Extracting sign language articulation from videos with MediaPipe

Carl “Calle” Börstell

UNIVERSITETET I BERGEN

Sign languages

- **Sign languages** are natural, full-fledged human languages
  - ~ **200 different languages documented** (so far)

- Signs have a **place of articulation** (location relative to the body), a **dominant hand** (left or right) and can be either **one- or two-handed**

Sign languages

- **Sign languages** are natural, full-fledged human languages
  - ~ **200 different languages documented** (so far)

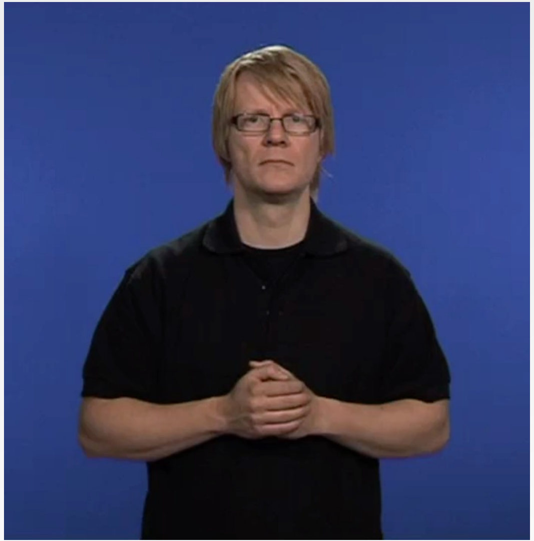
- Signs have a **place of articulation** (location relative to the body), a **dominant hand** (left or right) and can be either **one- or two-handed**

TAXI

Sign languages

- **Sign languages** are natural, full-fledged human languages
  - ~ **200 different languages documented** (so far)

- Signs have a **place of articulation** (location relative to the body), a **dominant hand** (left or right) and can be either **one- or two-handed**

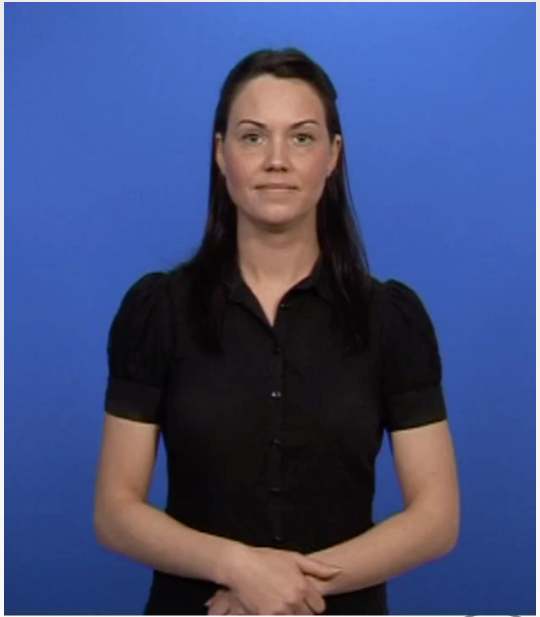


TO

Sign languages

- **Sign languages** are natural, full-fledged human languages
  - ~ **200 different languages documented** (so far)

- Signs have a **place of articulation** (location relative to the body), a **dominant hand** (left or right) and can be either **one- or two-handed**



DINNER

However, we don’t have reliable methods for automatically extracting or classifying phonological data!

Automatic extraction of “gesturing”

- **Motion Capture**
- **3D cameras** (e.g. Kinect)

Automatic extraction of “gesturing”

- **Motion Capture**
- **3D cameras** (e.g. Kinect)
  - Requires **special** hardware and (**proprietary**) software
  - Part of recording session (**pre**-planning)

Automatic extraction of “gesturing”

- **Motion Capture**
- **3D cameras** (e.g. Kinect)
  - Requires **special** hardware and (**proprietary**) software
  - Part of recording session (**pre**-planning)

- **Computer vision models**
  - Requires **general** hardware and **free** software
  - Can be done in post-processing (**after** recording)

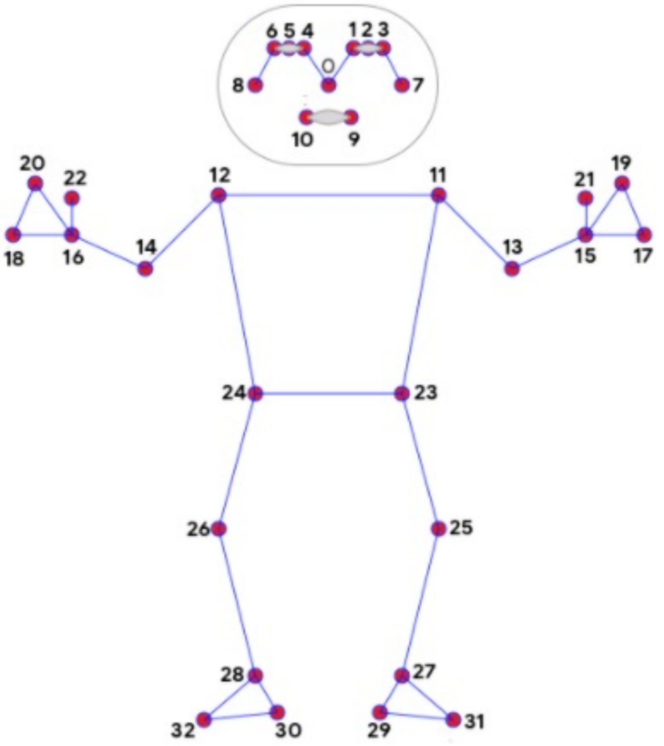
MediaPipe (by Google)

- **Free software** with many implementations (e.g. **Python**)

