Snapture - a Hybrid Hand Gesture Recognition System

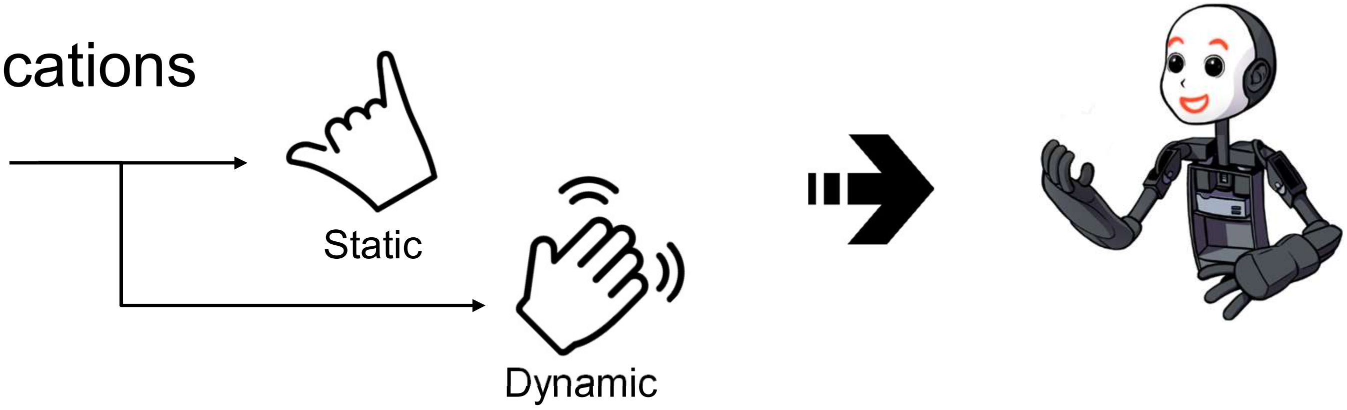
Hassan Ali

KNOWLEDGE
TECHNOLOGY

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

Motivation

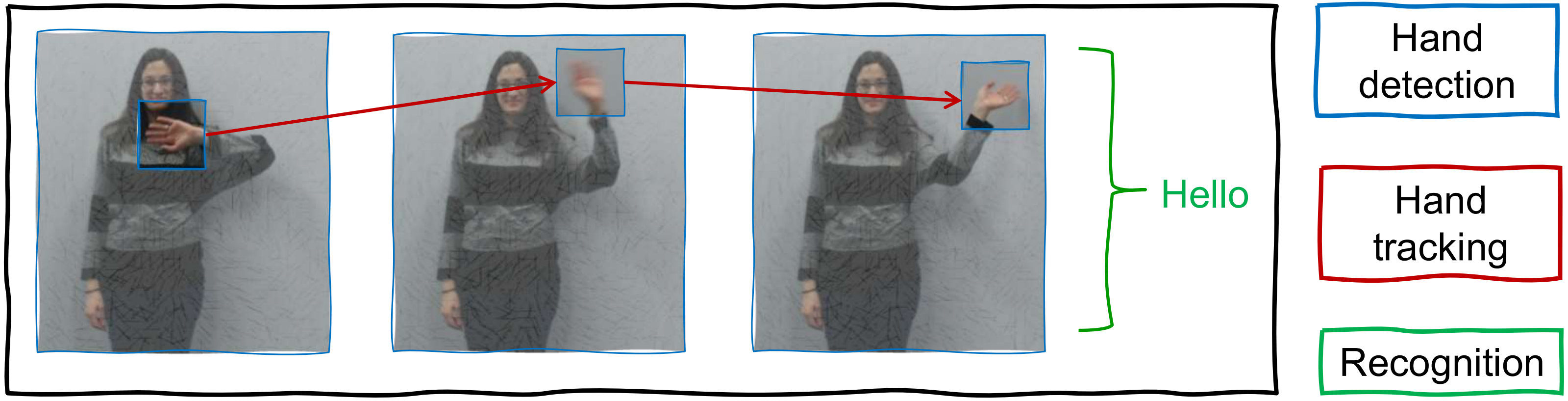
- Hand gesture applications
- Gesture taxonomy



Static

Dynamic

- Vision-based systems



Hello

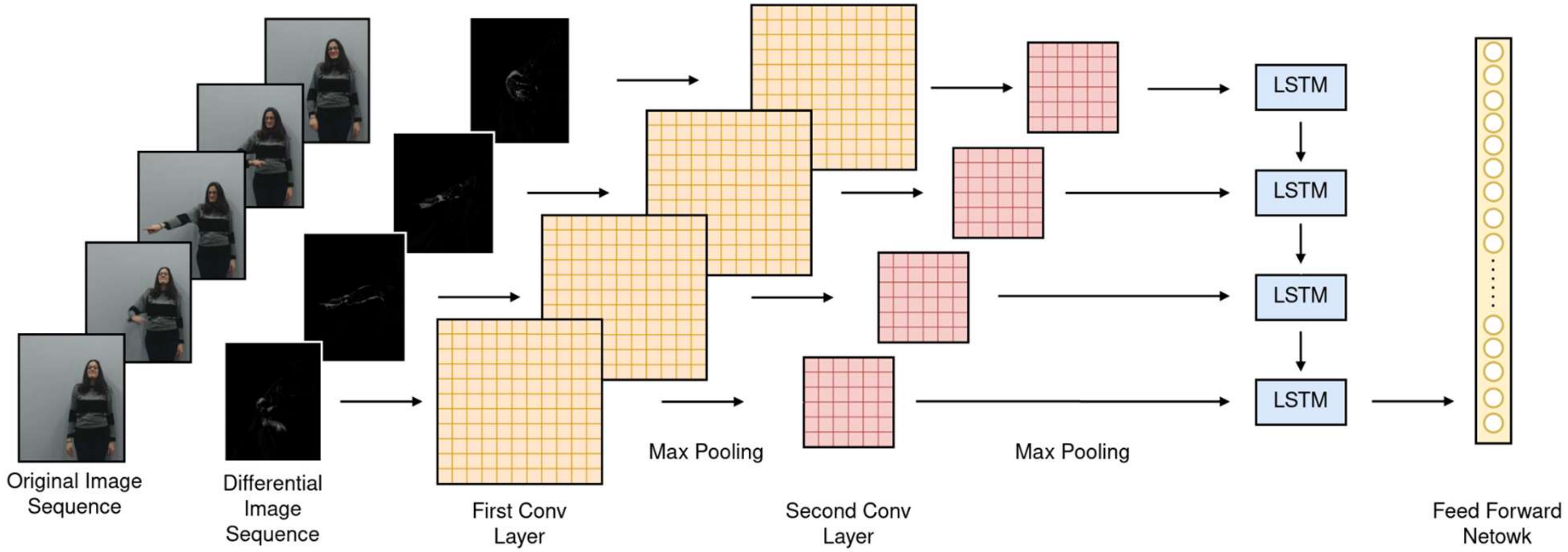
Hand detection

Hand tracking

Recognition

CNNLSTM

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



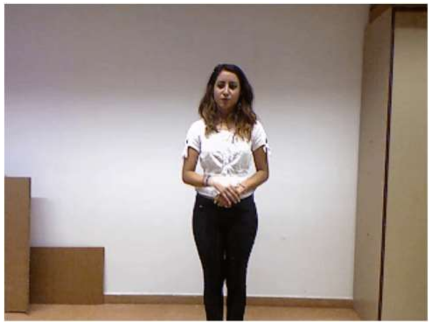
1. How influenced is the CNNLSTM network by subject variability?

