Snapture - a Hybrid Hand Gesture Recognition System

Hassan Ali

Examiners: Prof. Dr. S. Wermter, Dr. D. Jirak

KNOWLEDGE TECHNOLOGY

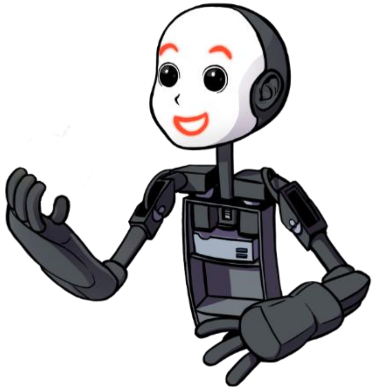
http://www.informatik.uni-hamburg.de/WTM/

Motivation

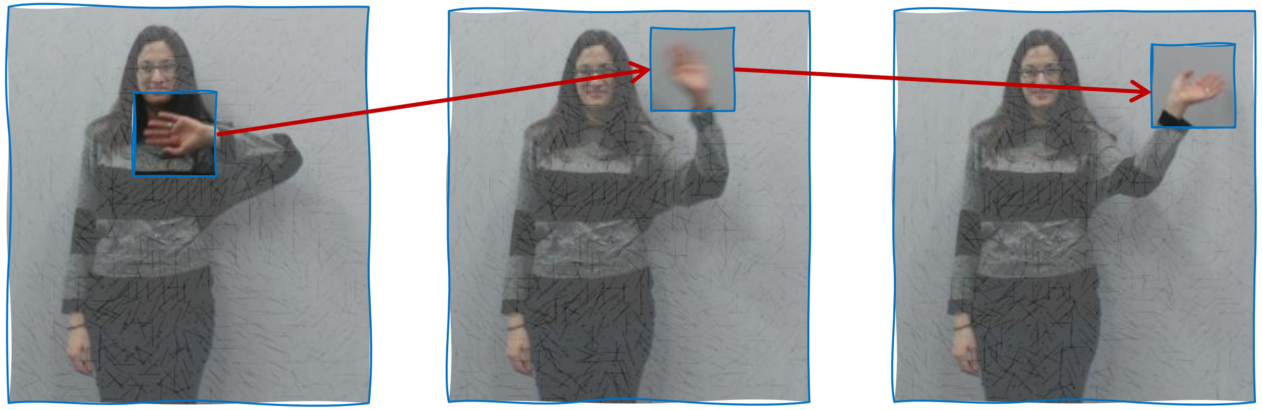
- Hand gesture applications
- Gesture taxonomy

Static

Dynamic



- Vision-based systems



Hello

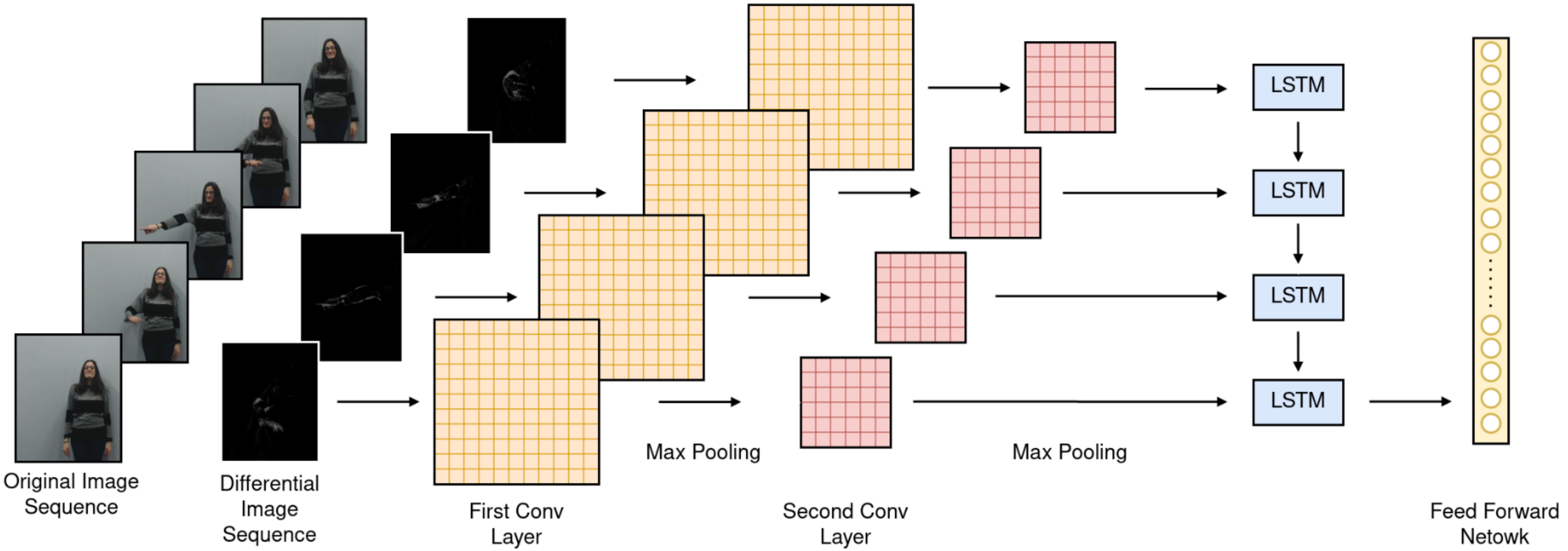
Hand detection

Hand tracking

Recognition

CNNLSTM

- Evaluated using the Tsironi GRIT dataset (available on WTM website)
- We reproduce the results.



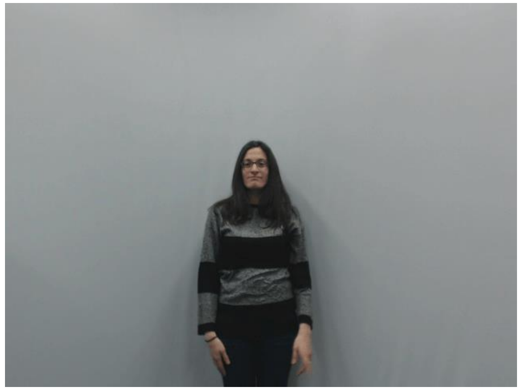
Research Questions

- CNNLSTM
  - How influenced is the CNNLSTM network by subject variability?
  - How efficient is the CNNLSTM network at learning co-speech gestures?

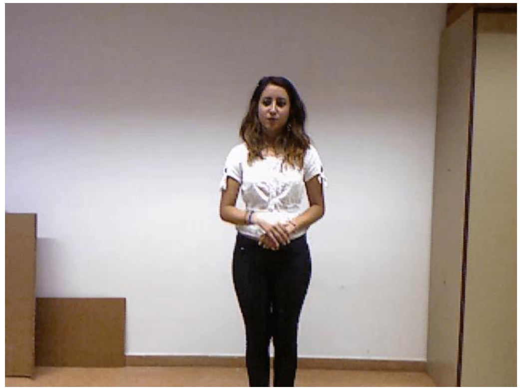
- Snapture
  - How to identify the peak of the gesture and extract the handshape using RGB data only?
  - How to regulate the integration of the hand details into a dynamic gesture recognition system?