- **Pre-trained model** that recognizes human location/movement in video

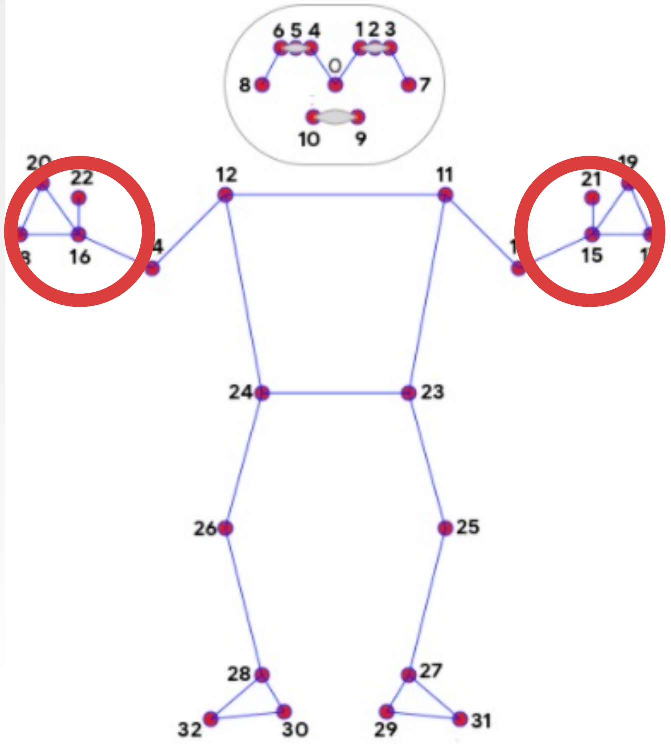
- Can be used with detailed models estimating face and finger landmarks, or more basic **body pose estimation**

MediaPipe output



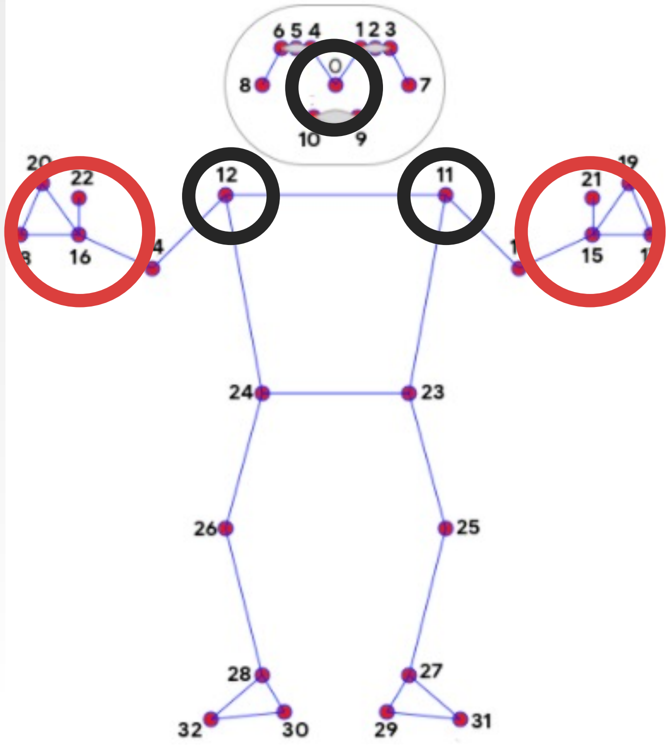
0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

MediaPipe output



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

MediaPipe output



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

My goals

- What **form information can be extracted** with MediaPipe?
  - The **articulation phase** of (STS) signs
  - The **dominant hand** (left or right)
  - The **number of hands** (one- or two-handed)
  - The main **place of articulation**

My goals

- What **form information can be extracted** with MediaPipe?

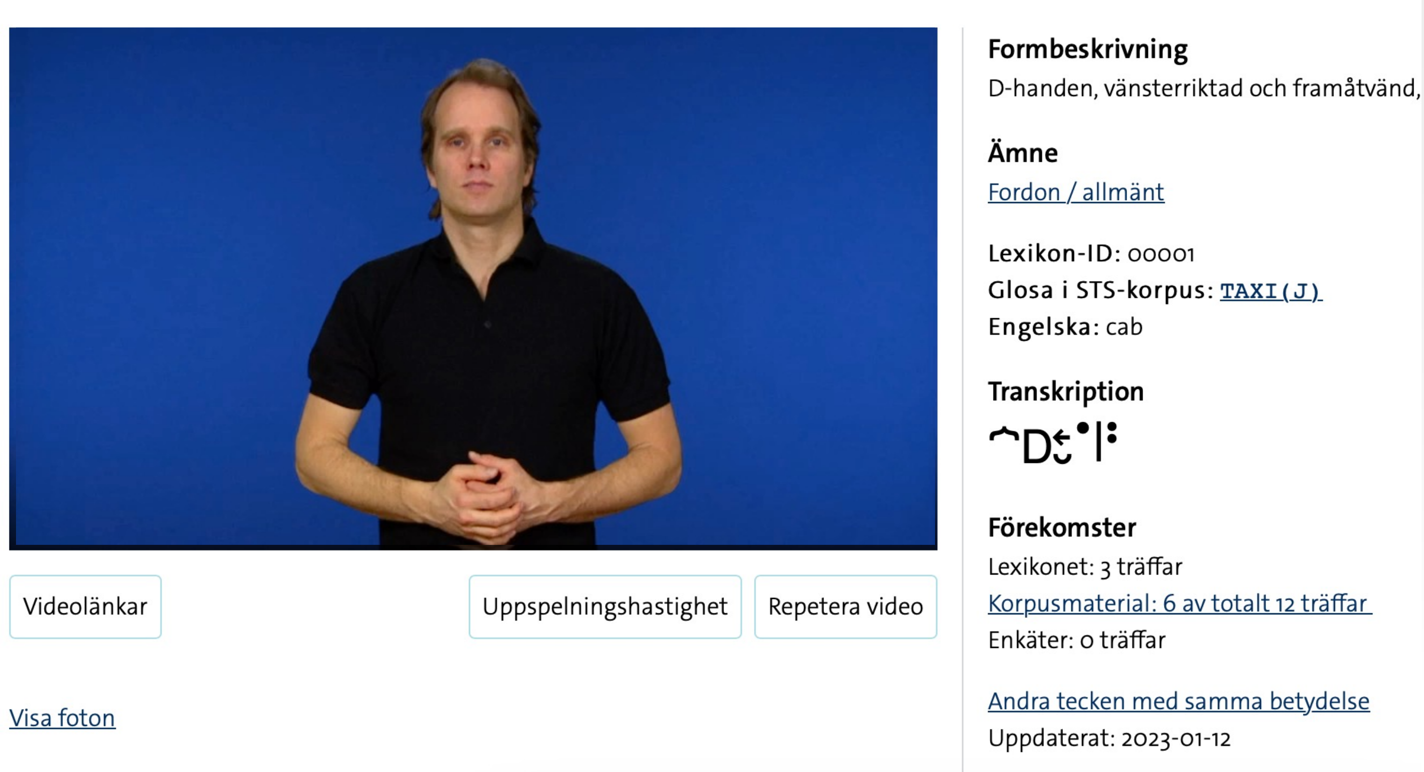
Could potentially lead to a quick (but dirty) way of annotating existing datasets of sign language videos

