Multimodal Hand Pose Enhancement for Sign Language

Alvaro Budria
alvaro.francesc.budria@estudiantat.upc.edu

Advisors: Laia Tarrés & Xavier Giró

Introduction to Research (I2RCED)
GCED 2021 - 2022

Outline

1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

Intro - Sign Language

Sign languages (SL) spoken by >450 million people

Texts has a discrete representation
SL is continuous

Lack of support for SL in communication technologies

Intro - SL processing

Lately, some promising attempts at SL processing through ML and DL

DL systems benefit from large amounts of data

Deep Learning 

New SL datasets, such as How2Sign

Outline

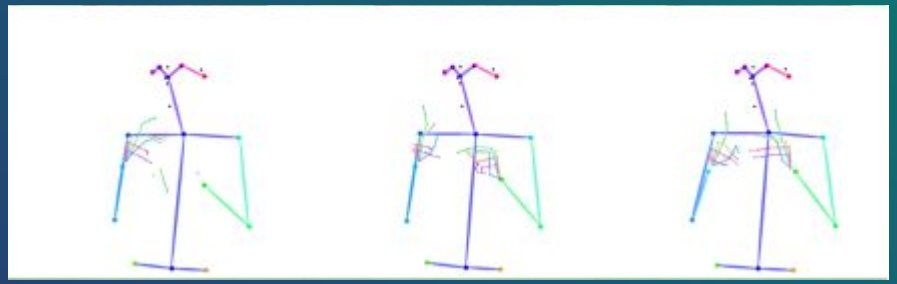
1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

How2Sign

Large-scale collection of multimodal sign language videos in American Sign Language (ASL)

Sentence-level alignment for >35k sentences

Annotations include **category labels, text annotations,** and automatically extracted **2D keypoints**

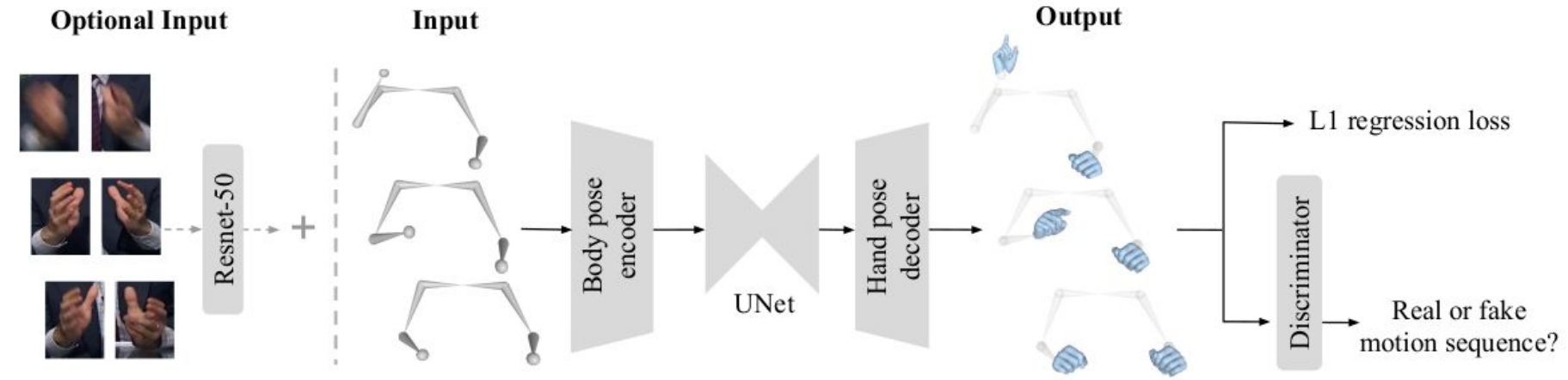


Duarte et al. *How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language.* CVPR 2021

Outline

1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

Body2Hands



Generative Adversarial Network (GAN)

Fully 1d-convolutional encoder-decoder generator

Image feature vector as optional input

Evonne Ng et al. *Body2Hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics*. CVPR 2021