1. How influenced is the CNNLSTM network by subject variability?

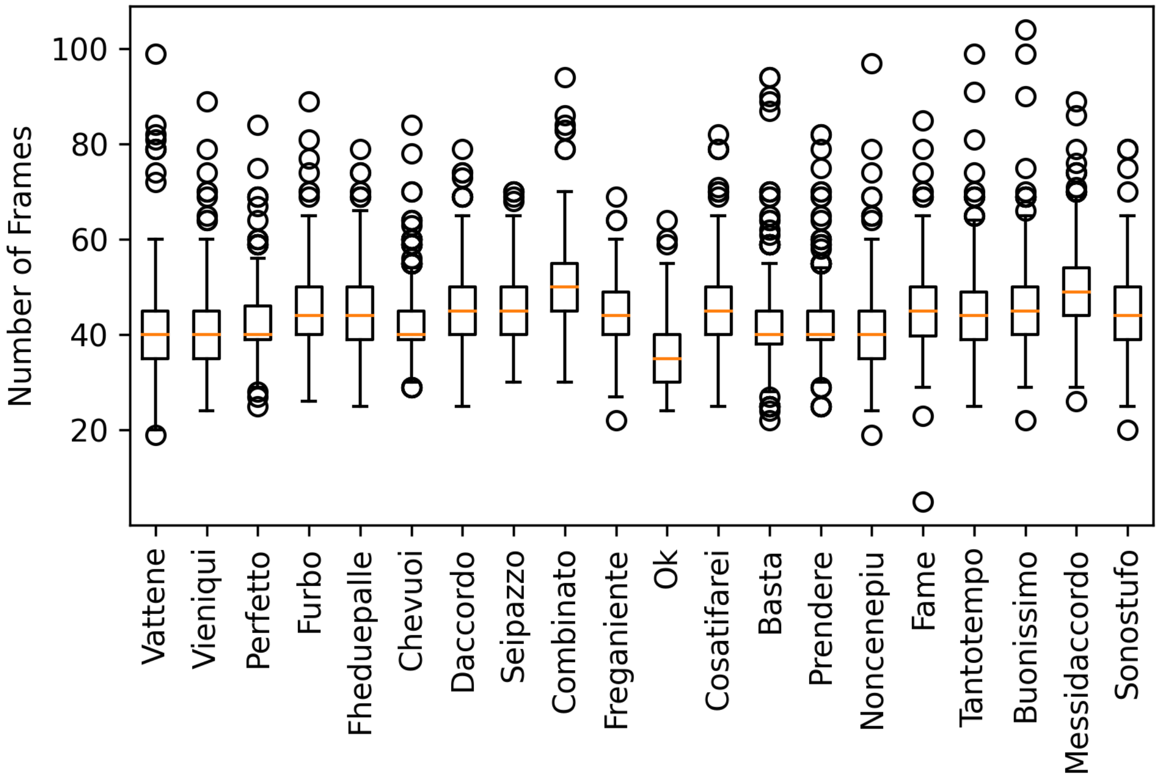


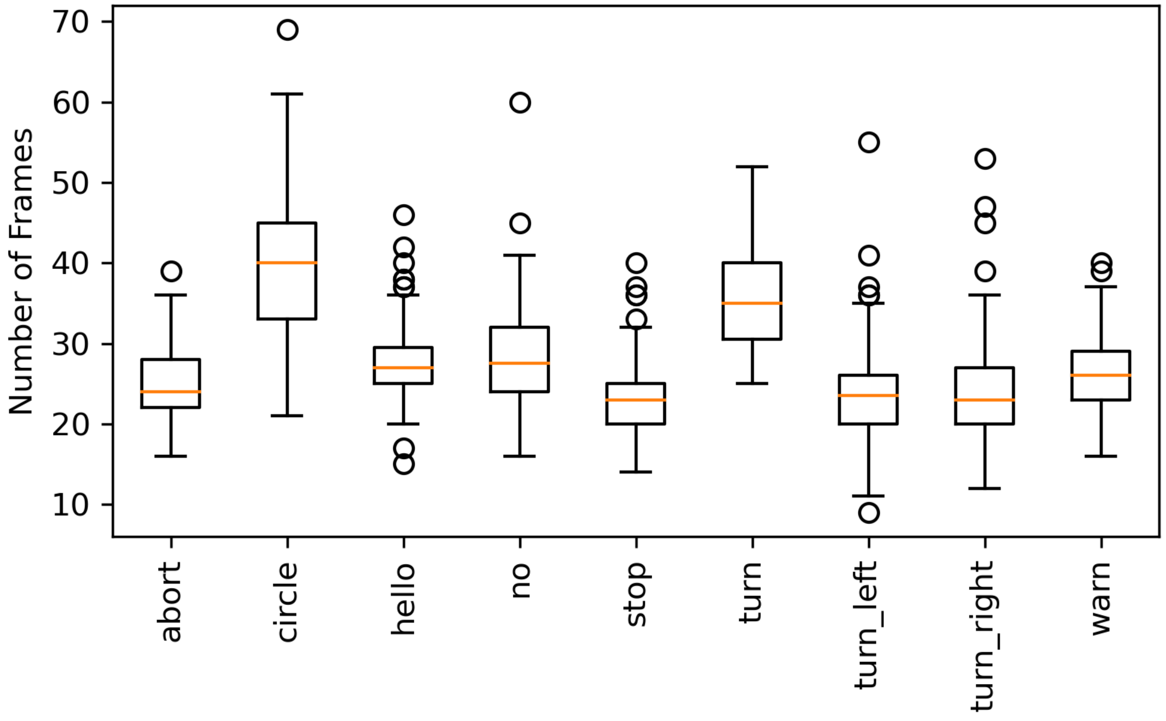
Prendere Daccordo Turn Left No

CNNLSTM – Subject Variability



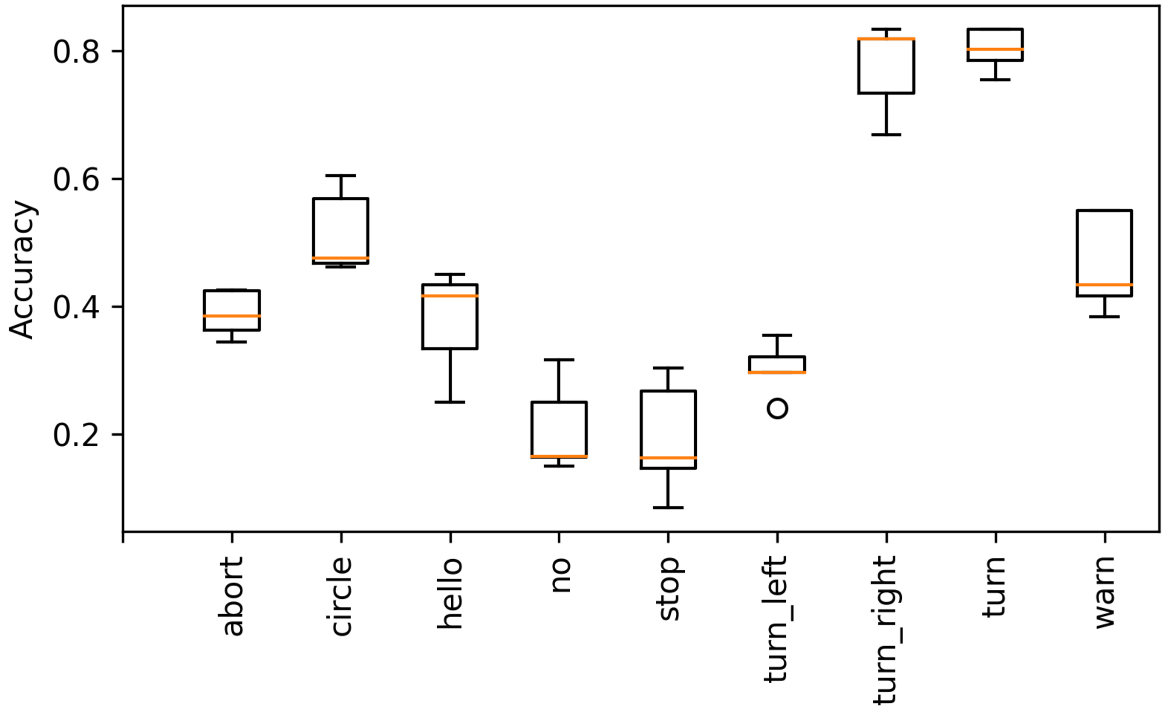
Montalbano (Chalearn Challenge)



Tsironi GRIT

CNNLSTM – Subject Variability

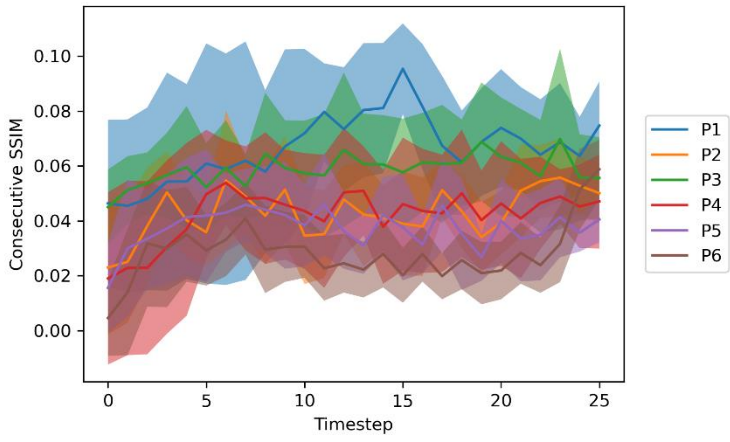
- Experiment using the GRIT dataset
- Evaluate on unseen subjects (Leave-one-out approach)
- Low accuracy for most classes



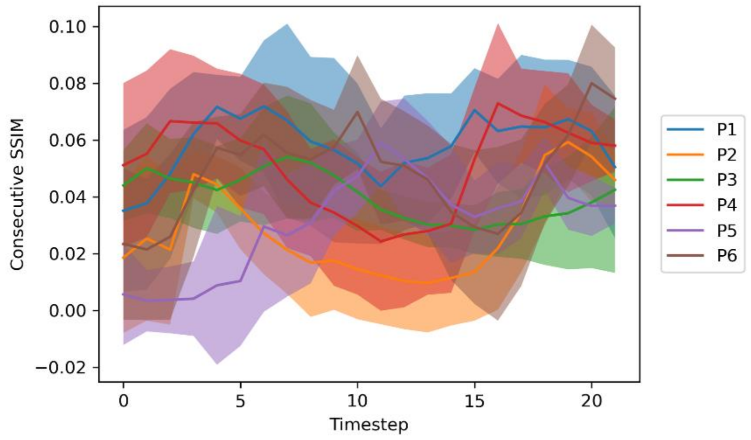
Accuracy per class, avg. over all subjects, avg. of 5 trials