Prendere

Messi D’accordo

Turn Left

No

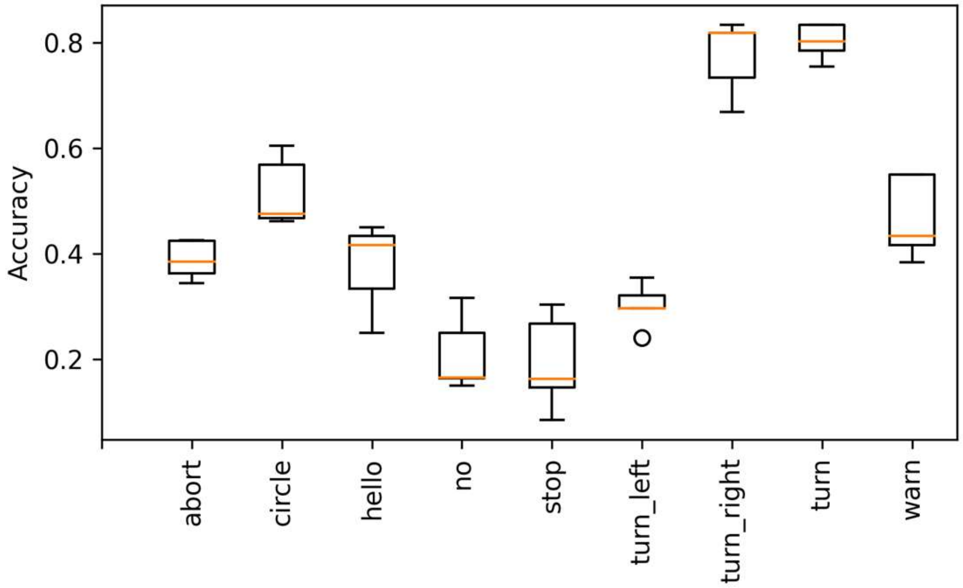


Montalbano

GRIT

CNNLSTM – Subject Variability

- Experiment using the GRIT dataset
- Evaluate on unseen subjects (Leave-one-out approach)
- Low accuracy for most classes
- **Consequence**: train on data of all subjects.



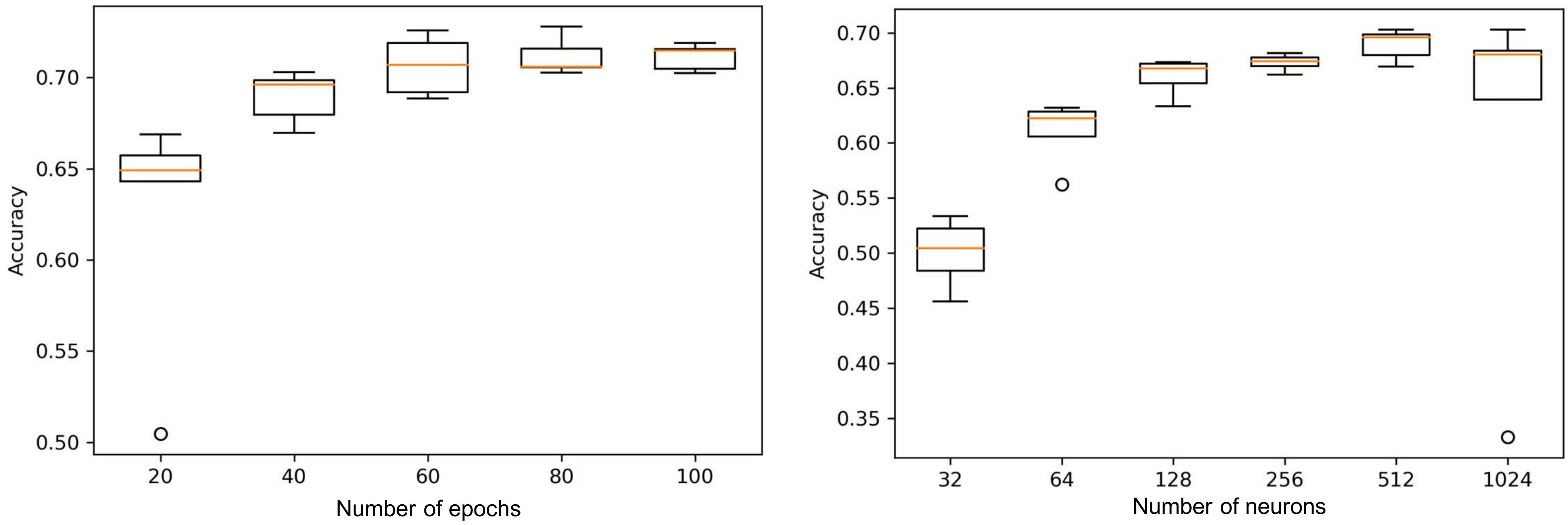
Accuracy per class, avg. over all subjects, avg. of 5 trials

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



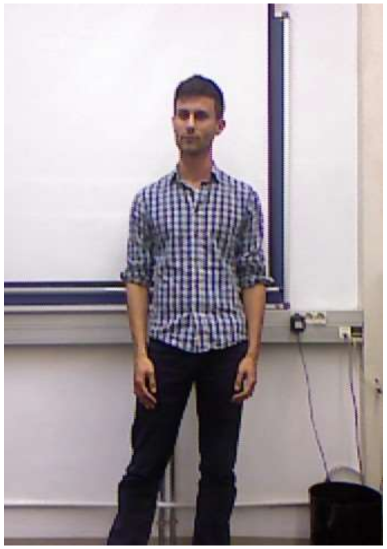
Results on the Montalbano dataset (avg. of 5 trials)

CNNLSTM – Upscaling

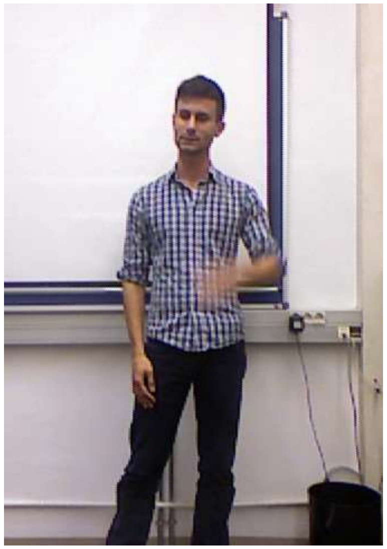
- Two Issues (Hypotheses):
  - *It is challenging for the CNNLSTM model to distinguish classes with similar movement patterns.*
  - *It is challenging for the CNNLSTM model to distinguish subtle movements done at the peak of the gesture.*
- **Solution:** *Snapture* architecture

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

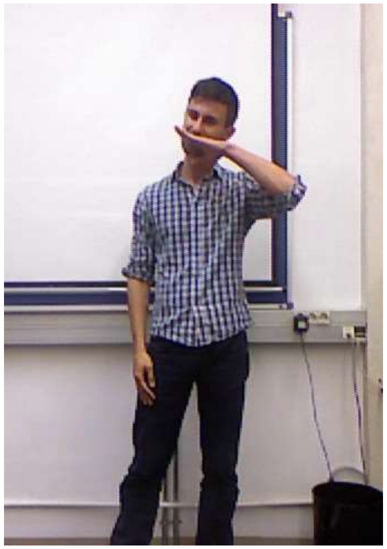
Gesture phases (Kendon), ex: *cosatifarei*



