Hearing the Silent: Using Deep Learning to Interpret American Sign Language

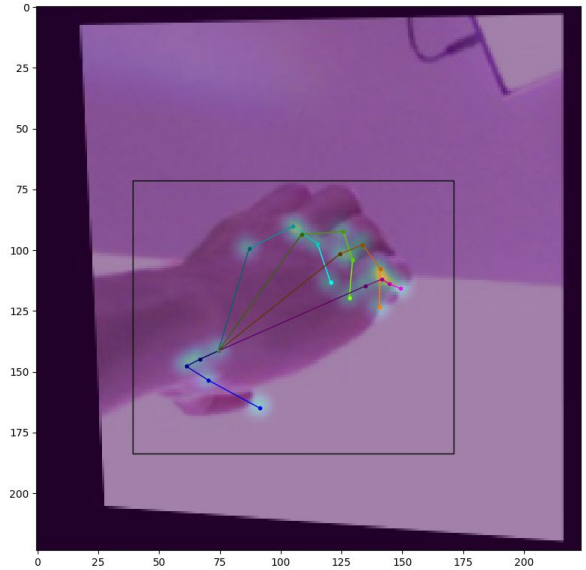
Christoffer Dyrssen

Project Goal

- Utilize publicly available datasets for the development and training of a deep learning model to interpret American Sign Language (ASL)
- Existing datasets include features such as the following:
  - 2D/3D hand joint positions
  - Depth maps
  - Foreground/background masks
  - Corresponding RGB image data
- To create a more robust model, 3D hand pose estimation is accomplished through combining depth and 2D joint position predictions.

2D Joint Position Estimation

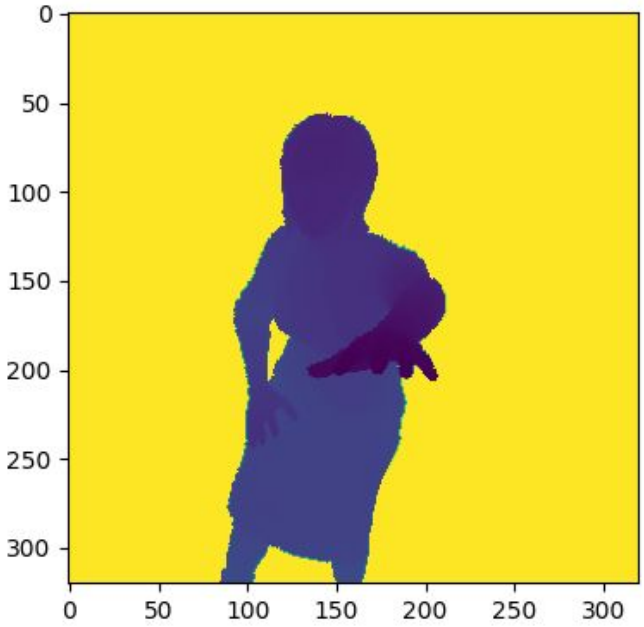
- Encoder/decoder neural network architecture with skip connections for a multiscale analysis
- Target variables:
  - Bounding box center
  - Ratio of bounding box width/height compared to the original image width/height
  - 21 hand keypoints (joint positions)
- Point heatmaps are learned instead of directly regressing towards keypoint positions



Freihand Dataset: https://lmb.informatik.uni-freiburg.de/projects/freihand/

Depth Estimation

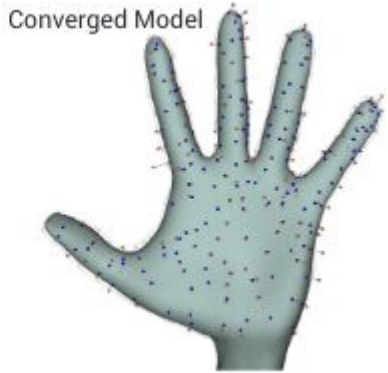
- Encoder/decoder neural network architecture with skip connections for a multiscale analysis
- Input data will be masked to represent hand pixels only and remove background noise during training
- Target variable:
  - Relative depth



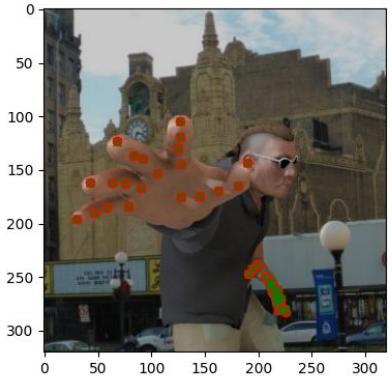
Rendered Handpose Dataset: https://lmb.informatik.uni-freiburg.de/resources/datasets/RenderedHandposeDataset.en.html

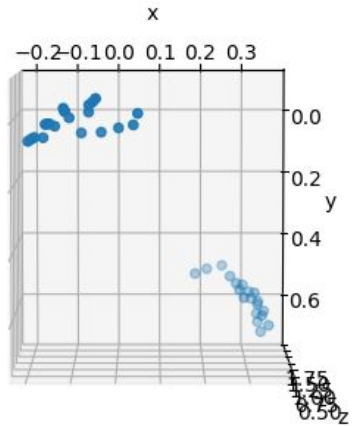
3D Pose Estimation

- Bayesian model
- A set of 3D points are sampled from the predicted depth map and an initial pose is created based on the 2D hand keypoint predictions
- A 3D hand model is then generated to minimize an energy function with the following constraints:
  - Distance from subsampled 3D points to hand model is minimized
  - Joint keypoints are in a physically viable and statistically probable configuration



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/SIGGRAPH2016-SmoothHandTracking.pdf





Rendered Handpose Dataset:
https://lmb.informatik.uni-freiburg.de/resources/datasets/RenderedHandposeDataset.en.html

Future Work

http://www.cedus.it/documents/SicurezzaUrbana/Videosorveglianza_e_altre_tecnologie_video_di_controllo/3ZChi_ACV-1.pdf

- Hand pose estimation is only one part of interpreting ASL
- Other steps required:
  - Track hand positions relative to subject's head/body to capture gestures over time
  - Facial expression classification since ASL interpretation often times relies on reading expressed emotions
  - Creation of transcribed ASL video dataset
- Overall pipeline would utilize a recurrent neural network trained on 3D hand poses, hand tracking paths, and facial expressions for each frame of a video feed

http://people.csail.mit.edu/sybor/research_abstract/sybor_2006.html

https://www.youtube.com/user/billvicars/featured