CNNLSTM – Subject Variability

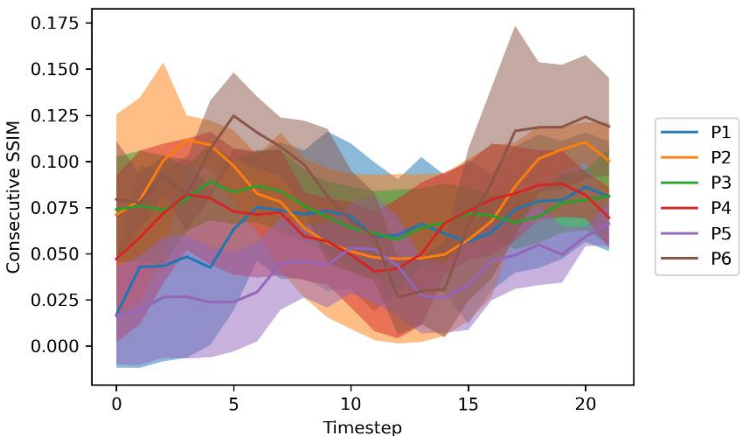
- Motion profile of the worst-case classes: *No*, *Stop*,
- and the best-case classes: *Turn Right*, *Turn*
- **Consequence**: train on data of all subjects.



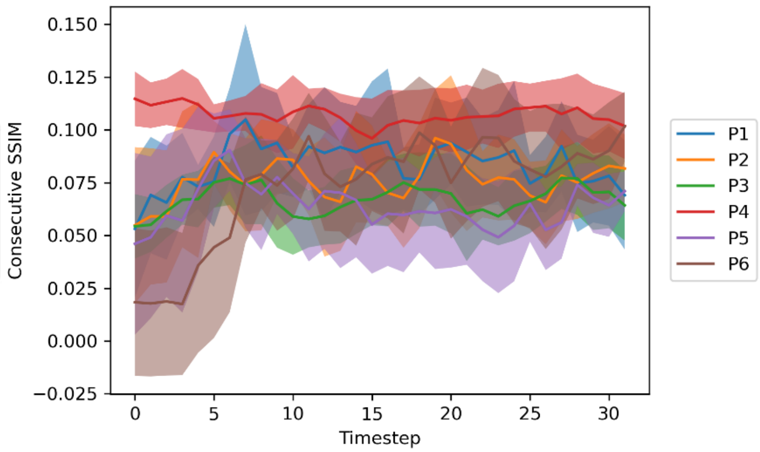
No



Stop



Turn Right



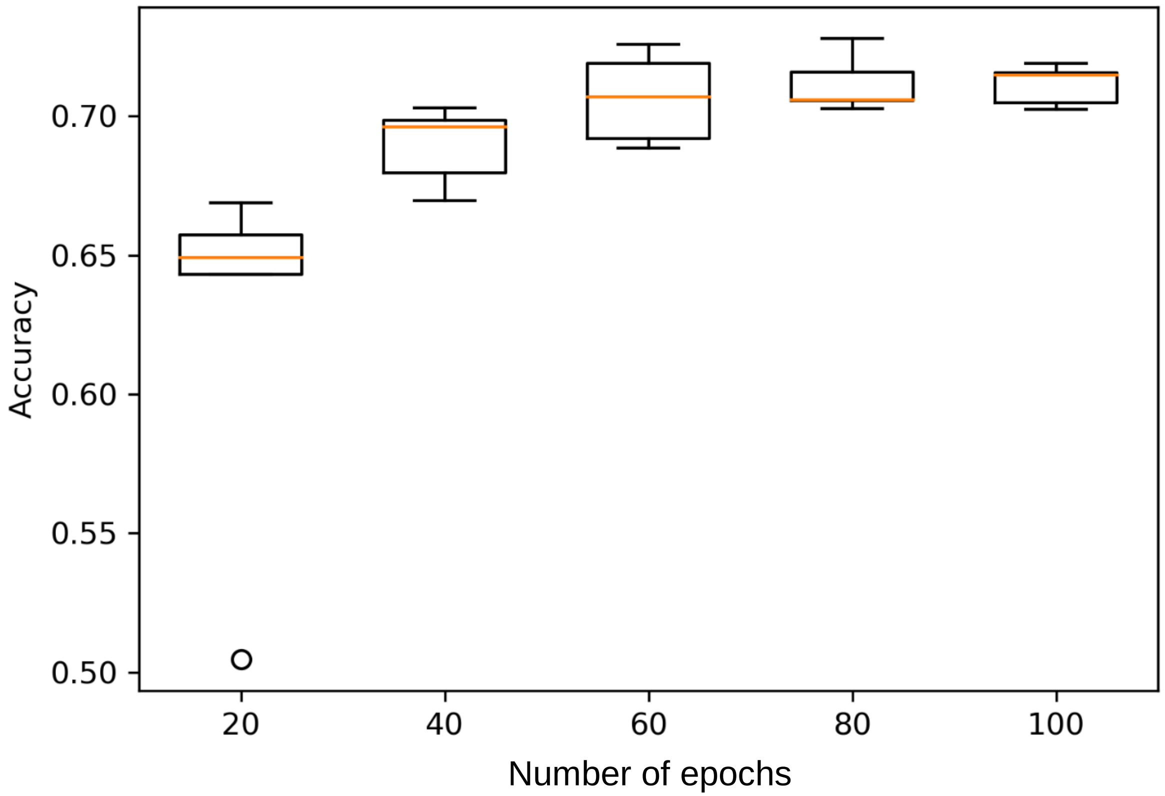
Turn

2. How efficient is the CNNLSTM network at learning co-speech gestures?

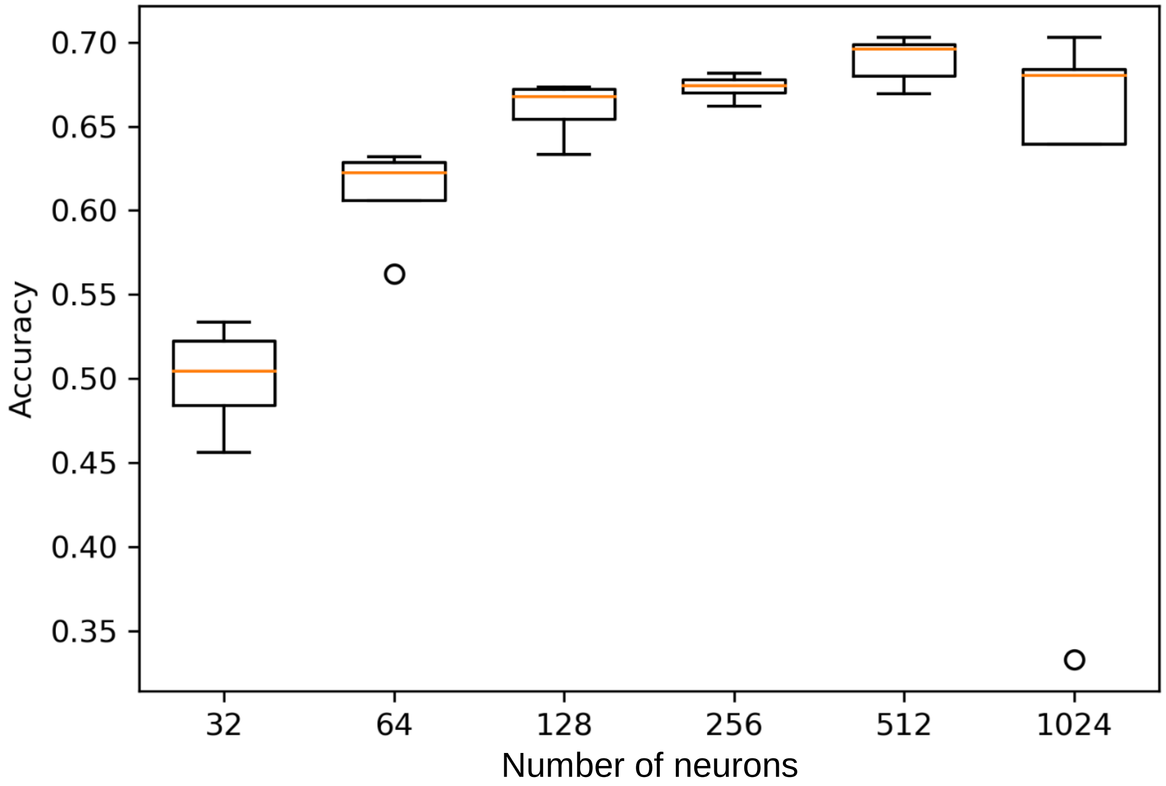
- Grit: suitable for robot commands, unique motion paths, lab settings
- Montalbano: co-speech (more profound)