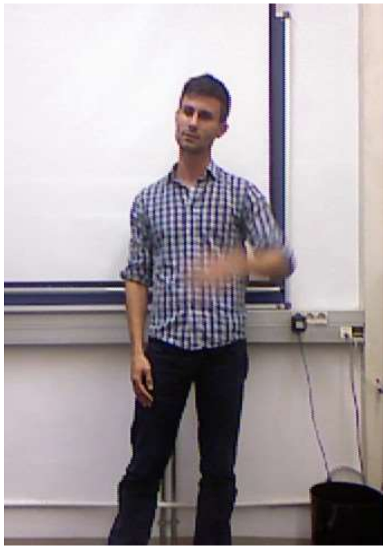
1. Rest position



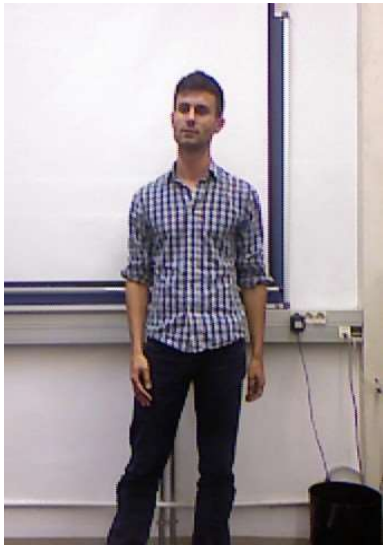
2. Pre-stroke



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

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

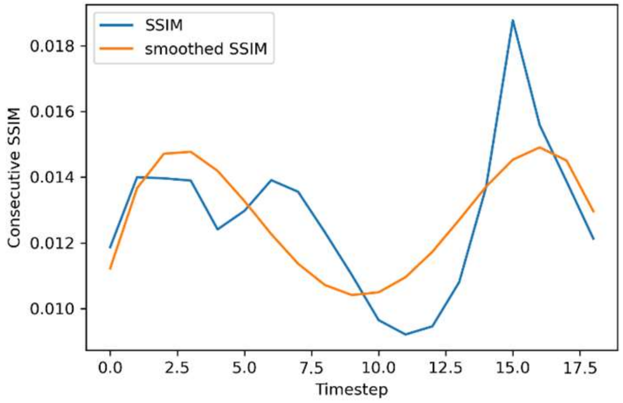
$$inverted_{SSIM} = 1 - \sum SSIM(\Delta_i, \Delta_{i-1})$$

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

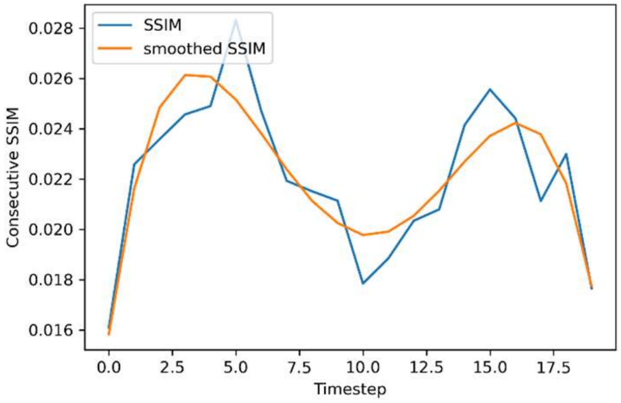
μ: avg. intensity,σ^2 :variance
C_1, C_1: stability constants
Δ_i, Δ_{i-1}: differential images
I_{i-1}, I_i, I_{i+1}: original frames

Motion Profile

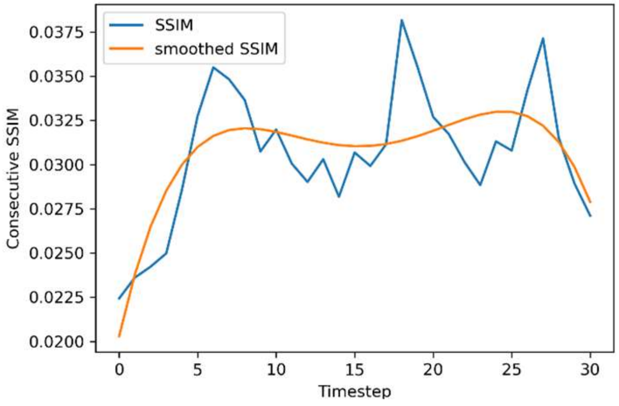
- Tsironi GRIT data:
  - *paused-gestures:* the arm remains briefly in a fixed position at the peak
  - or gestures with *repeated-pattern:* include a motion pattern, usually circular



Stop



Turn Left

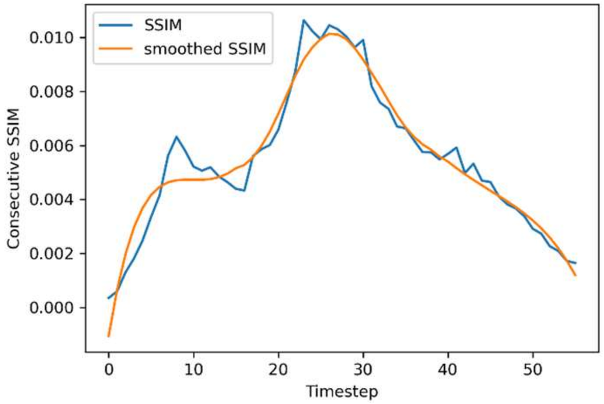


Turn

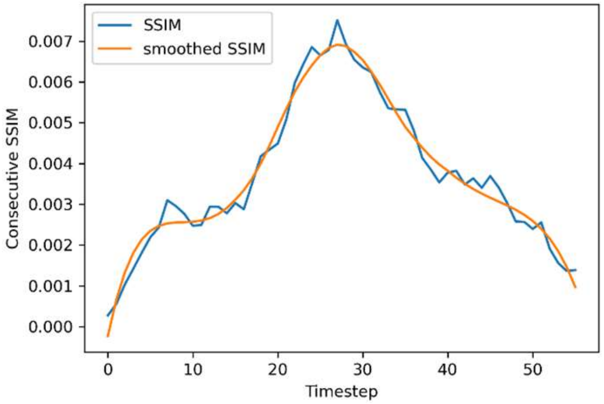
Motion Profile

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

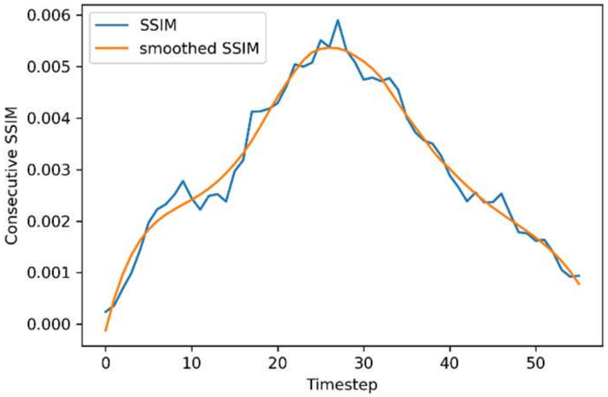
  → Peak around the mean of the sequence length