  - The **number of hands** (one- or two-handed)
  - The main **place of articulation**

STS dictionary

- >20,000 sign **videos**
- >40 different **signers** in the videos (mostly right-handed)
- An extensive **lexical database** behind it **describing form and meaning** (and linking with the corpus)
- Thomas Björkstrand is the manager and he provided data about the signs and signers for this study (thanks!)

STS dictionary: sign TAXI



Visa foton

Formbeskrivning

D-handen, vänsterriktad och framåtvänd,

Ämne

Fordon / allmänt

Lexikon-ID: 00001
Glosa i STS-korpus: TAXI(J)
Engelska: cab

Transkription

⏜D←◡•|:

Förekomster

Lexikonet: 3 träffar
Korpusmaterial: 6 av totalt 12 träffar
Enkäter: 0 träffar

Andra tecken med samma betydelse
Uppdaterat: 2023-01-12

The sample of signs

- 1,292 sign videos that were
  - **Non-compounds**
  - Represented a **diverse** set of **signers** (**handedness**)
  - Represented diverse **locations**

- These were all downloaded from the dictionary and then **processed with MediaPipe**

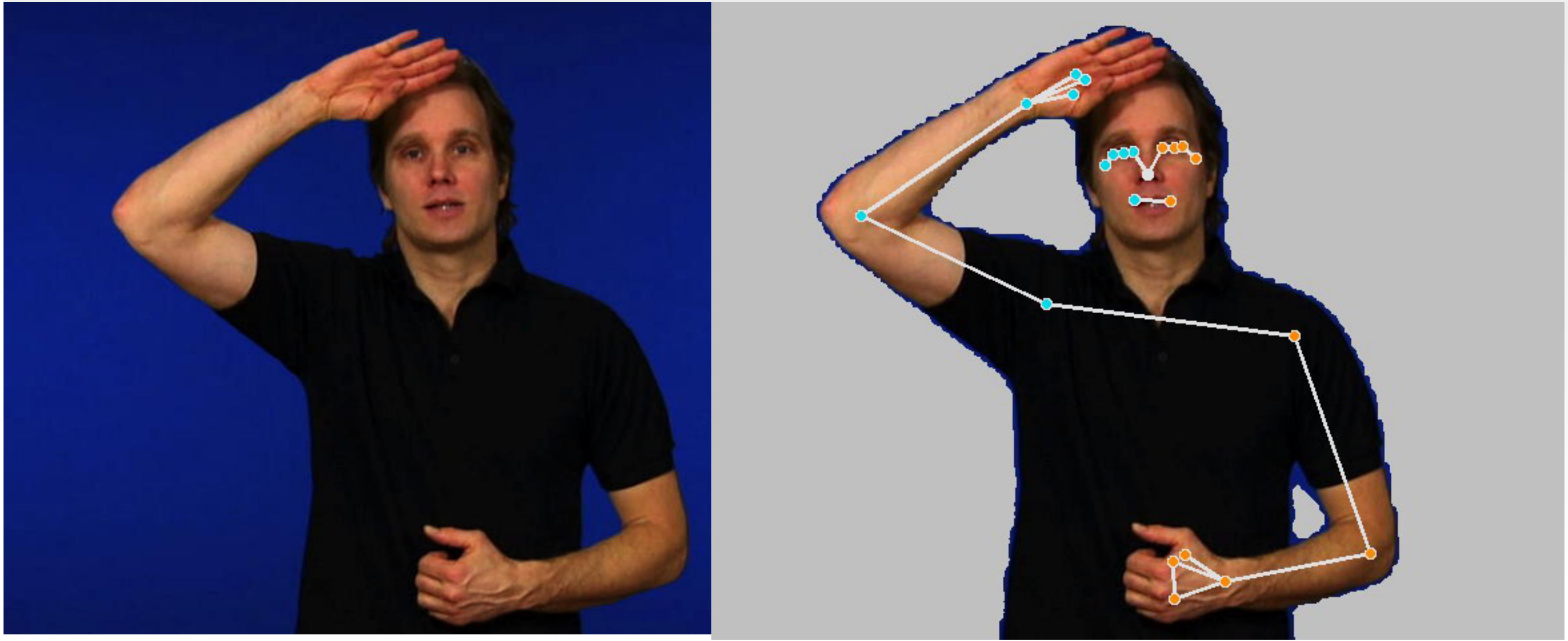
The data

- 1,292 **videos** (approximately 2 secs long each)

- 107,955 video **frames** (videos are 25 fps or 50 fps)