Dataset	Chalearn Montalbano	Tsironi GRIT
#classes	**20**	9
#observations	**13 342**	542
#participants	**48**	6
#scenes	**5**	1

CNNLSTM – Upscaling



Avg. of 5 trials



Avg. of 5 trials

CNNLSTM – Upscaling

- Two Issues (Hypotheses):
  - *It is challenging for the CNNLSTM model to distinguish classes with similar movement patterns.*
  - *It is challenging for the CNNLSTM model to distinguish subtle movements done at the peak of the gesture.*
- **Solution:** *Snapture* architecture

3. How to identify the peak of the gesture and extract the handshape using RGB data only?

Gesture phases (Kendon)

1. Rest position

2. Pre-stroke

3. Stroke

4. Post-stroke

5. Rest position

Motion Profile

- **Problem:** analysis of motion/pause carried in a movement sequence
- **Solution:** structure similarity (SSIM) index

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

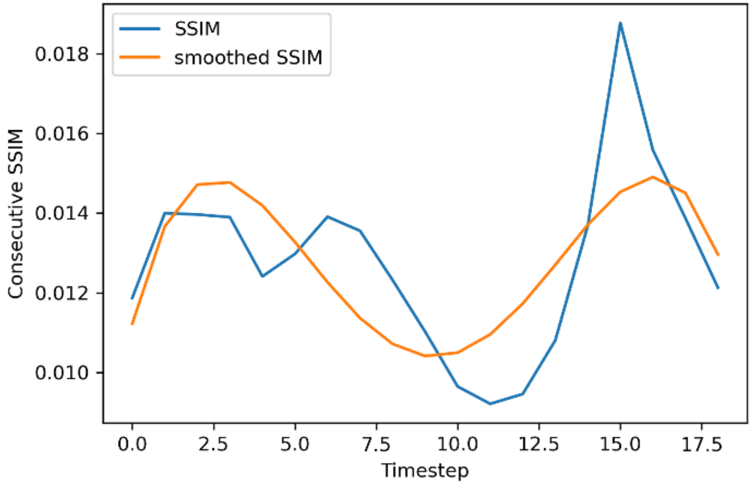
$$inverted_{SSIM} = 1 - \sum SSIM(\Delta_i, \Delta_{i-1})$$

$$\Delta_i = (I_i - I_{i-1}) \wedge (I_{i+1} - I_i)$$

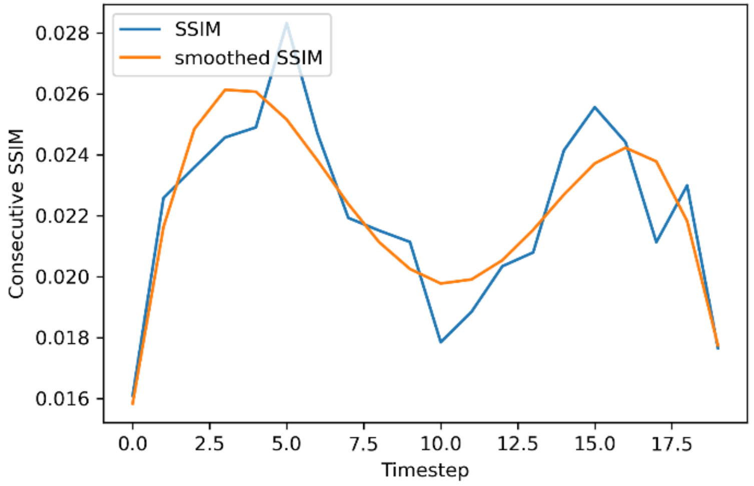
μ: avg. intensity, σ^2 :variance
C_1, C_1: stability constants
Δ_i, Δ_{i-1}: differential images
I_{i-1}, I_i, I_{i+1}: original frames

Motion Profile

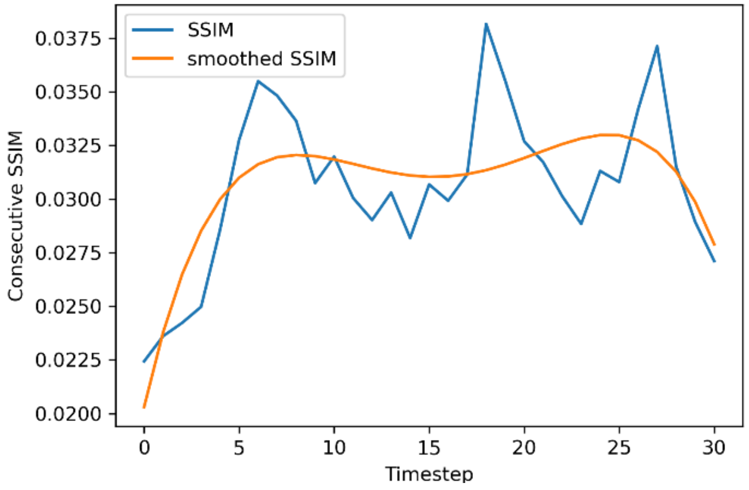
- Tsironi GRIT data:
  - *paused-gestures:* the arm remains briefly in a fixed position at the peak
  - or gestures with *repeated-pattern:* include a motion pattern, usually circular