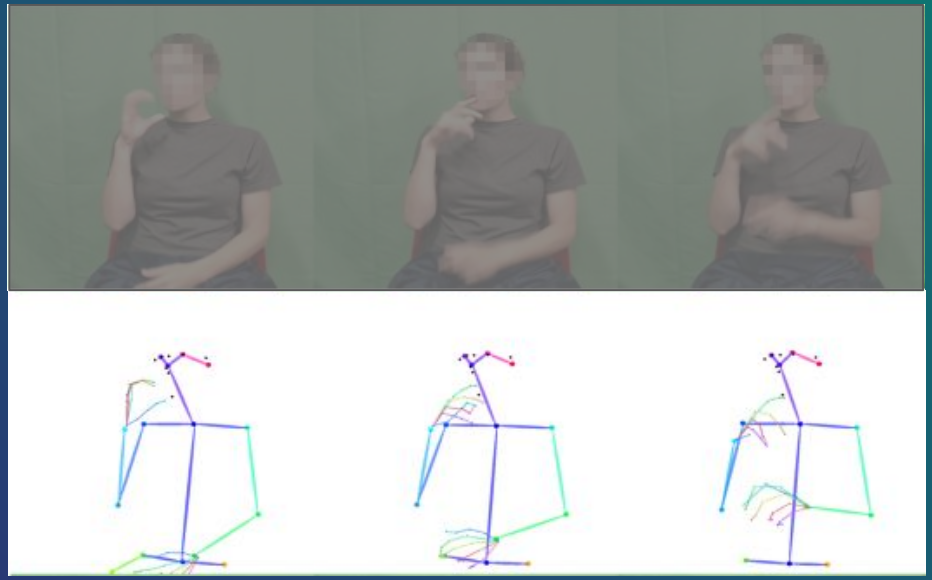
Outline

1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

Data Representation

SL does not admit a discrete representation, but...
directly processing video is too time-consuming and resource intensive

SL does not admit a discrete representation, but...
directly processing video is too time-consuming and resource intensive

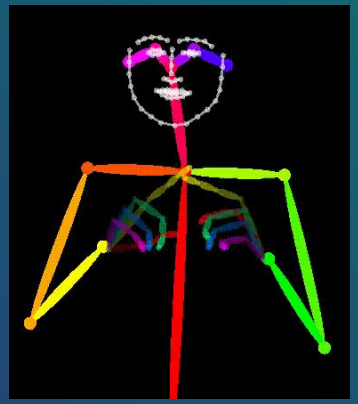


Duarte et al. *How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language.* CVPR 2021

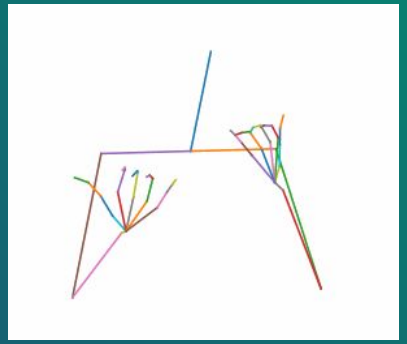
...so keypoints are extracted, making systems more robust to changes in background and variability among signers

RGB video

OpenPose



2D to 3D lifting

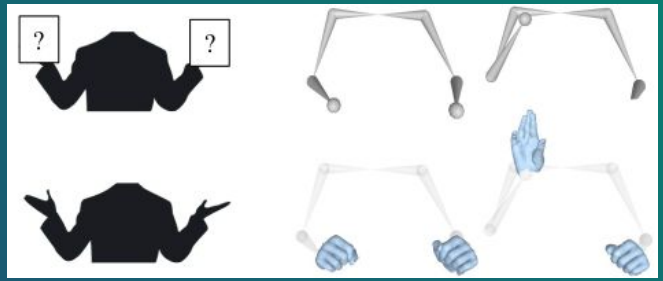


$v_t = (R_0, G_0, B_0, \ldots, R_{255}, G_{255}, B_{255})$

$v_t = (x_0, y_0, \ldots, x_{49}, y_{49})$

$v_t = (x_0, y_0, z_0, \ldots, x_{49}, y_{49}, z_{49})$

Body2Hands
(conversational settings)



Body2Hands
(sign language)