- Only **5 landmarks** included = 539,775 **data points**

The output



The output

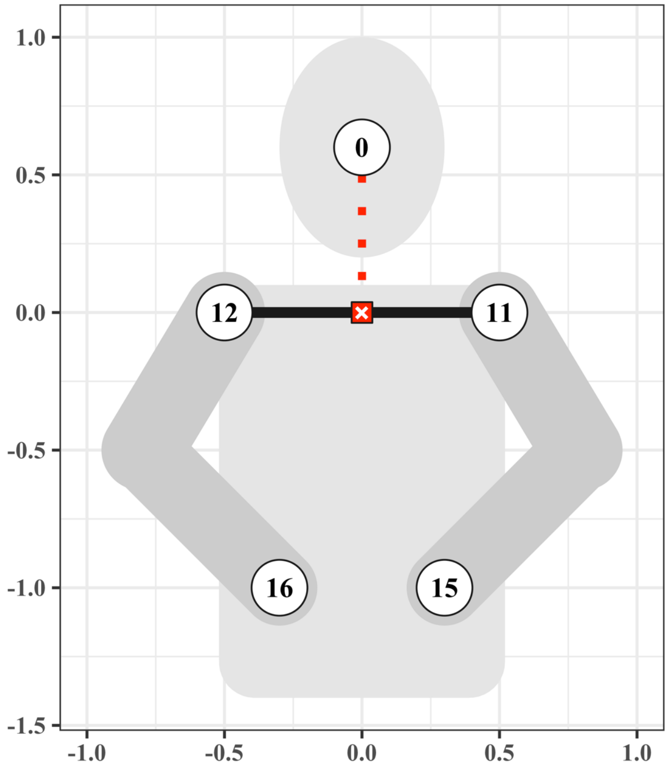
```
language,id,frame,video_height,video_width,landmark,x,y,hand,hands,movement,location
STS,4165,1,720,960,0,0.4843829870223999,0.22800175845623016,right,1_1,sym,neutral
STS,4165,1,720,960,11,0.639739453792572,0.4301624298095703,right,1_1,sym,neutral
STS,4165,1,720,960,12,0.3643210232257843,0.44197335839271545,right,1_1,sym,neutral
STS,4165,1,720,960,15,0.5331403613090515,0.852113664150238,right,1_1,sym,neutral
STS,4165,1,720,960,16,0.48698344826698303,0.8003442287445068,right,1_1,sym,neutral
STS,4165,2,720,960,0,0.4856907427310944,0.22824083268642423,right,1_1,sym,neutral
```

…

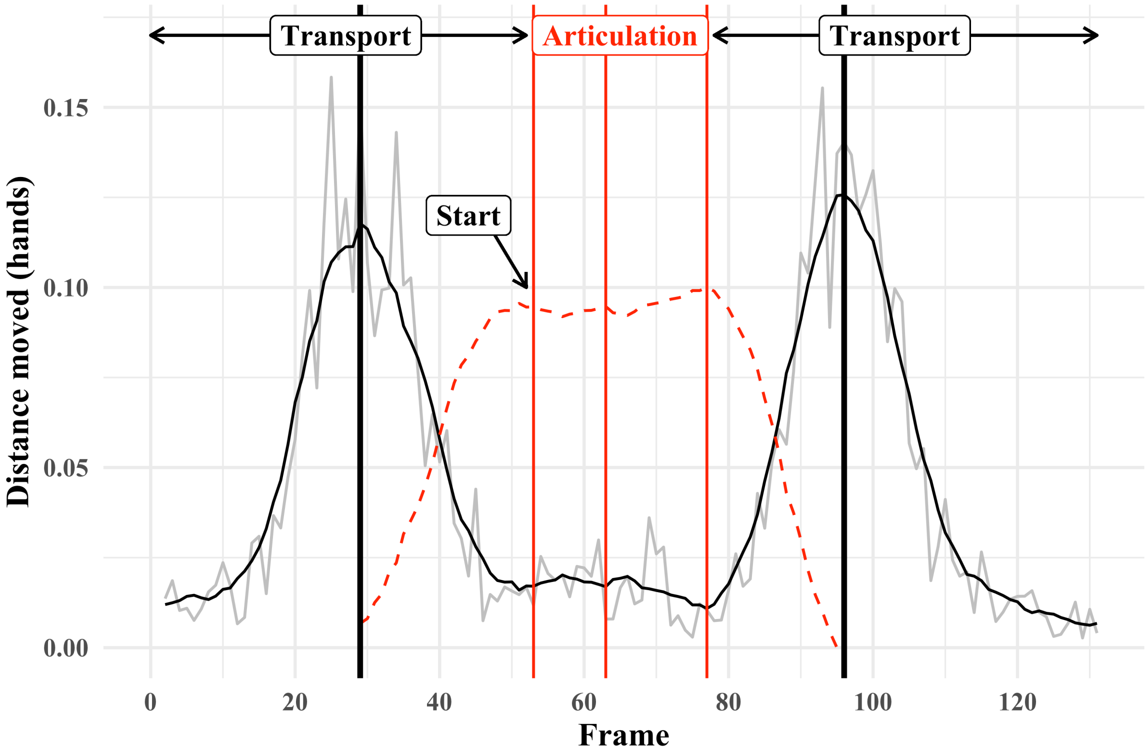
Full data and code: https://osf.io/x3pvq/

Normalizing the data

- **Shoulder distance** = norm
- Mean midpoint between shoulders is **origo**: everything's center
- **X axis** is scaled to norm = 1
- **Y axis** is scaled to norm = 0.6