Stop



Turn Left

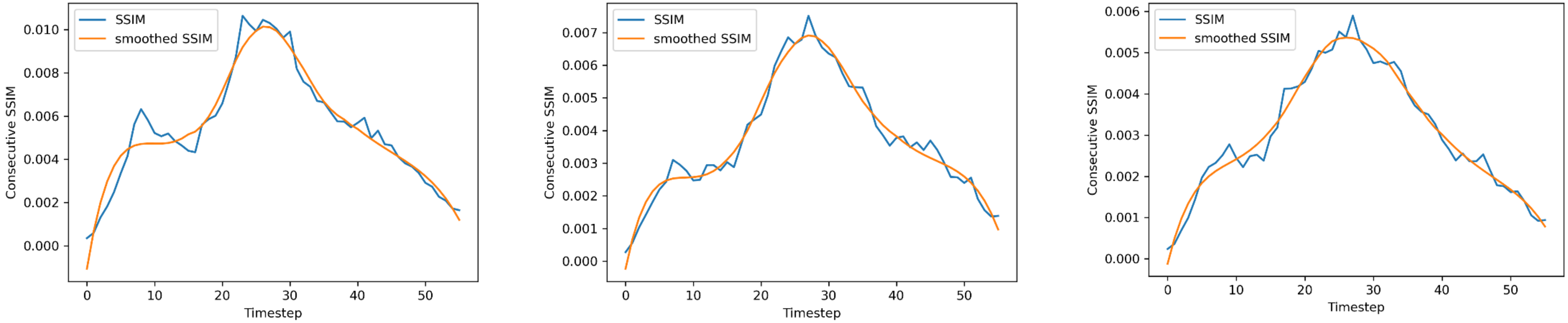


Turn

Motion Profile

- Chalearn Montalbano movements have comparable profile with pause at the peak

  → Peak around the mean of the sequence length



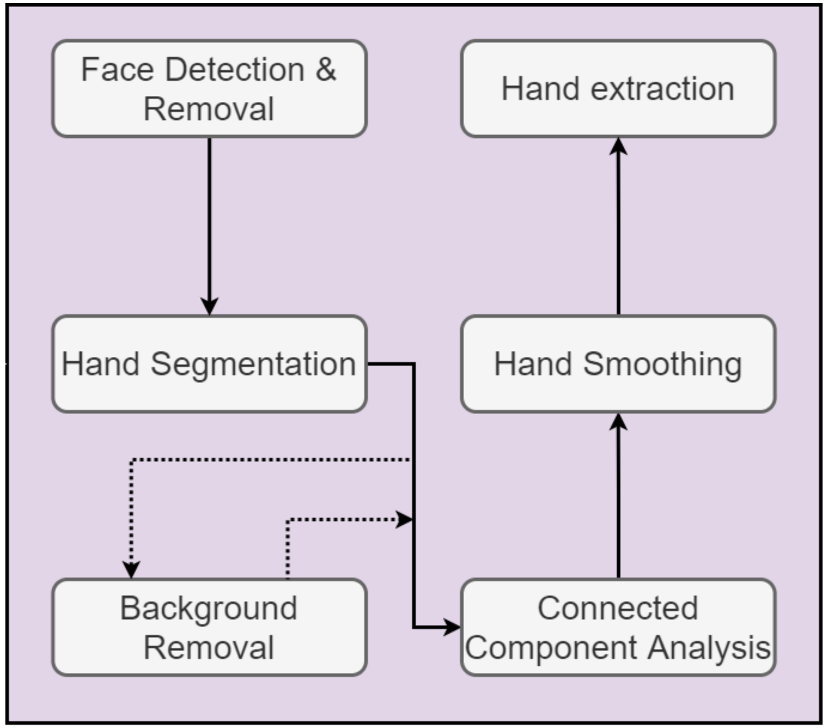
Vattene Vieniqui Ok

Snapture Architecture

Static Channel

- Gesture Peak Detection
- Gesture Peak Extraction

Frame at the peak of the gesture

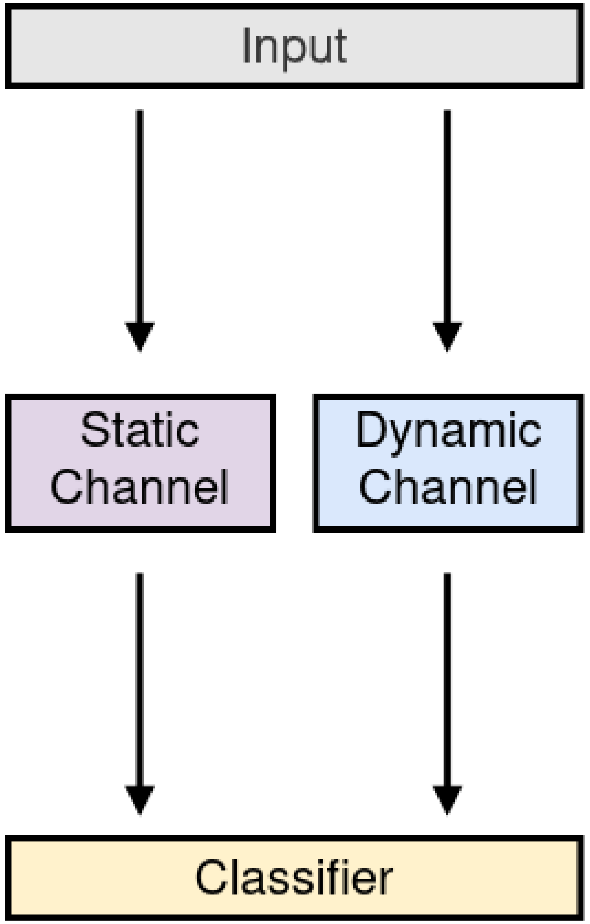


Snapshot Extraction

- Independent of subject's dominant hand

Snapture Architecture

- SNAPshot capTURE is our proposed architecture
- Hybrid (static/ dynamic) gesture recognition
- **Input:** *isolated sequences*



Snapture Architecture

Dynamic Channel

sequence-level CNNLSTM

Freganiente

Ok

Noncenepiu

Static Channel

a *snapshot* at the peak, CNN-based

4. How to regulate the integration of the hand details into a dynamic gesture recognition system?

- Some gestures, e.g., *Circle* are strictly dynamic .
- Low camera frame-rate

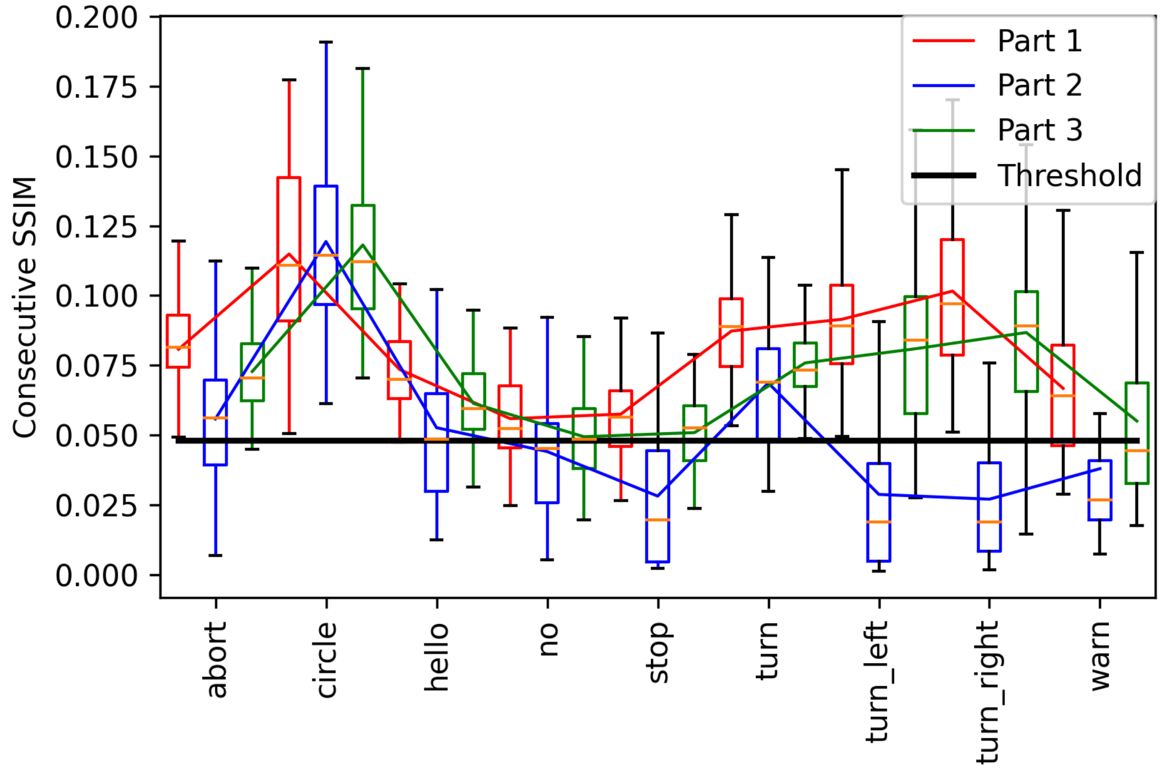
  → *blurriness* issue
- **Solution:** *threshold-controlled* approach based on sufficient pause

Circle

Stop

Regulating the static channel

- Lower values represent longer pauses.
- More pronounced for Stop, Turn Left and Turn Right
- Approx. only 44% of the GRIT samples include a pause

