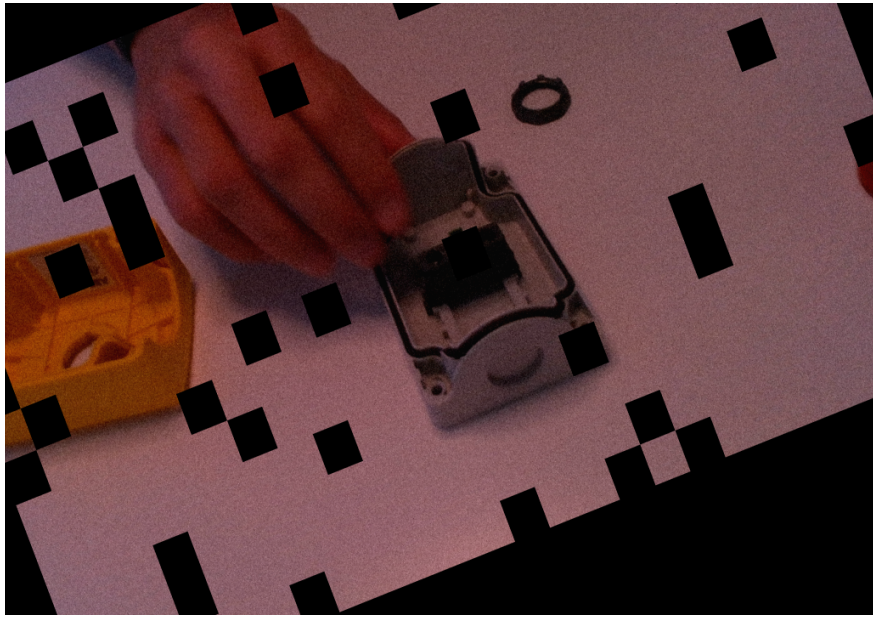
- ”You only look once”



1. YOLO - training

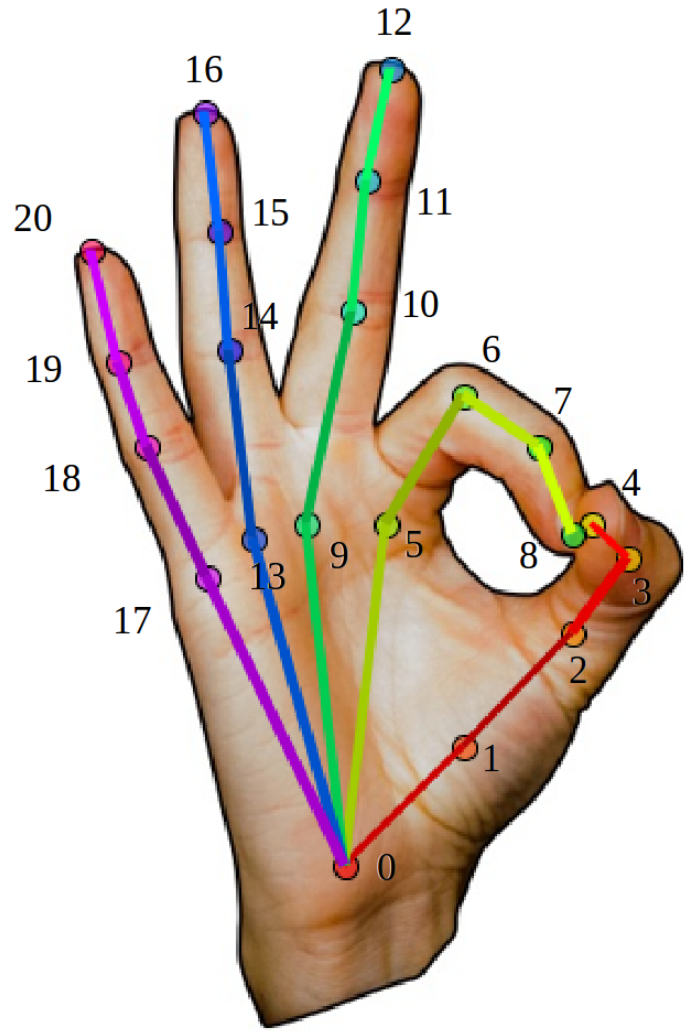
- About 1000 annotated images.
- Image augmentation, increased to about 6000 images.
- Mean average precision of 99%.