Estimating articulation phase

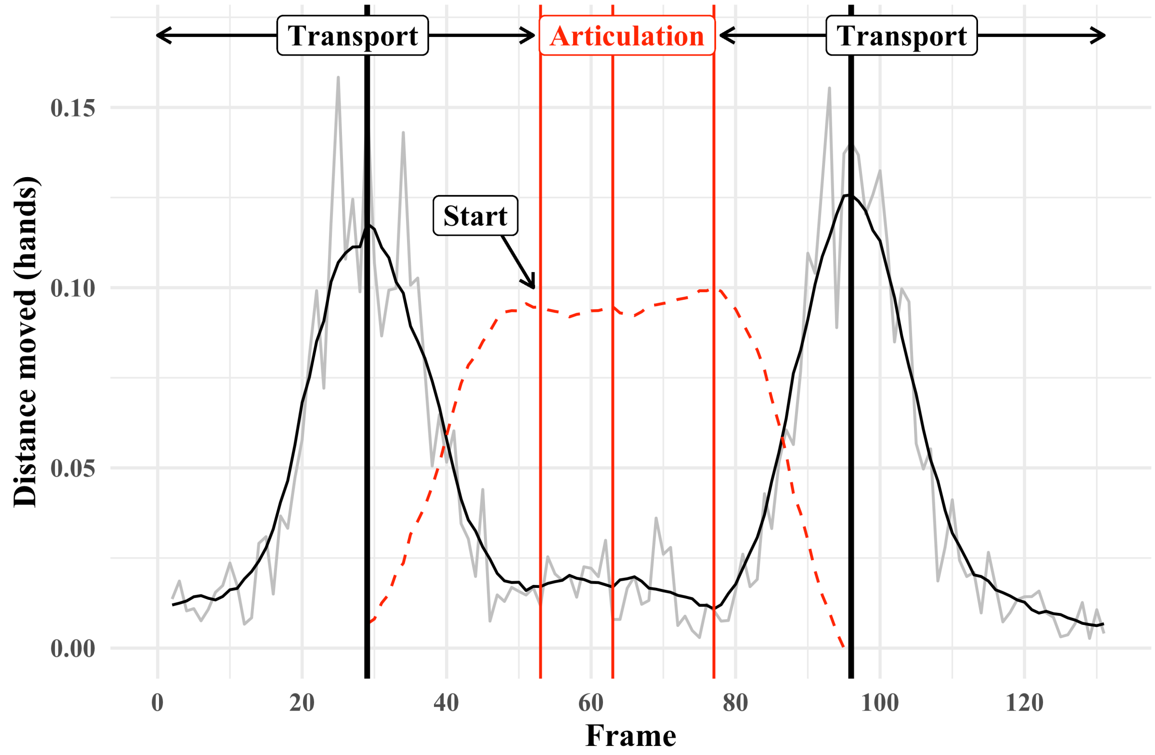


Estimating articulation phase

- Total **distance traveled** by **both hands**

- Peaks in smoothed curve ≈ transport

- Valleys between peaks ≈ articulation

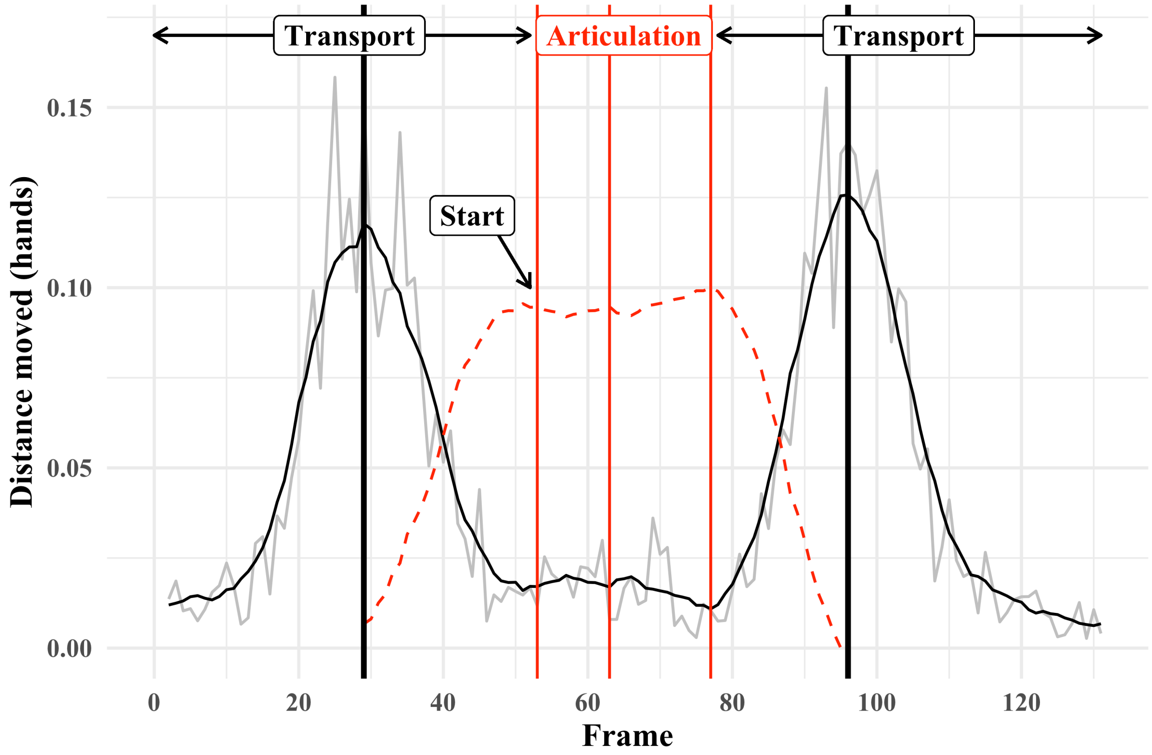


Estimating articulation phase

- Articulation phase is the **short sign**

- The first valley is the **start**