Vattene



Vieniqui

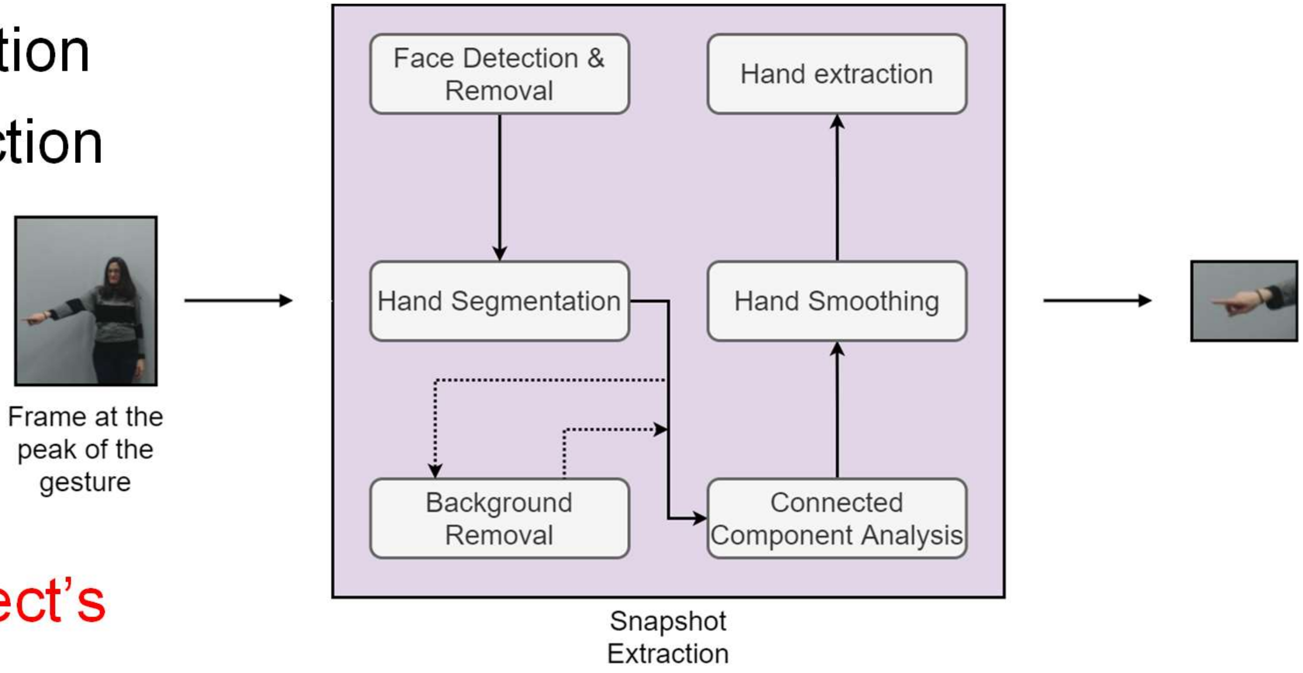


Ok

Snapture Architecture

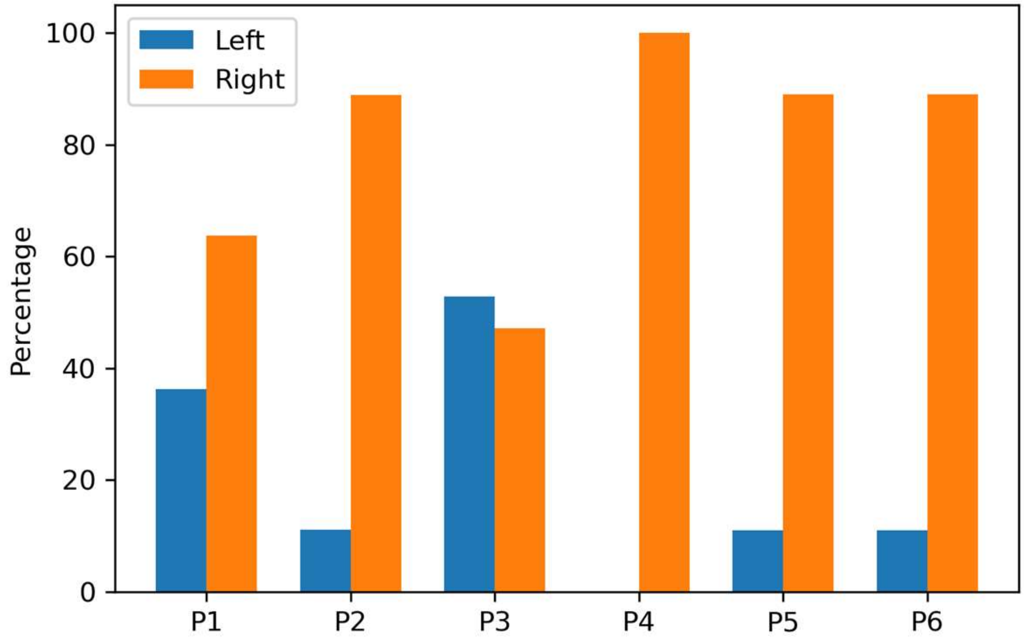
Static Channel

- Gesture Peak Detection
- Gesture Peak Extraction



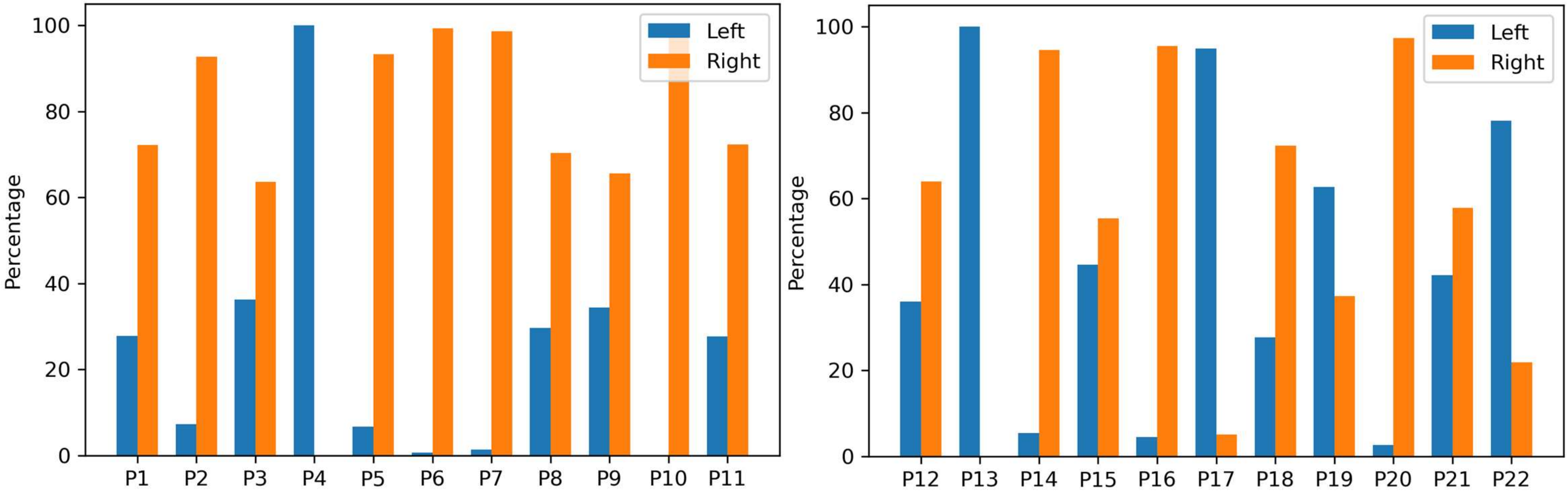
- Independent of subject's dominant hand

Hand Preference



The ratio of the gestures performed using the left and right hand per subject in the GRIT dataset.

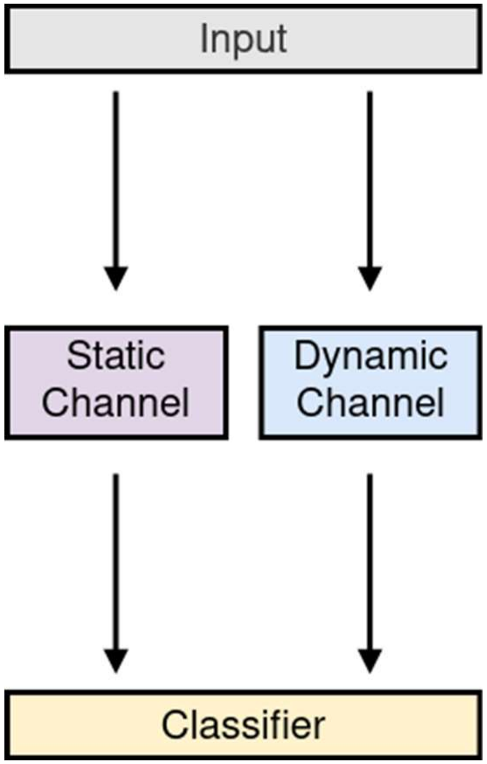
Hand Preference



The ratio of the gestures performed using the left and right hand for 22 subjects of the Montalbano dataset.