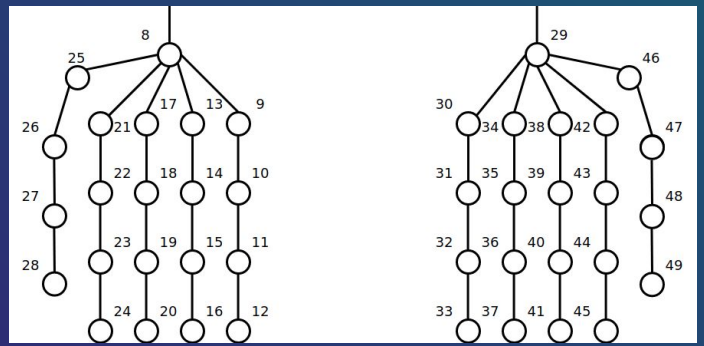
Evonne Ng et al. *Body2Hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics*. CVPR 2021

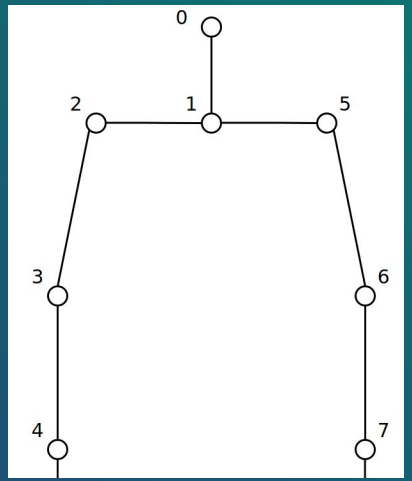
Skeletal Model & Data Representation

2D keypoints not suitable → not robust against occlusions and changes in angle

2D KPs are lifted to 3D

We defined a kinematic tree





Zelinka et al. *NN-Based Czech Sign Language Synthesis.* 2019

Cartesian to Rotational Representation

3D KPs not robust against changes in scale and length of the speaker's limbs

Rotational representation solves this problem

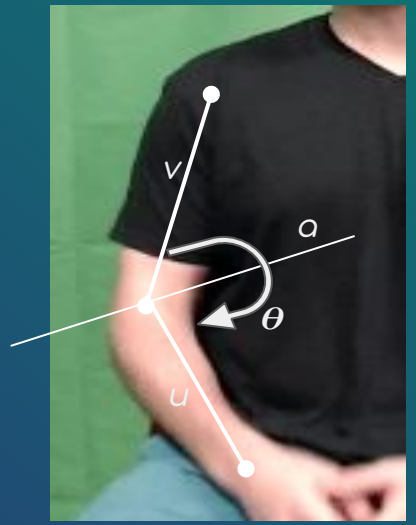
3D (Cartesian) → axis-angle (rotational) → R6D (rotational)

Pseudo-code

for each bone iB, traversing the kinematic tree:
    $u \leftarrow$ bone iB;
    $v \leftarrow$ its *parent* bone