- The entire video is the **full sign**

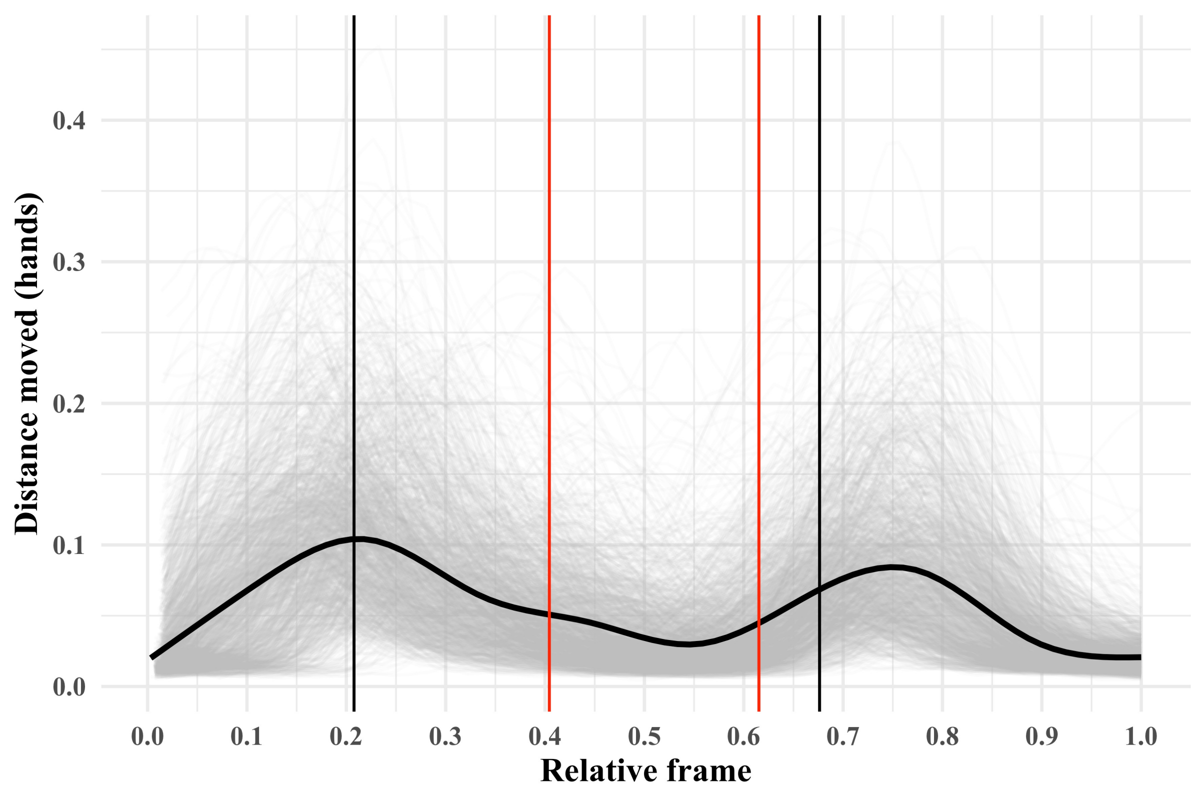


Estimating articulation phase

- All signs have **peaks**

- 96.4% of signs had at least one **valley**

- If only one valley, end was defined as 2nd peak OR start+10 frames

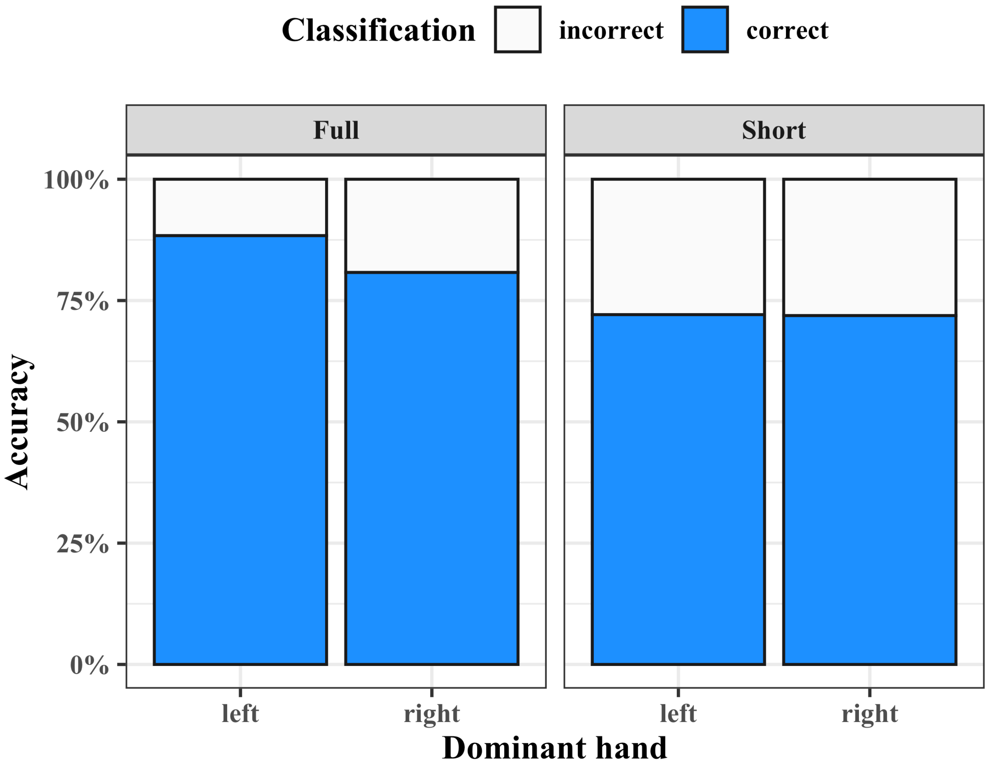


Estimating hand dominance

- Hand dominance = which hand **traveled a longer distance** (right is default)