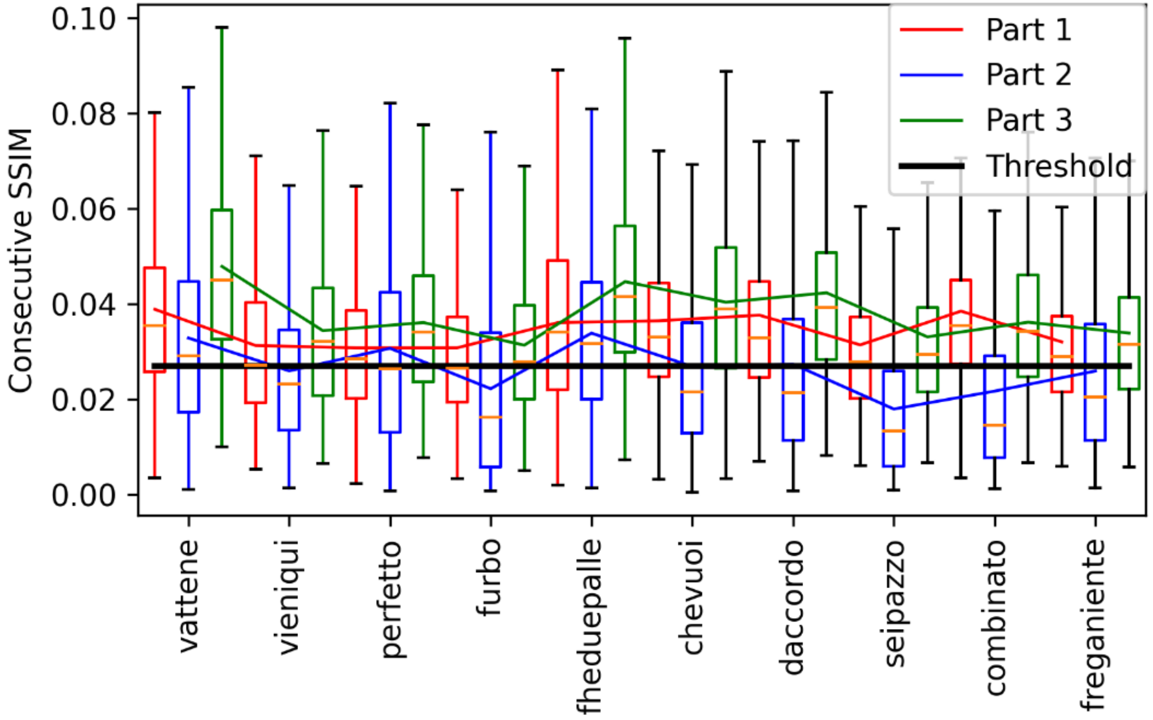


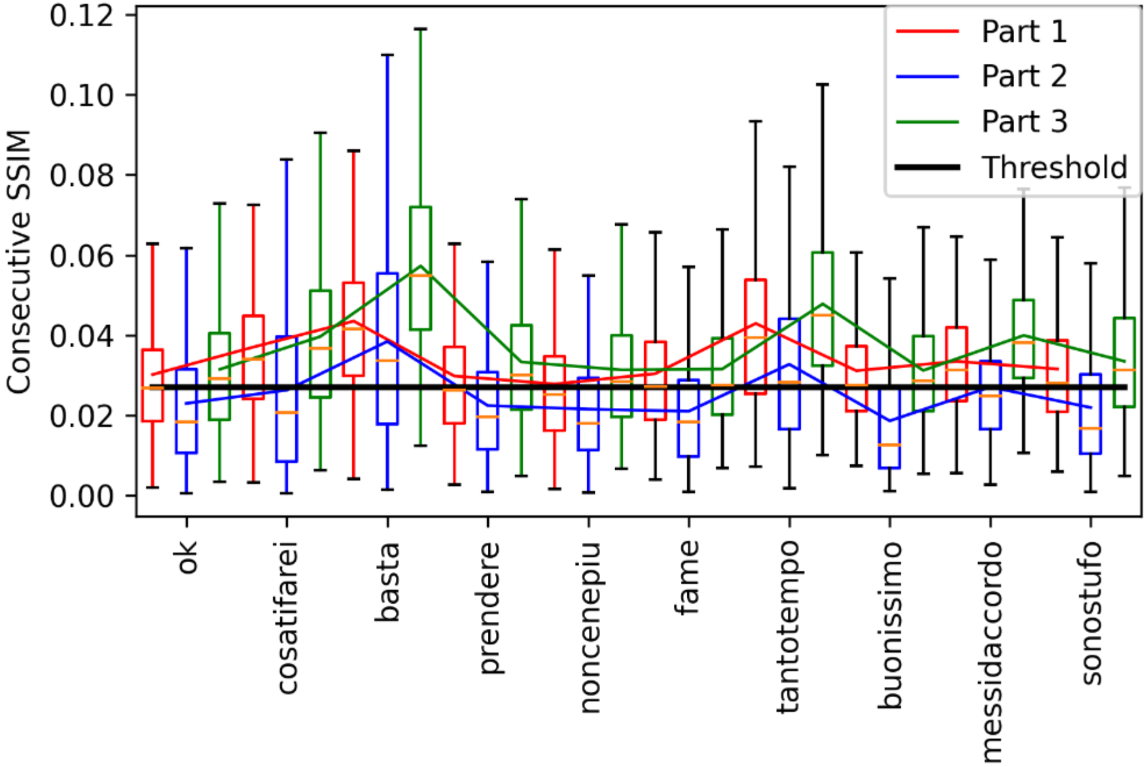
Part 1: rest to pre-stroke, Part 2: pre-stroke to post-stroke, Part 3: post-stroke to rest

Regulating the static channel

- Co-speech movements include more pause at the peak (approx. 70% of the Montalbano samples).

Part 1: rest to pre-stroke, Part 2: pre-stroke to post-stroke, Part 3: post-stroke to rest





Results

Tsironi GRIT Dataset

Model	Accuracy	F1-score	Time*
CNNLSTM	0.91 (0.012)	0.913 (0.012)	140.612 (0.255)
Snapture	0.924 (0.006)	**0.927** (0.005)	170.012 (1.027)
Snapture $_{thold}$	**0.926** (0.008)	0.913 (0.012)	125.156 (1.117)

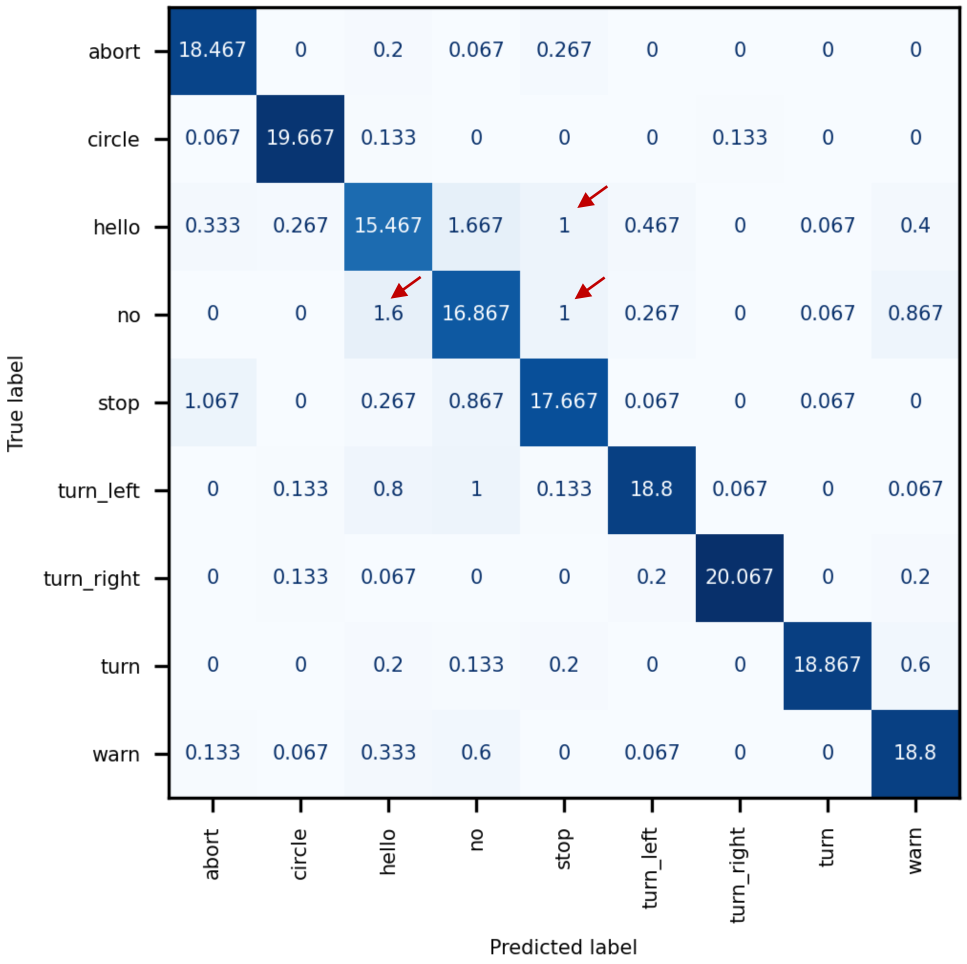
***In seconds.**

Chalearn Montalbano Dataset

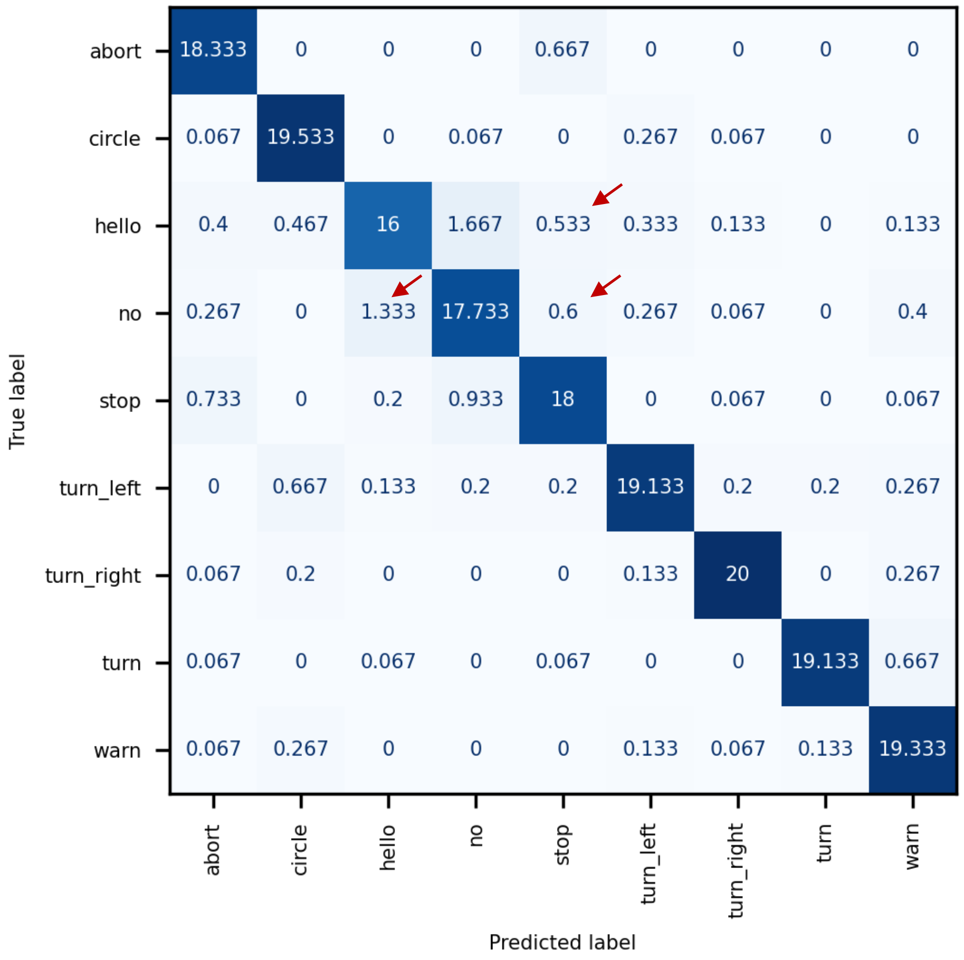
Model	Accuracy	F1-score	Time*
CNNLSTM	0.699 (0.014)	0.701 (0.013)	234.762 (0.115)
Snapture	0.755 (0.021)	0.752 (0.021)	318.578 (0.428)
Snapture $_{thold}$	**0.77** (0.008)	**0.772** (0.007)	744.953 (0.724)