Snapture Architecture

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



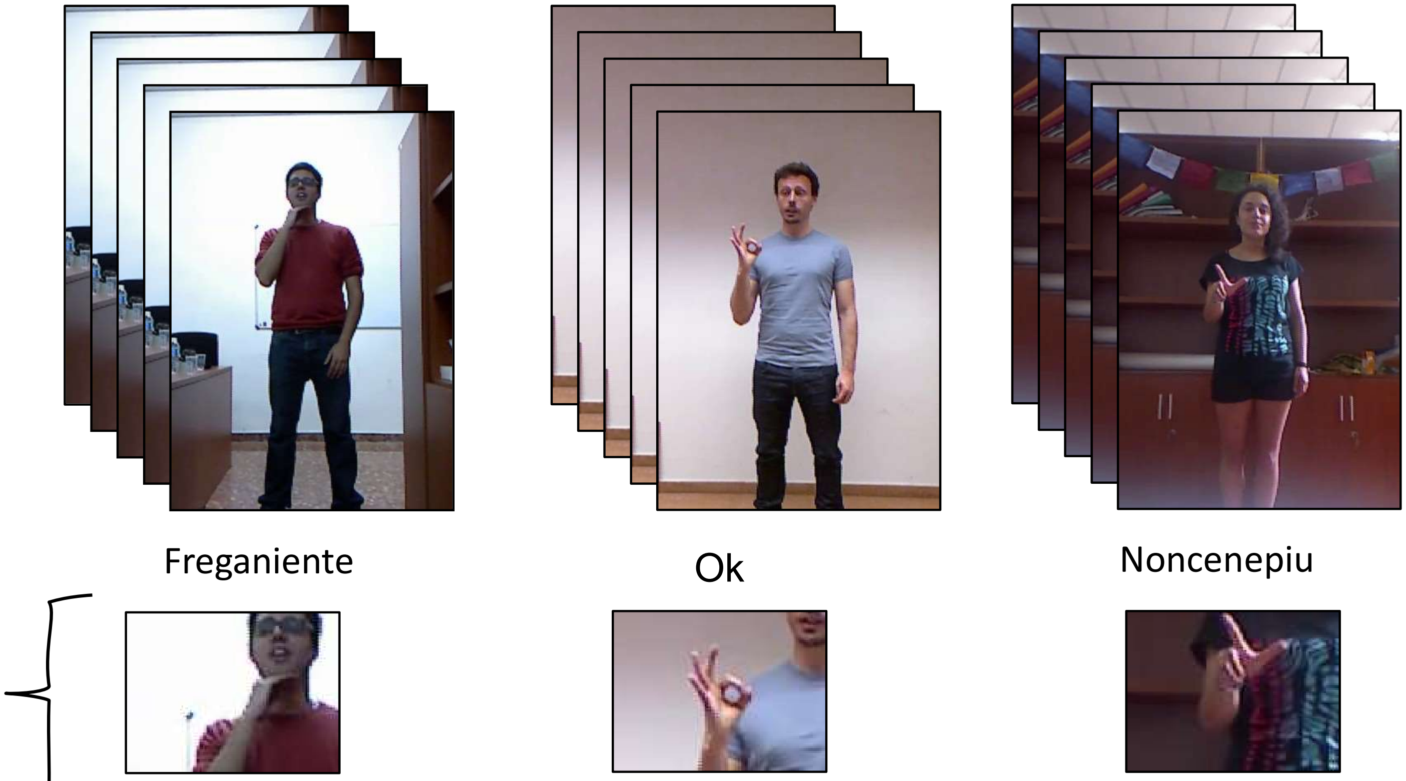
Snapture Architecture

Dynamic Channel

sequence-level CNNLSTM

Static Channel

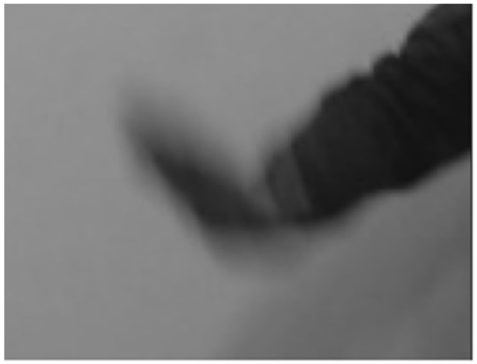
a *snapshot* at the peak, CNN-based



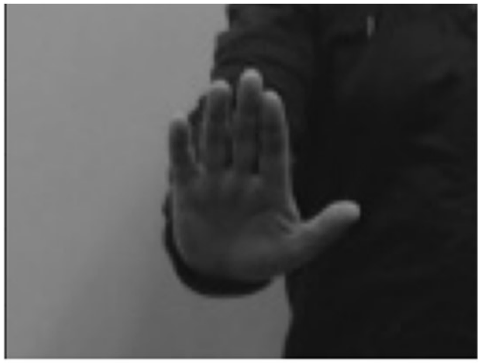
4. How to regulate the integration of the hand details into a dynamic gesture recognition system?