- Estimating hand dominance is more accurate with **full sign**

- No obvious difference between **left/right**

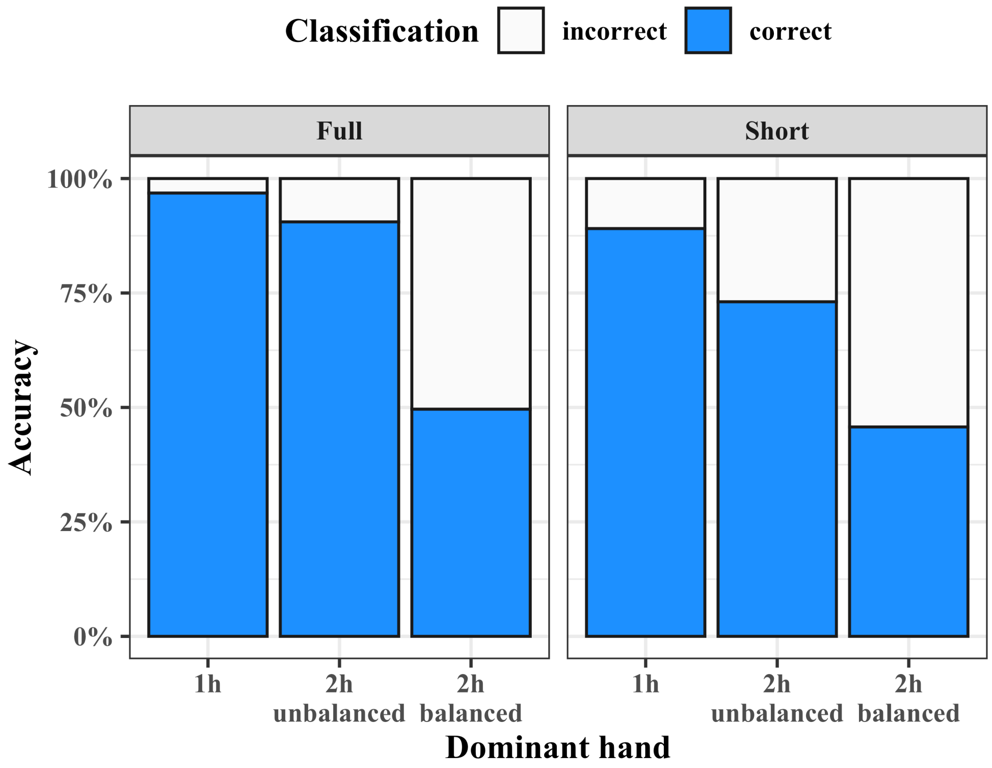


Estimating hand dominance

- Hand dominance estimation is **more accurate with one-handed signs**

- **Full method** still better

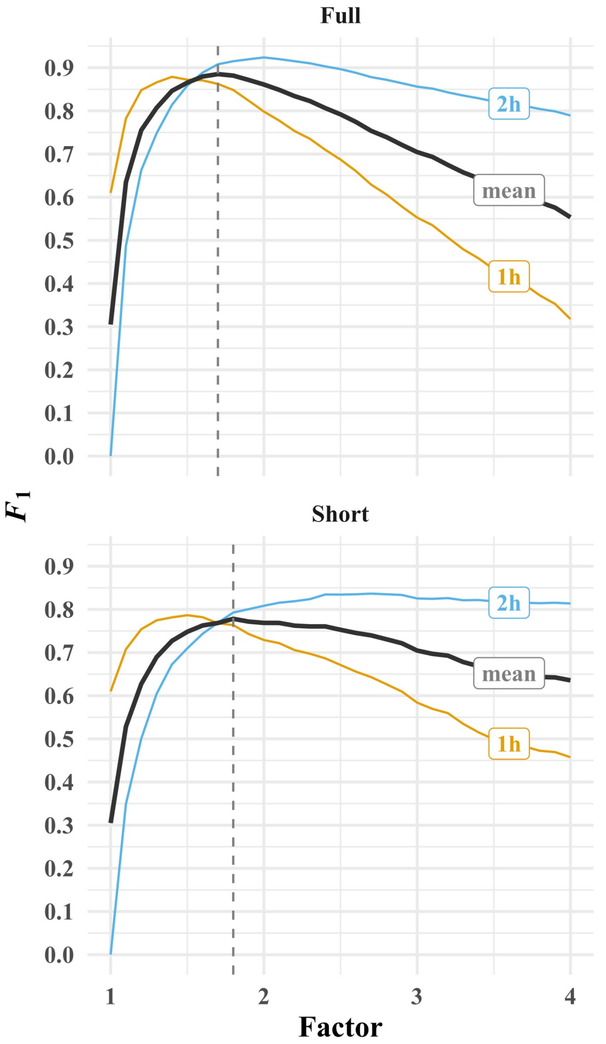
- **Two-handed** dominance doesn’t really matter



Estimating # of hands

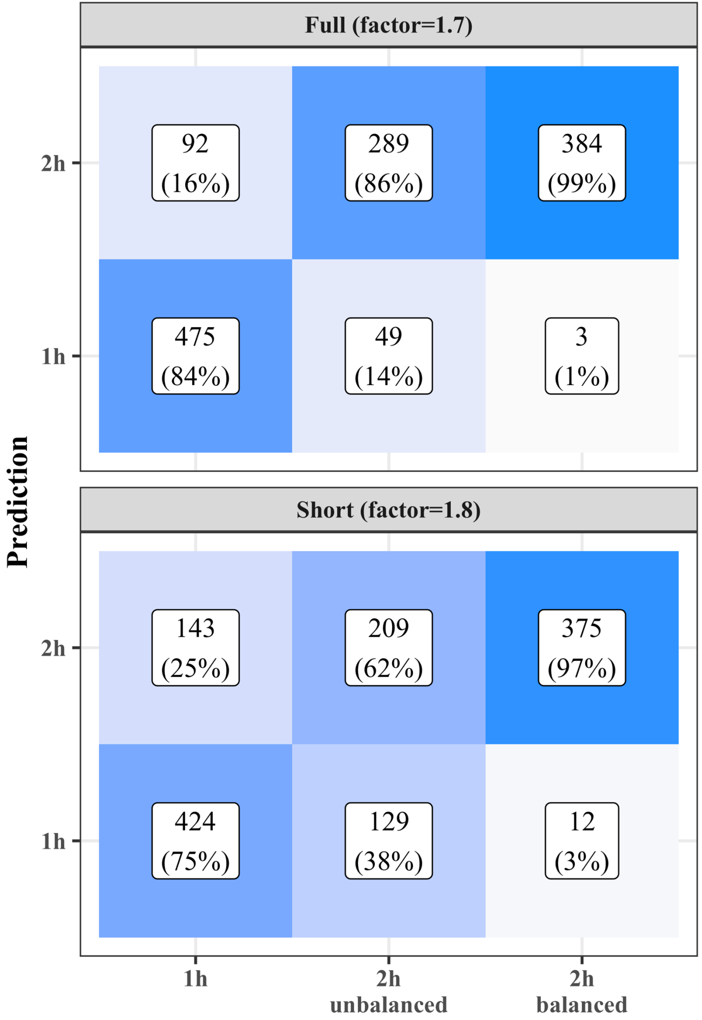
- Previous work (Östling et al. 2018) used a factor of 3 as the cut-off point in deciding number of hands
    - If one hand **moved 3x longer** than the other, it is a **one-handed** sign

- I tested the most accurate factor for the STS signs:
    - A **factor of ≈2** seems best!