2. OpenPose – Pose estimation

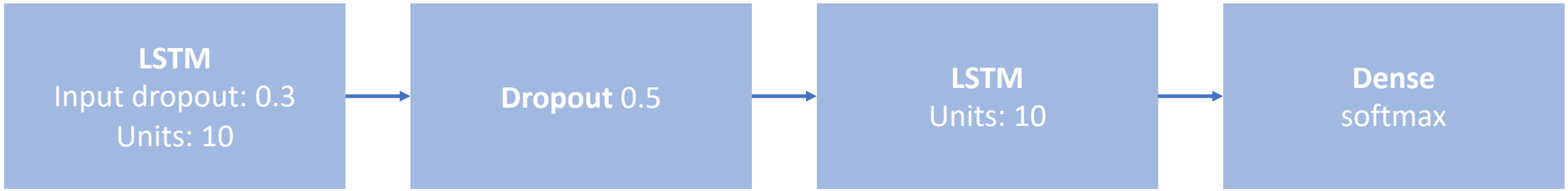
- Estimate body, foot, face and **hand** key points.
- 21 key points per hand.



3. Hand activity recognition

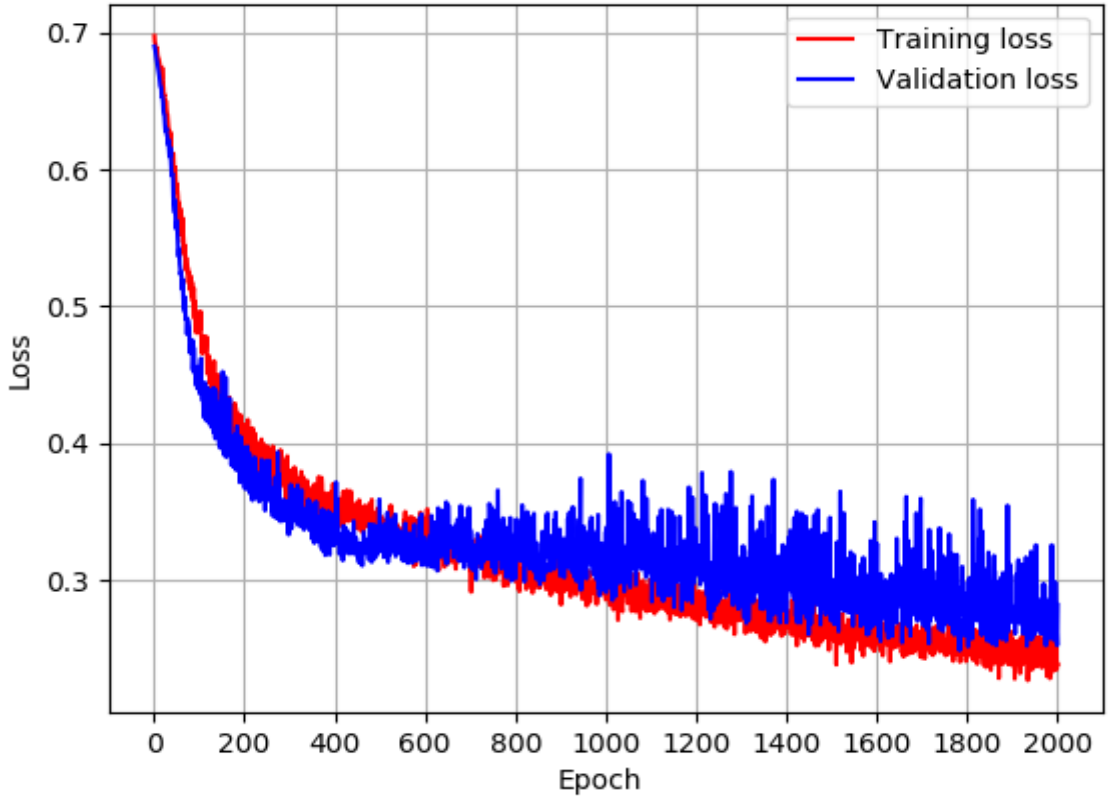
- Classify sequences of the 32 frames.
  - Each frame containing hand key points from OpenPose.
  - Only used the 5 fingertips.
- Two classes: grip and drop