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

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



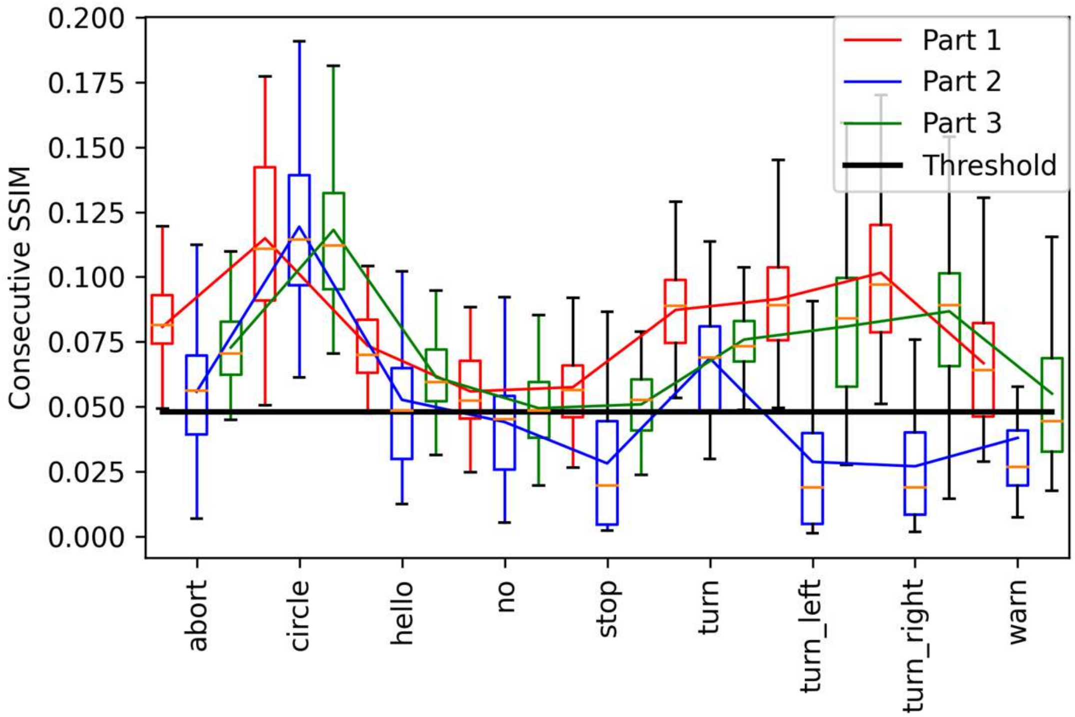
Circle



Stop

Regulating the static channel

- Approx. only 44% of the GRIT samples include a pause

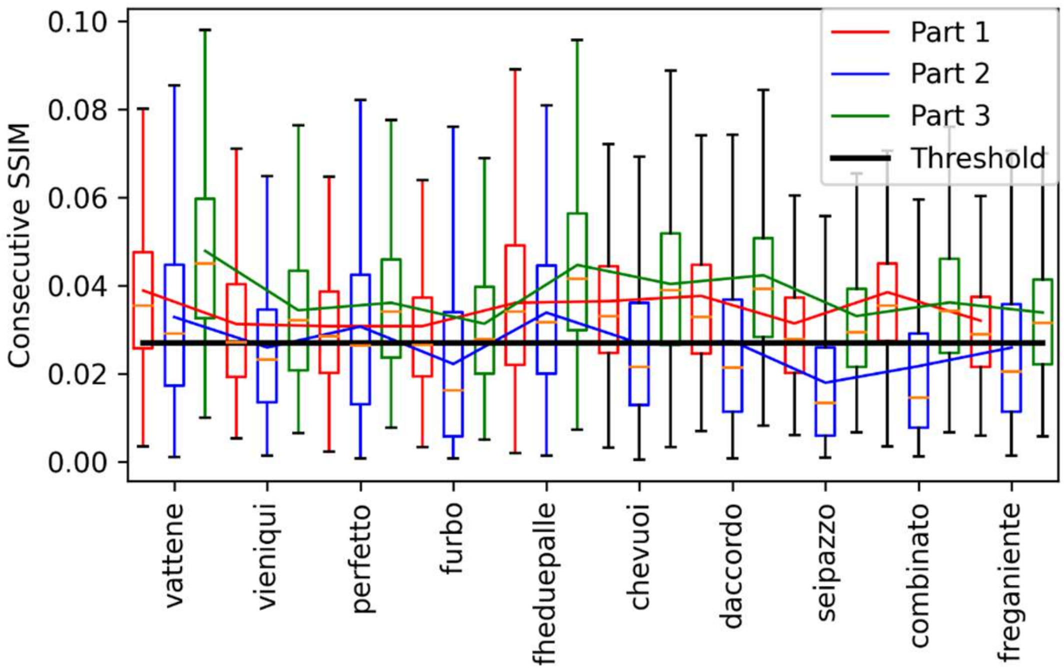


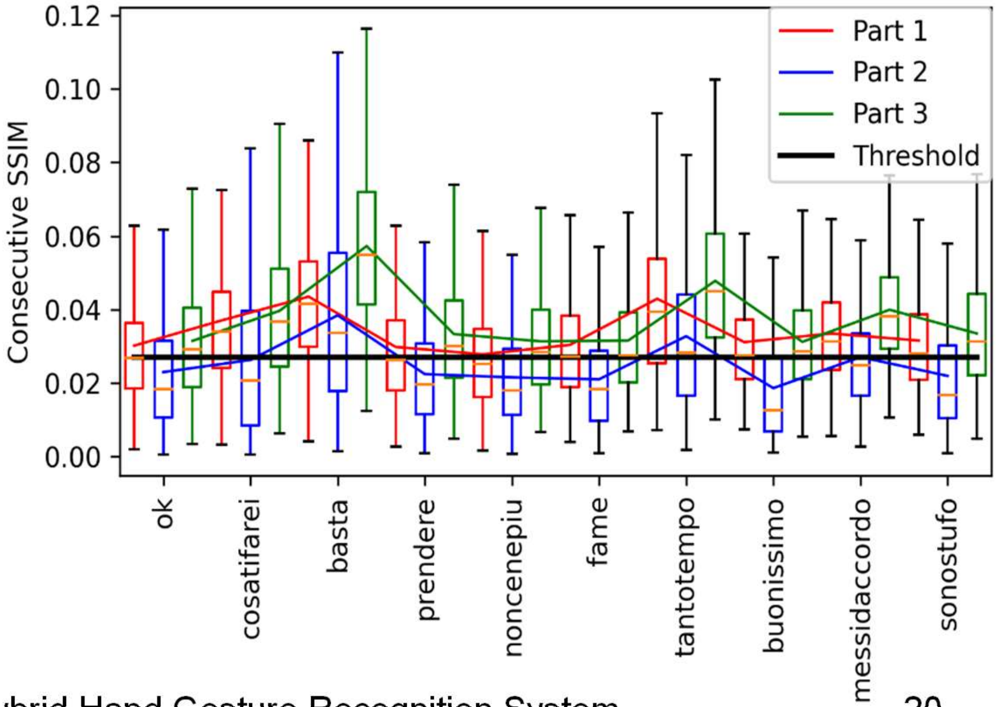
Part 1: rest to pre-stroke,
Part 2: pre-stroke to post-stroke,
Part 3: post-stroke to rest

Regulating the static channel

- Co-speech movements include more pause at the peak (≈70% 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 $_{\text{thold}}$	**0.926** (0.008)	0.913 (0.012)	125.156 (1.117)

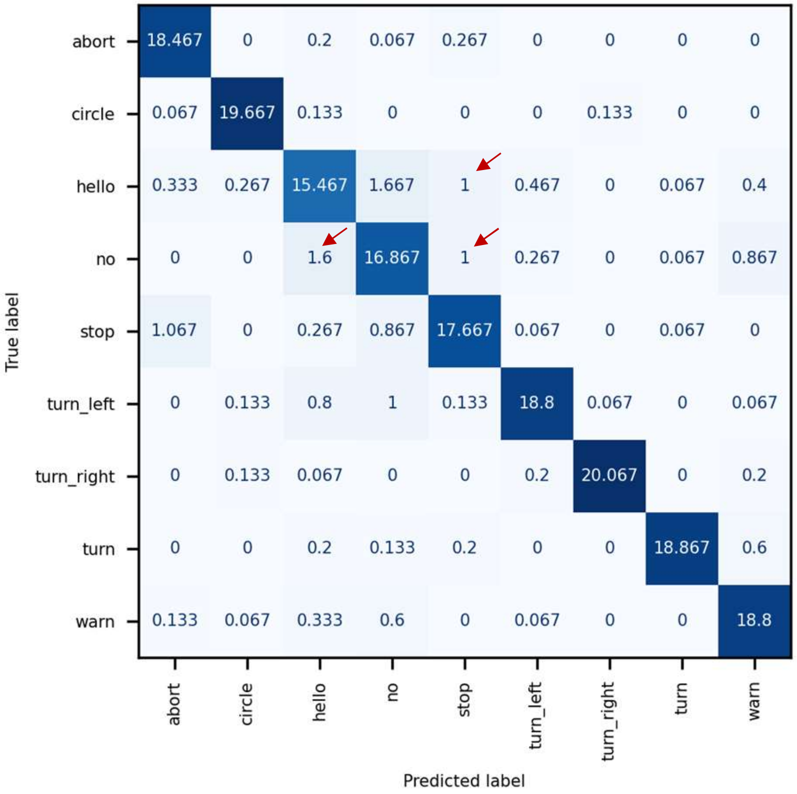
***In seconds.**

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 $_{\text{thold}}$	**0.77** (0.008)	**0.772** (0.007)	744.953 (0.724)

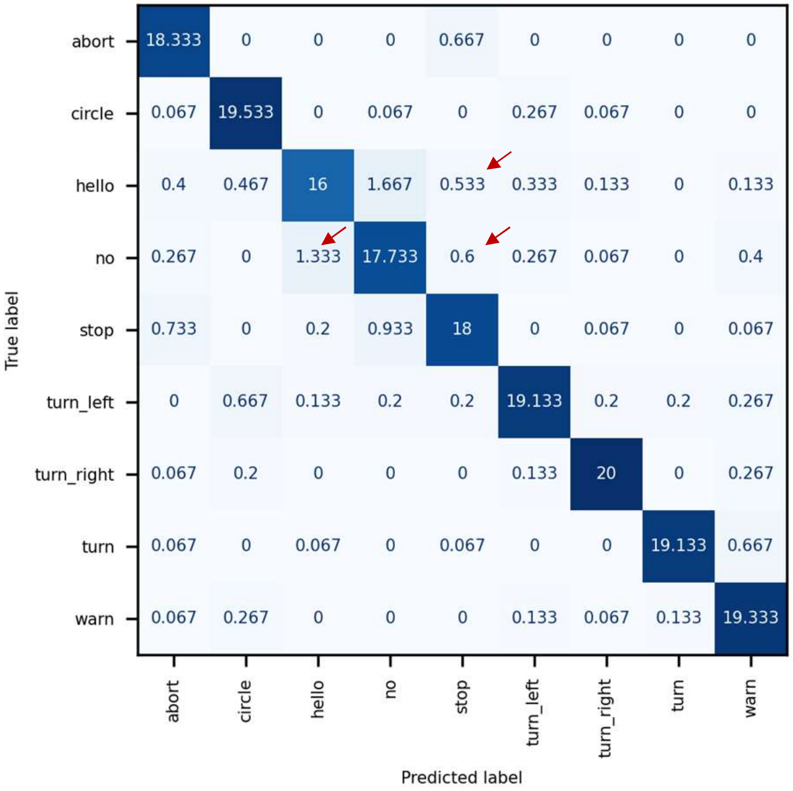
***In minutes.**

Results Analysis - GRIT



	abort	circle	hello	no	stop	turn_left	turn_right	turn	warn
abort	18.467	0	0.2	0.067	0.267	0	0	0	0
circle	0.067	19.667	0.133	0	0	0	0.133	0	0
hello	0.333	0.267	15.467	1.667	1	0.467	0	0.067	0.4
no	0	0	1.6	16.867	1	0.267	0	0.067	0.867
stop	1.067	0	0.267	0.867	17.667	0.067	0	0.067	0
turn_left	0	0.133	0.8	1	0.133	18.8	0.067	0	0.067
turn_right	0	0.133	0.067	0	0	0.2	20.067	0	0.2
turn	0	0	0.2	0.133	0.2	0	0	18.867	0.6
warn	0.133	0.067	0.333	0.6	0	0.067	0	0	18.8