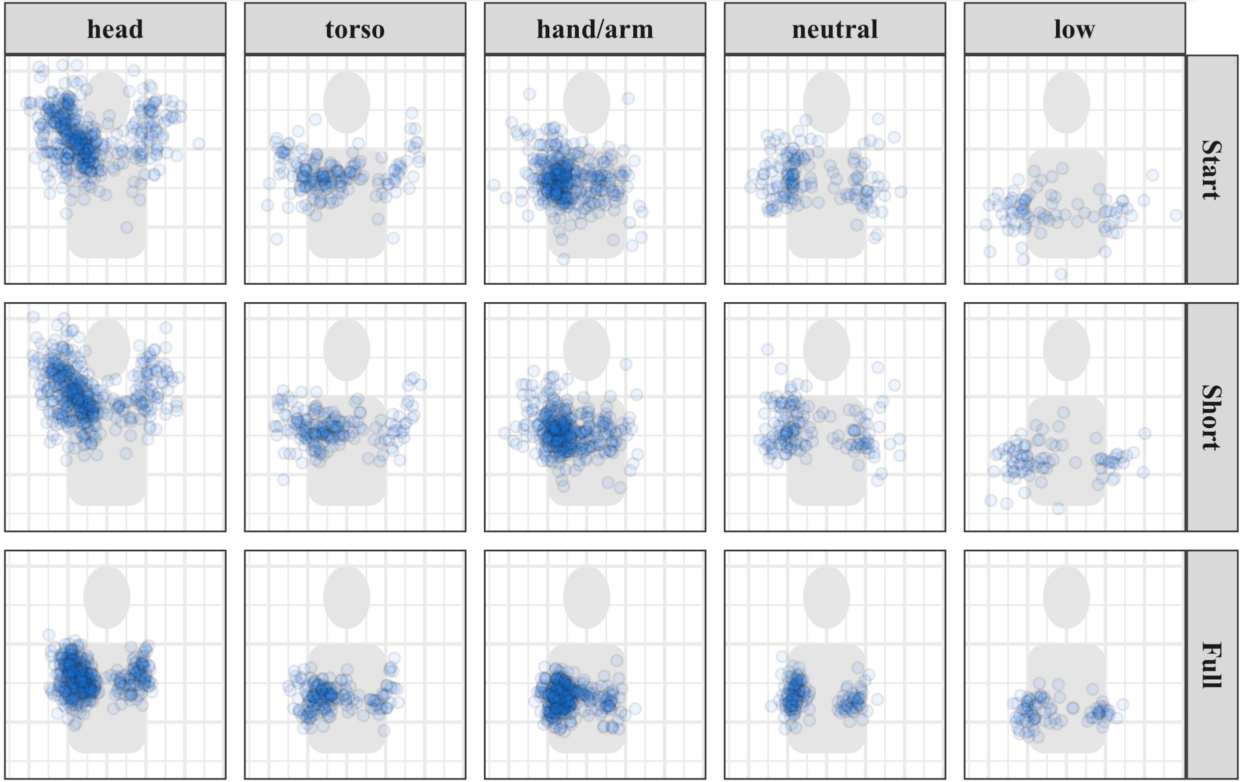
Estimating # of hands

- The **full method** is still the best
- The method is very **accurate with two-handed** signs, but struggles a little with one-handed signs and unbalanced two-handed signs
  - Unbalanced signs are in a way both one- and two-handed!

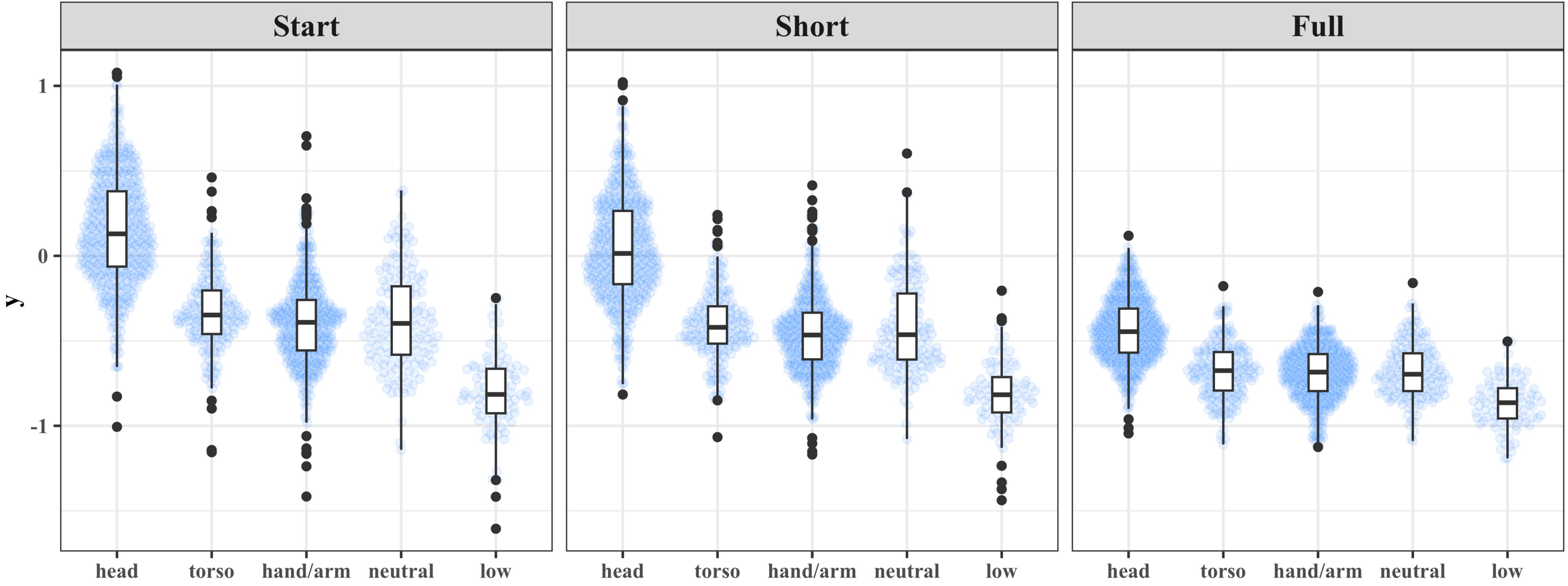


Estimating place of articulation

- Finally, the **short method** paid off!
- Also **finding the start** is useful
- **Full method = useless**
  - Shows transport and rest



Estimating place of articulation



Conclusions

- **MediaPipe can be used to extract information** about **sign form** directly from videos

- **Transport movements** (in dictionary signs) are useful for estimating **hand dominance** and **number of hands**
  - We simply get **more data** (and a **bigger difference**)

- Estimating place of articulation requires estimation of the **key part** of an actual **articulation phase** (e.g. the start)

Thanks!