    # compute rotation angle w.r.t to parent
    $\theta \leftarrow \arccos[\,(u \cdot v) \,/\, (\,||u||\;||v||\,)\,]$

    # compute rotation axis w.r.t. to parent
    $a \leftarrow (u \times v) \,/\, (\,||u||\;||v||\,)$



Outline

1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

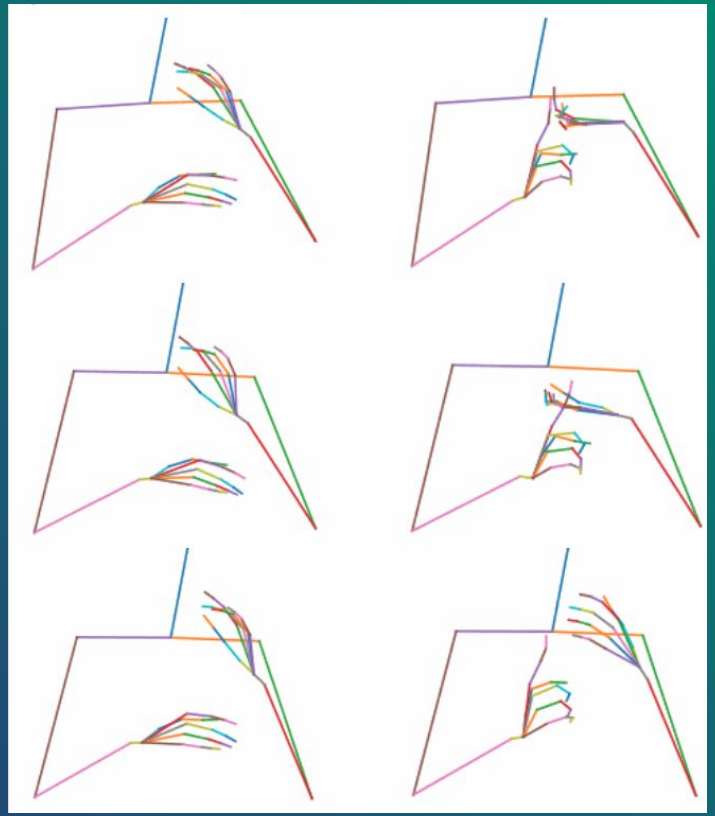
Transferring Body2Hands to SL

Directly trained Body2Hands model on How2Sign sign language data

Tried conditioning on textual sentence embedding

Tried conditioning on images features of the hands

Body2Hands cannot be directly applied to SL



Left: generated poses Right: ground-truth poses

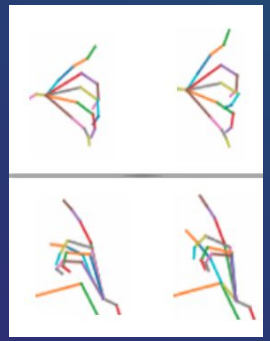
Outline

1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

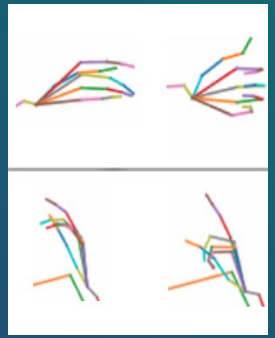
Finger Unmasking

Input: arms as well as hands, while masking out some of the fingers.
The model must reconstruct the masked fingers

The more masked fingers, the less varied are the reconstructions



Left: hand with 1 reconstructed finger
Right: ground-truth poses



Left: hand with 5 reconstructed fingers
Right: ground-truth poses

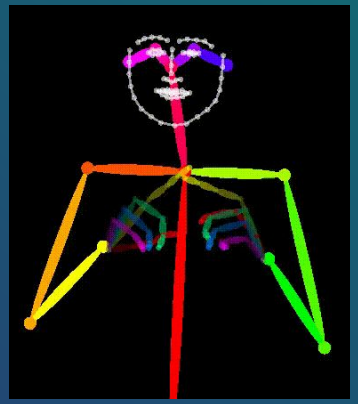
Outline

1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

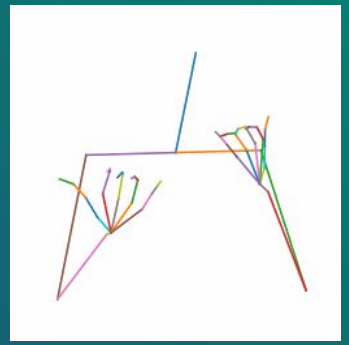
Hand Pose Enhancement

RGB video

OpenPose



2D to 3D lifting



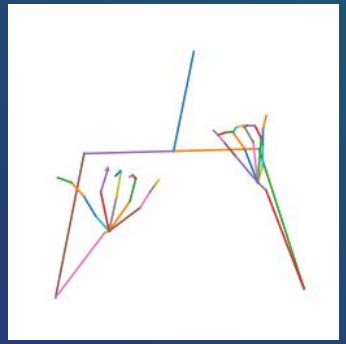
$v_t = (R_0, G_0, B_0, \ldots, R_{255}, G_{255}, B_{255})$