CNNLSTM
(avg. 5 trials)

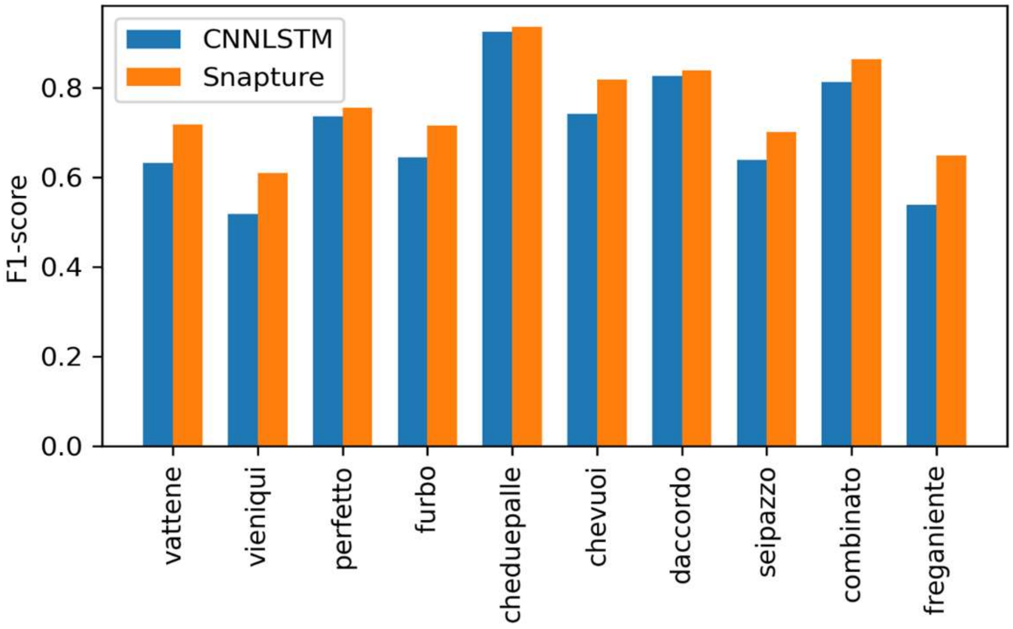


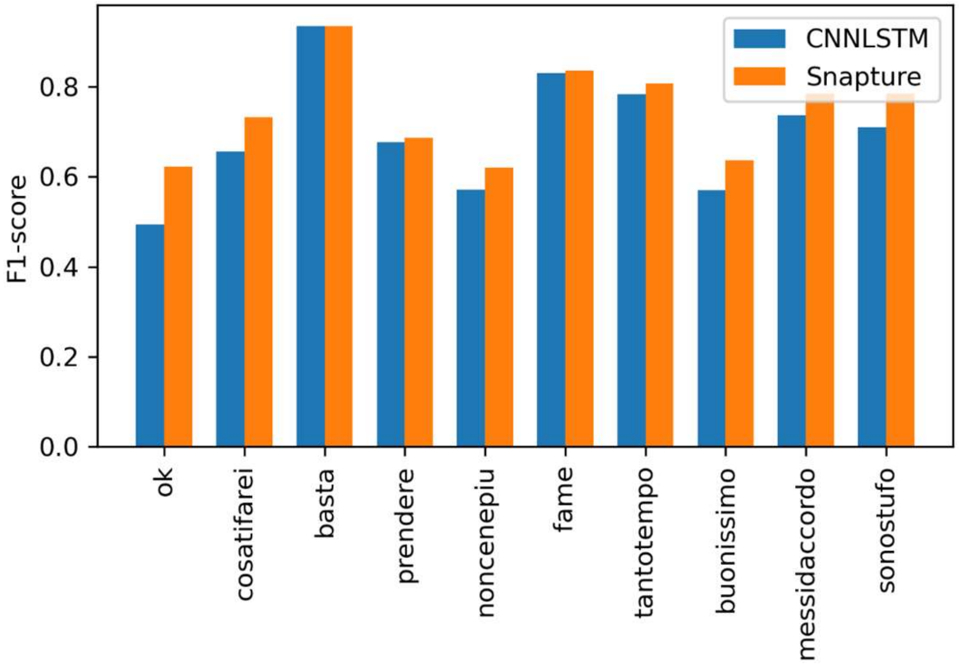
	abort	circle	hello	no	stop	turn_left	turn_right	turn	warn
abort	18.333	0	0	0	0.667	0	0	0	0
circle	0.067	19.533	0	0.067	0	0.267	0.067	0	0
hello	0.4	0.467	16	1.667	0.533	0.333	0.133	0	0.133
no	0.267	0	1.333	17.733	0.6	0.267	0.067	0	0.4
stop	0.733	0	0.2	0.933	18	0	0.067	0	0.067
turn_left	0	0.667	0.133	0.2	0.2	19.133	0.2	0.2	0.267
turn_right	0.067	0.2	0	0	0	0.133	20	0	0.267
turn	0.067	0	0.067	0	0.067	0	0	19.133	0.667
warn	0.067	0.267	0	0	0	0.133	0.067	0.133	19.333

Snapture
(avg. 5 trials)

Results Analysis - Montalbano

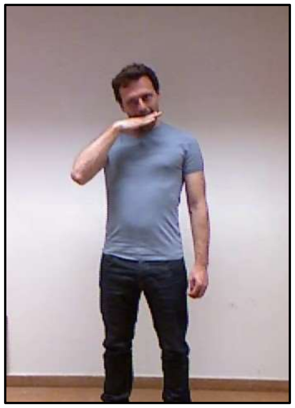
- *Snapture:* superior results on all classes except for *Basta*
- Boosted F1-score for unique handshape classes (ex: *Ok*)



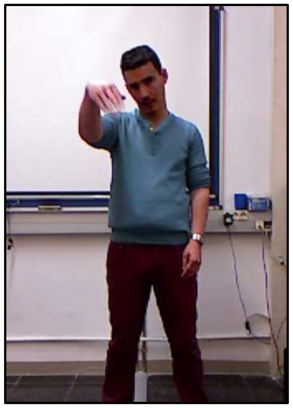


Results Analysis - Montalbano

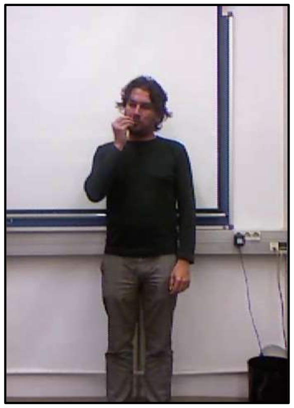
- *Snapture* boosts the classification of indistinctive movements.