***In minutes.**

Results Analysis - GRIT



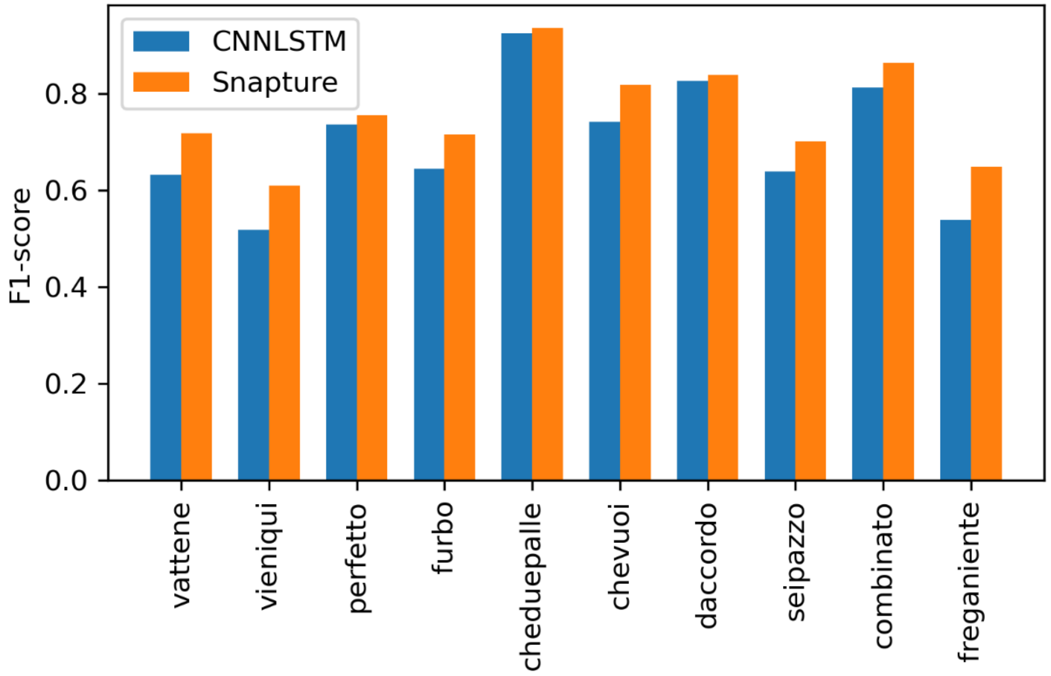
CNNLSTM
(avg. 5 trials)

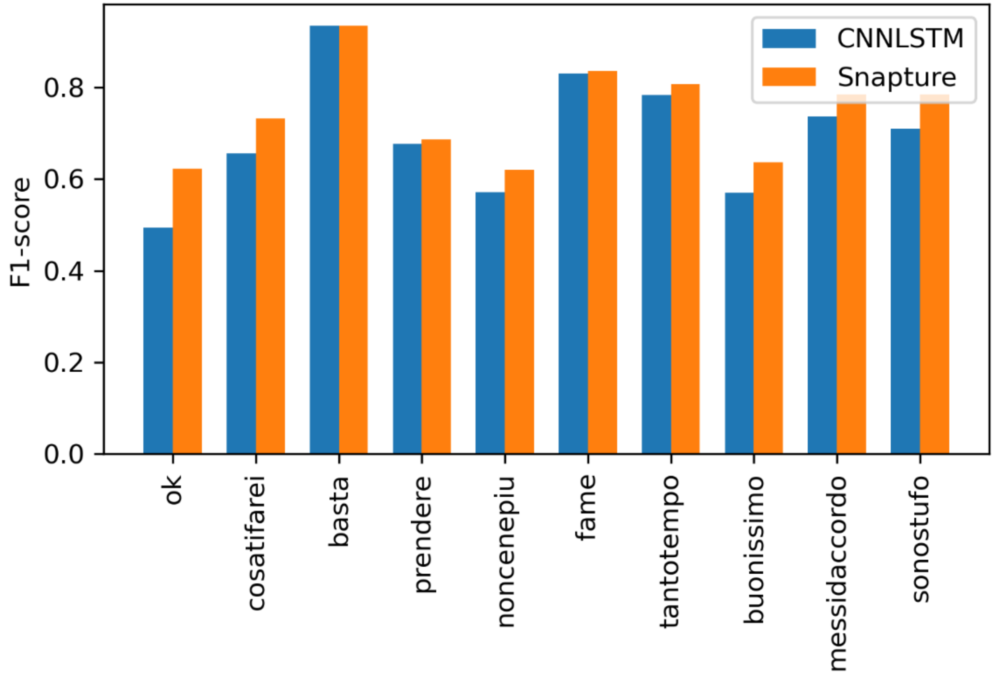


Snapture
(avg. 5 trials)

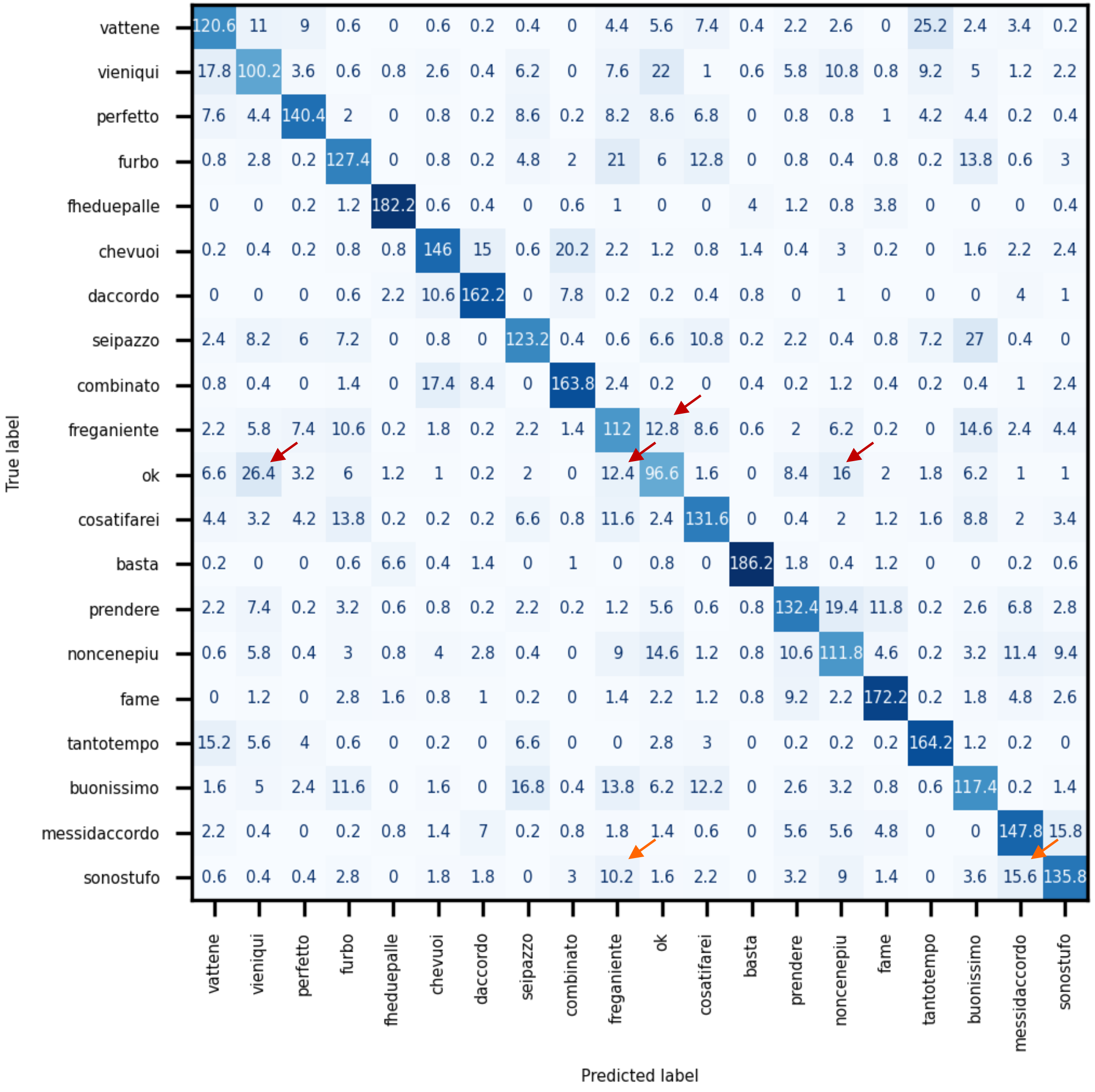
Results Analysis - Montalbano

- *Snapture:* superior results on all classes except for *Basta*
- Boosted F1-score for unique handshape classes (ex: *Ok*)