$v_t = (x_0, y_0, \ldots, x_{49}, y_{49})$

$v_t = (x_0, y_0, z_0, \ldots, x_{49}, y_{49}, z_{49})$

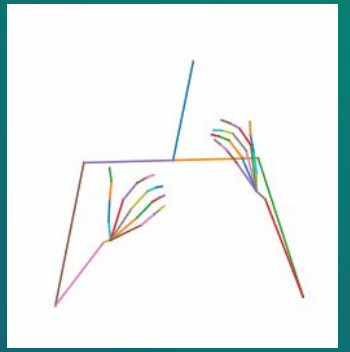
Hand Pose Enhancement

2D to 3D lifting



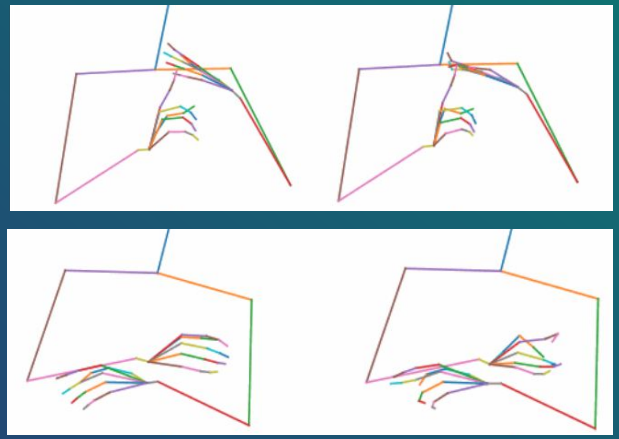
$v_t = (x_0, y_0, z_0, \ldots, x_{49}, y_{49}, z_{49})$

Body2Hands



$v_t = (x_0, y_0, z_0, \ldots, x_{49}, y_{49}, z_{49})$

Qualitative evaluation of the results



Left: enhanced hand poses

Right: original hand poses

Unable to evaluate quantitatively →

evaluate against surrogate task: topic detection

Outline

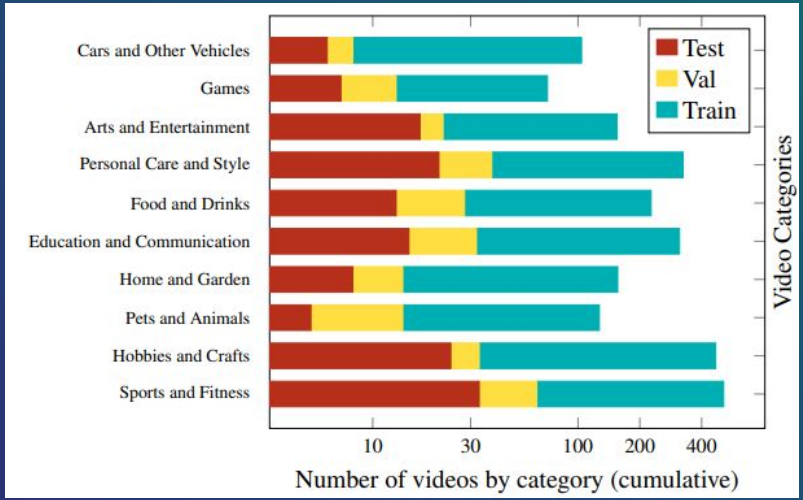
1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

Topic Detection

Motivation: provide a more quantitative evaluation of the method

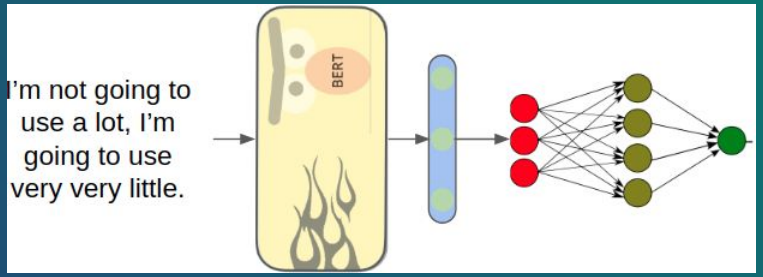
By default, data in How2Sign dataset is at *sentence level*

