How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language

Paper

Contains:

1. Green screen studio RGB videos
2. Green screen studio RGB-D videos
3. Body-face-hands 2D keypoints (Multiple views)
4. Panoptic studio data (Subset of dataset containing multi-view VGA/HD videos and 3D keypoint estimation)

The RGB-D videos, 2D keypoints, and 3D keypoints, can each be used to form frames of 3D keypoints.

1. 2D pose estimation can be applied to the RGB-D videos and the depth values can be used to lift the 2D keypoints to 3D.
2. The 2D keypoints can be retrieved from the front-facing 2D keypoint video and the side-facing 2D keypoint video can be used to determine the depth values.
3. The 3D keypoints are already provided for a subset of the dataset.

The dataset of 3D keypoint frames can be split by ASL sign and a transformer-based model can be designed to learn the ASL language, converting ASL Gloss (or English) tokens into frames of ASL signs as 3D pose keypoints.