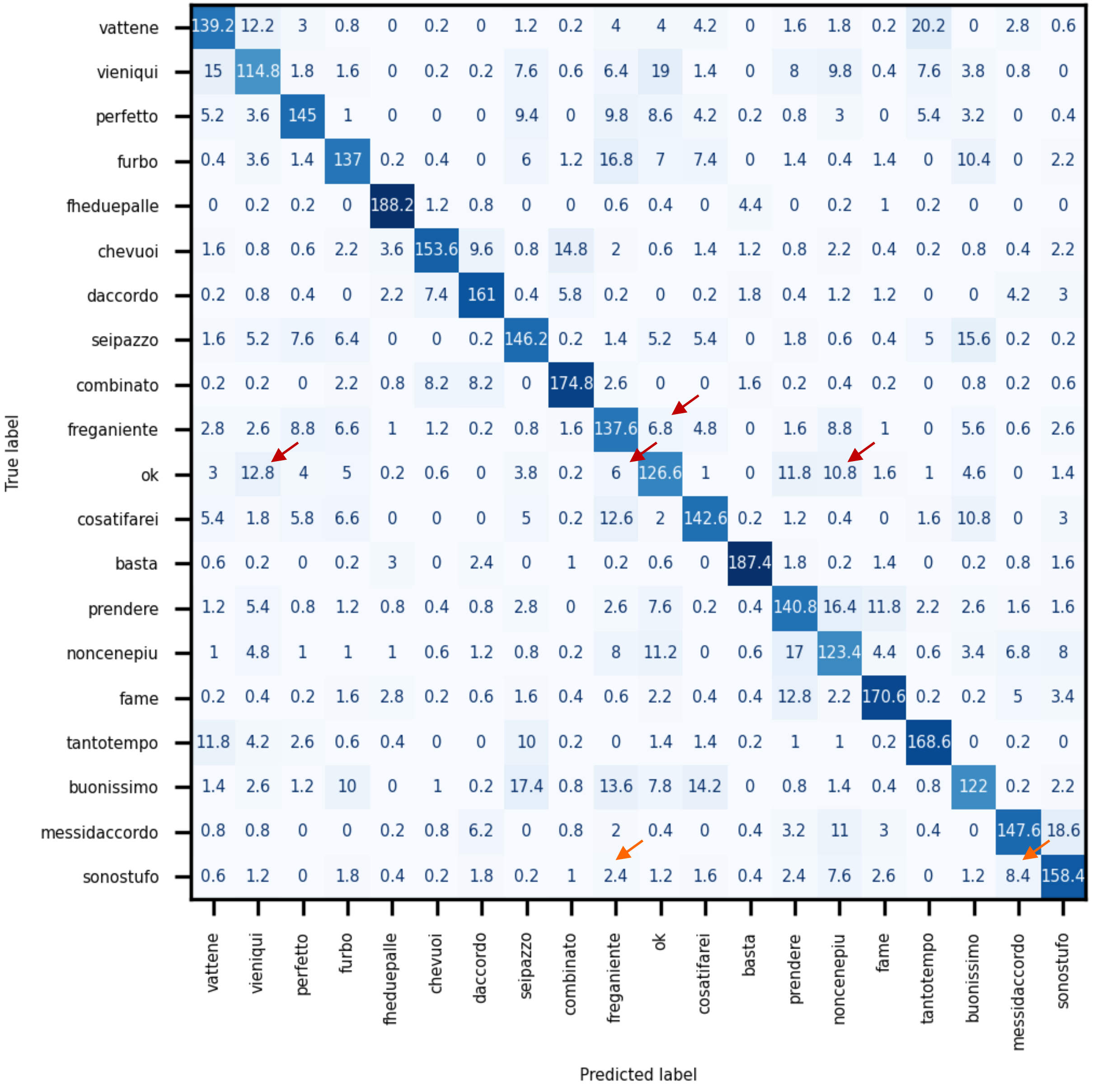




Results Analysis - Montalbano



CNNLSTM (avg. 5 trials)



Snapture (avg. 5 trials)

Results Analysis - Montalbano

- *Snapture* boosts the classification of indistinctive movements.






Cosatifarei | Vattene | Perfetto | Freganiente | Ok | Noncenepiu






snapshot | *snapshot* | *snapshot* | *snapshot* | *snapshot* | *snapshot*

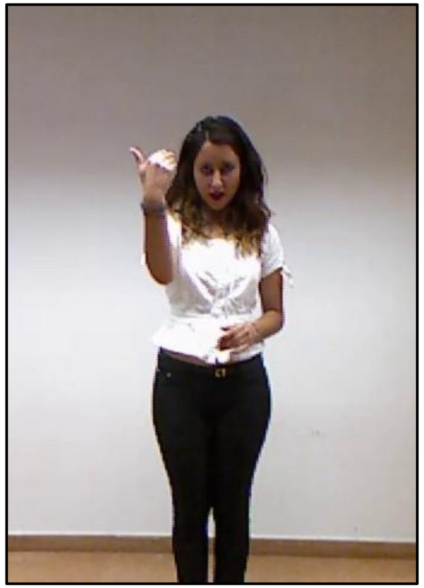
Results Analysis - Montalbano

- *Snapture* boosts the classification of subtle movements.

Basta (explicit hand movement

Sonostufo (subtle hand movement)

Results Analysis - Montalbano



Vieniqui or *Tantotempo*?

snapshot

snapshot

Furbo or *Buonissimo*?

snapshot

snapshot

Conclusion

How influenced is the CNNLSTM network by subject variability?

- Network's limitation when presented with data of new subjects

  → Train *Snapture* on data from all participants

  → Satisfactory solution in the context of our work (we analyzed architecture rather instead of comparing to benchmark)

Conclusion

How efficient is the CNNLSTM network at learning co-speech gestures?

- Network's limitation
  - Main issues: similar motion patterns, missing hand details (pre-processing)

- Our *Snapture* architecture achieved superior results to CNNLSTM due to the static channel

Conclusion

How to identify the peak of the gesture and extract the handshape using RGB data only?

- Our algorithm worked well under the Kendon's model assumption.
- Independent of the hand
- Future work: additional channels (facial features, speech, body pose)

  → simple (modularity of our architecture)

Conclusion

How to regulate the integration of the hand details into a dynamic gesture recognition system?

- *Snapture* $_{thold}$ bypassed the *blurriness issue and improved* the performance of *Snapture even further.*
- Future work: improve robustness of threshold values.

Thesis Contribution

- Thorough analysis of Tsironi GRIT and Chalearn Montalbano gestures
- New architecture proposal: *Snapture* (code available soon).
  - Emblematic gestures & co-speech domains.
- New algorithm based on SSIM for analyzing a gesture's motion profile and identifying its pause.
- Montalbano temporal segmentations (publicly available soon).
- A concrete step to support an immersive HRI scenarios without the lab restrictions.

The End

Thank you for your attention.

Any question?

Literature

- Jorge Alberto Marcial Basilio, Gualberto Aguilar Torres, Gabriel Sanchez Perez, L. Karina Toscano Medina, and Hector M. Perez Meana. Explicit image detection using YCbCr space color model as skin detection. In Proceedings of the 2011 *American Conference on Applied Mathematics and the 5th WSEAS International Conference on Computer Engineering and Applications*, AMERICAN-MATH'11/CEA'11, page 123-128, Stevens Point, Wisconsin, USA, 2011. World Scientific and Engineering Academy and Society (WSEAS).
- Sergio Escalera, Xavier Baro, Jordi Gonzalez, Miguel A. Bautista, Meysam Madadi, Miguel Reyes, Victor Ponce-Lopez, Hugo J. Escalante, Jamie Shotton, and Isabelle Guyon. Chalearn looking at people challenge 2014: Dataset and results. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision* - ECCV 2014 Workshops, pages 459-473, Cham, 2015. Springer International Publishing.
- Adam Kendon. Gesticulation and Speech: *Two Aspects of the Process of Utterance*, pages 207-228. De Gruyter Mouton, 2011.
- E. Tsironi, P. Barros, and S. Wermter. Gesture recognition with a convolutional long short-term memory recurrent neural network. In *ESANN*, 2016.
- Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600-612, 2004.