Group sentence-level data based on its video of origin → data at video level

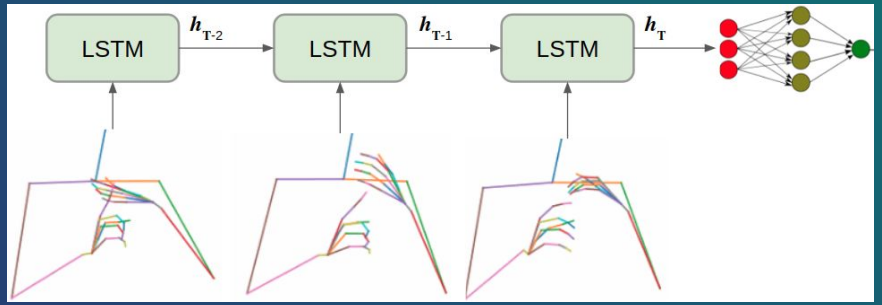


Topic Detection - Modeling Setup

Textual data



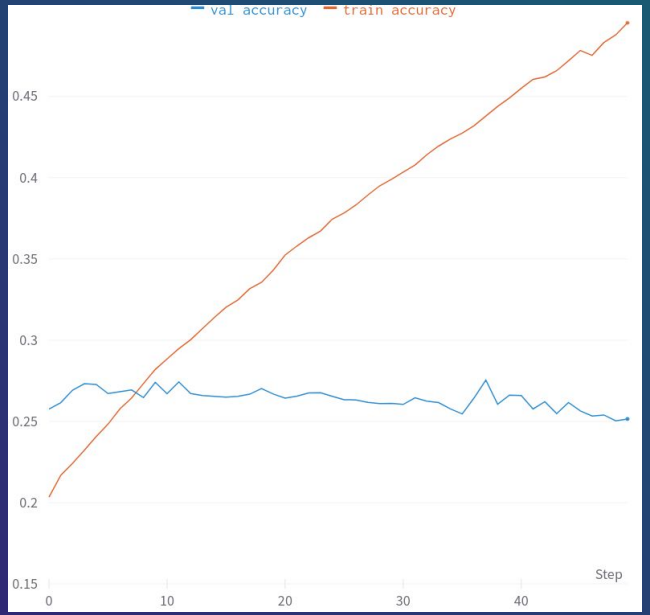
SL data



Topic Detection

1. determine if sentence level data is rich enough to classify

Train a classifier on textual sentence-level data

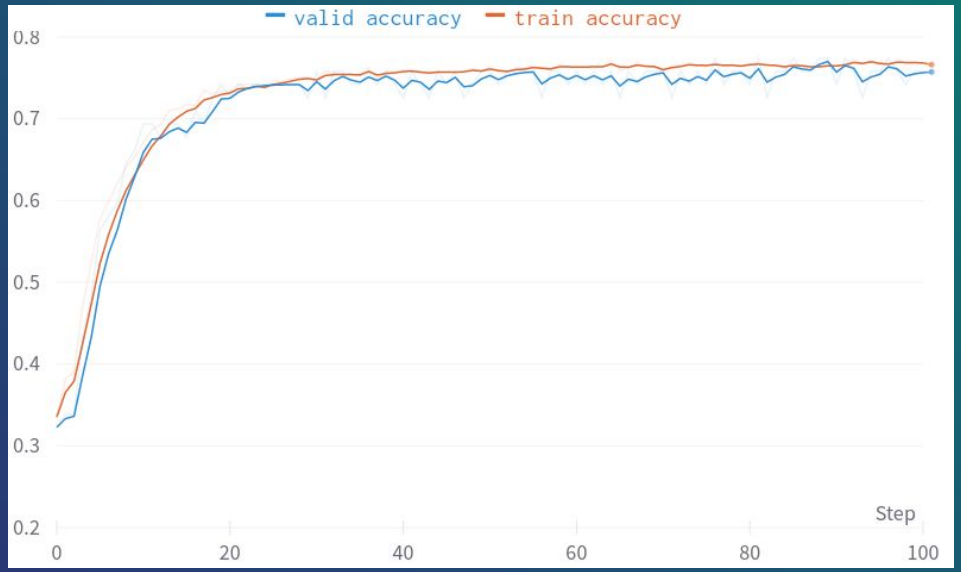


Conclusion:
not enough information in a single sentence

Therefore we group sentence level data into video level data

Topic Detection

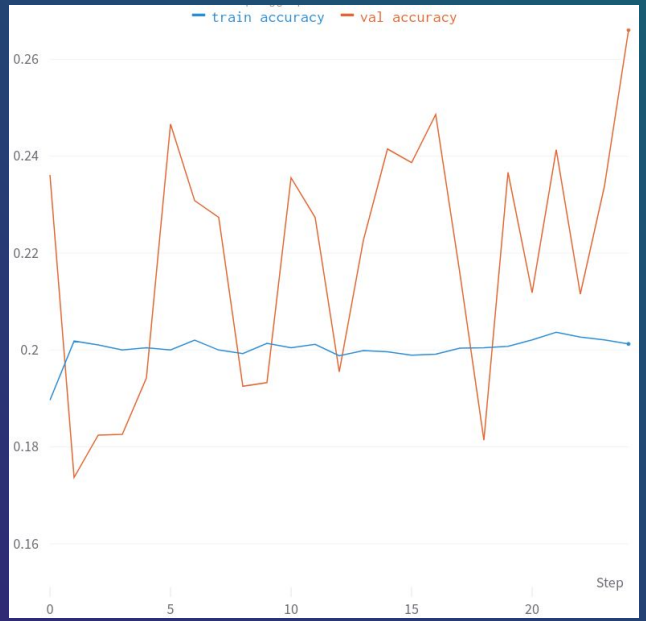
2. check if topic detection can be solved with textual data



After hyperparameter tuning, validation accuracy of 77% → topic detection task is solvable with video level data

Topic Detection

3. determine if 2D data is suitable for topic detection

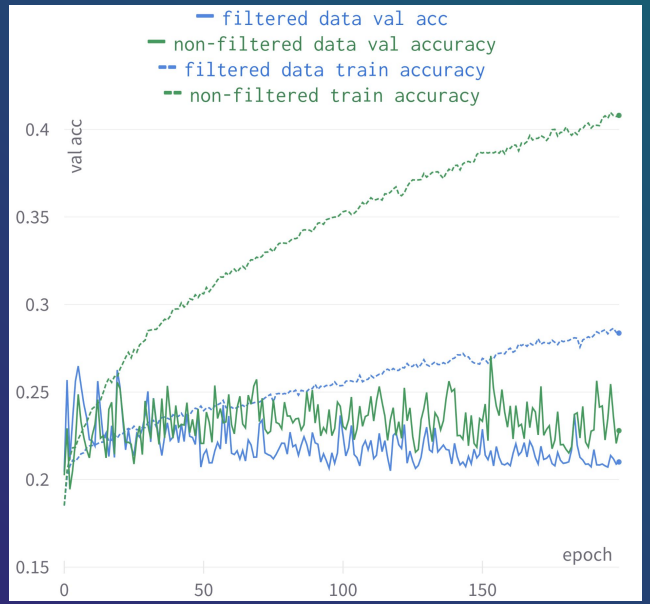


Train a classifier on SL 2D video-level data

With 2D keypoints, we cannot tackle the topic detection task

Topic Detection

4. compare classification with non-enhanced hand-poses vs. enhanced



No substantial improvement in terms of validation accuracy

Filtered poses seem to prevent over-fitting

Outline

1. Introduction
2. How2Sign
3. Body2Hands
4. Data Representation
5. Transferring Body2Hands to SL
6. Finger Unmasking
7. Hand Pose Enhancement
8. Topic Detection
9. Recap & Discussion

Recap & Discussion

We focus on Body2Hands, initially proposed for conversational settings, and transfer it to SL

Body2Hands is not adequate as an off-the-shelf method for our purposes

We encounter the limits of Body2Hands for generating SL hand poses

We show qualitatively promising results in hand pose enhancement

No gain was made on topic detection with enhanced hand poses → a more specific architecture such as SL transformers is in order

Thank you

Alvaro Budria

alvaro.francesc.budria@estudiantat.upc.edu

Advisors: Laia Tarrés & Xavier Giró

Introduction to Research (I2RCED)

GCED 2021 - 2022