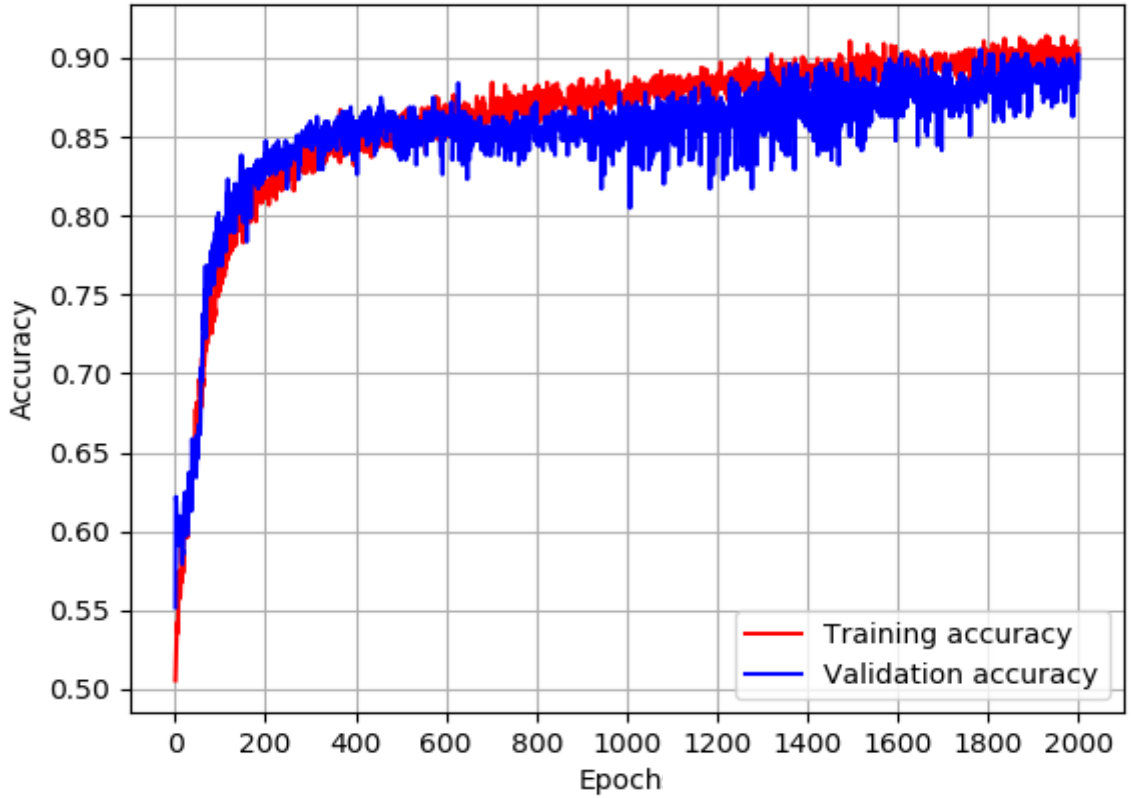
3. Hand activity recognition - architecture



3. Hand activity recognition - training

- About 1300 sequences, 70% training, 15% validation, 15% testing
- 2000 epochs, batch size of 256





3. Hand activity recognition - results

Class	Precision	Recall	F1-score
Grip	0.89	0.95	0.92
Drop	0.94	0.87	0.91
Average	0.92	0.91	0.91

Demo 1

4. Determine which object is being manipulated

- Depth camera.
- Look for sufficient close objects.
- Could add more, e.g. looking if object moves in same direction as hand.

Demo 2

Discussion

- Does not work that good when tested on video containing a lot of activities and movements.
  - Detects activities with high confidence even though the person is just moving his hands around.
  - Grip and drop are very subtle activities.

Discussion

- Incorrect classifications.
  - Grip and drop are similar.
  - Many of the frames in the sequences for both grip and drop are just the hands moving.

Discussion

- The system is made of a number of chained sub systems.
  - Every sub system adds some uncertainty to the final result.

Discussion

- It’s possible to detect subtle movements in 2D, even though improvements are required.

Future work

- Train the neural network not using fixed-sized frame sequences.
  - Stateful LSTM.
- Include depth of the hand key points in the data.
  - Would make it easier to detect subtle movements.

Future work

- Add more classes.
  - E.g. a screwing motion.
  - Would give a hint of how much the similarity between grip and drop affects the results.

Thank you!

Questions?