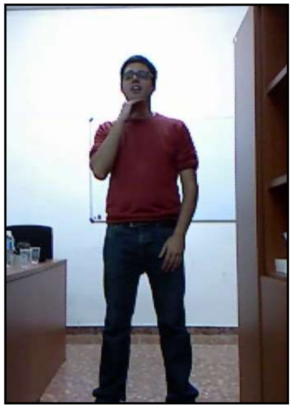
Cosatifarei



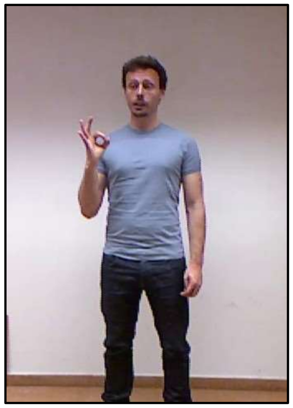
Vattene



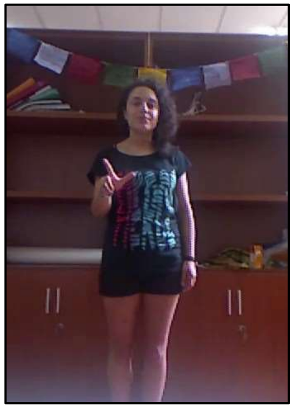
Perfetto



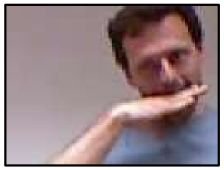
Freganiente



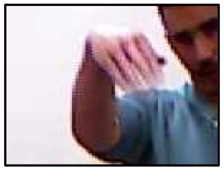
Ok



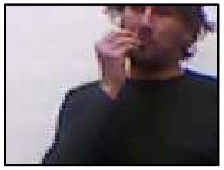
Noncenepiu



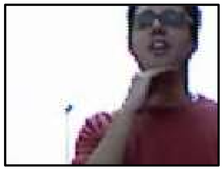
snapshot



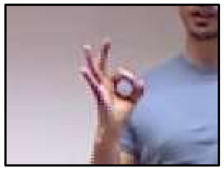
snapshot



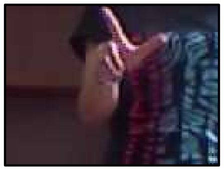
snapshot



snapshot



snapshot



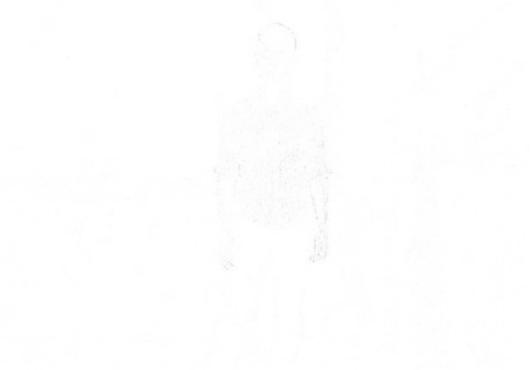
snapshot

Results Analysis - Montalbano

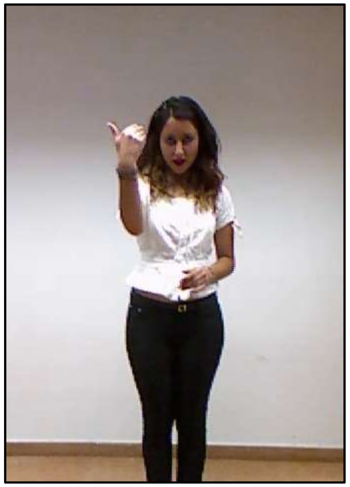
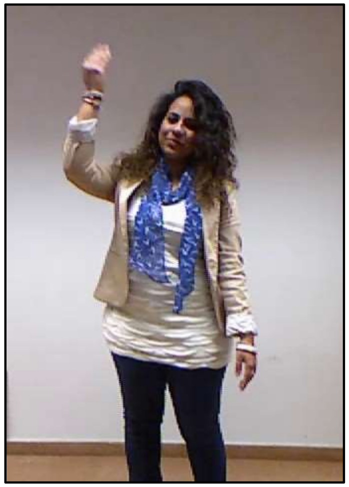
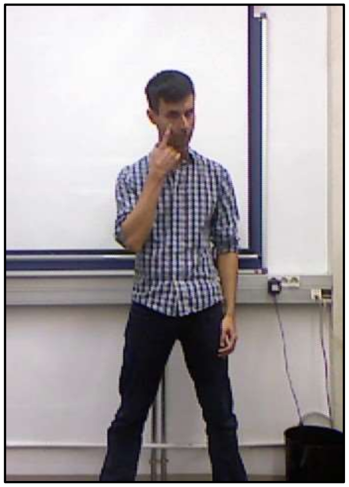
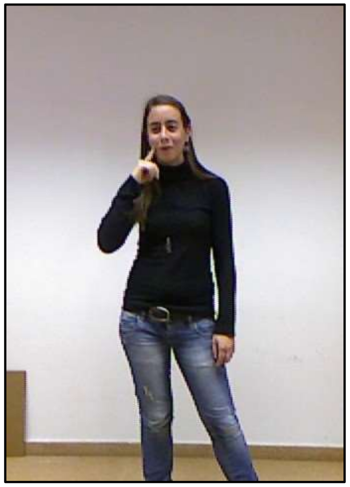
- *Snapture* boosts the classification of subtle movements.

Basta (explicit hand movement

Sonostufo (subtle hand movement)

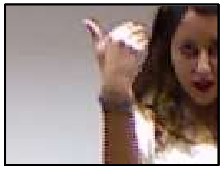


Limitations

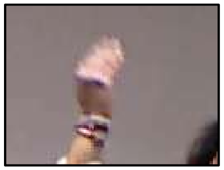





Vieniqui or Tantotempo?

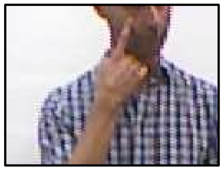
Furbo or Buonissimo?



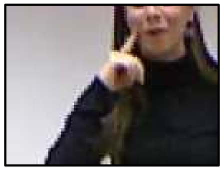
snapshot



snapshot

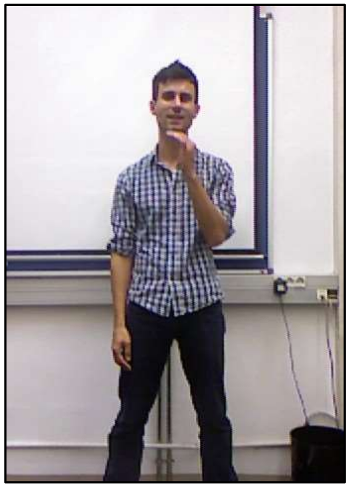


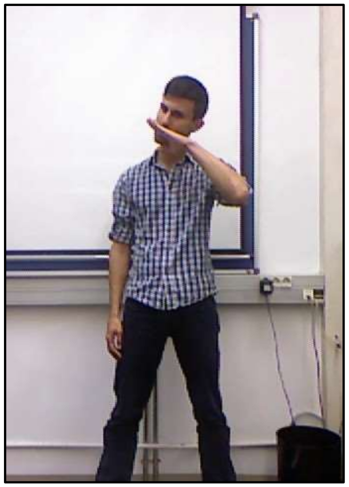
snapshot

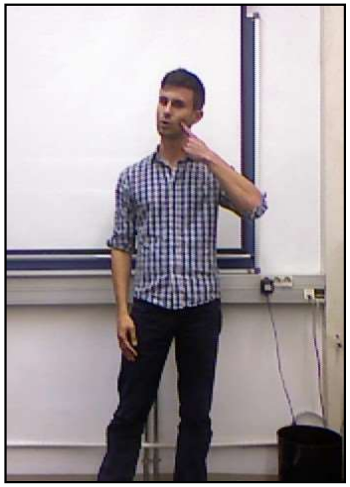


snapshot

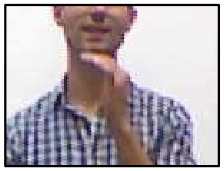
Limitations



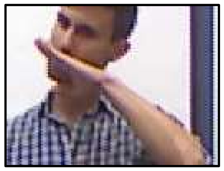




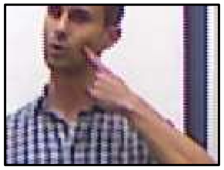
Open palm or an extended index finger?



snapshot

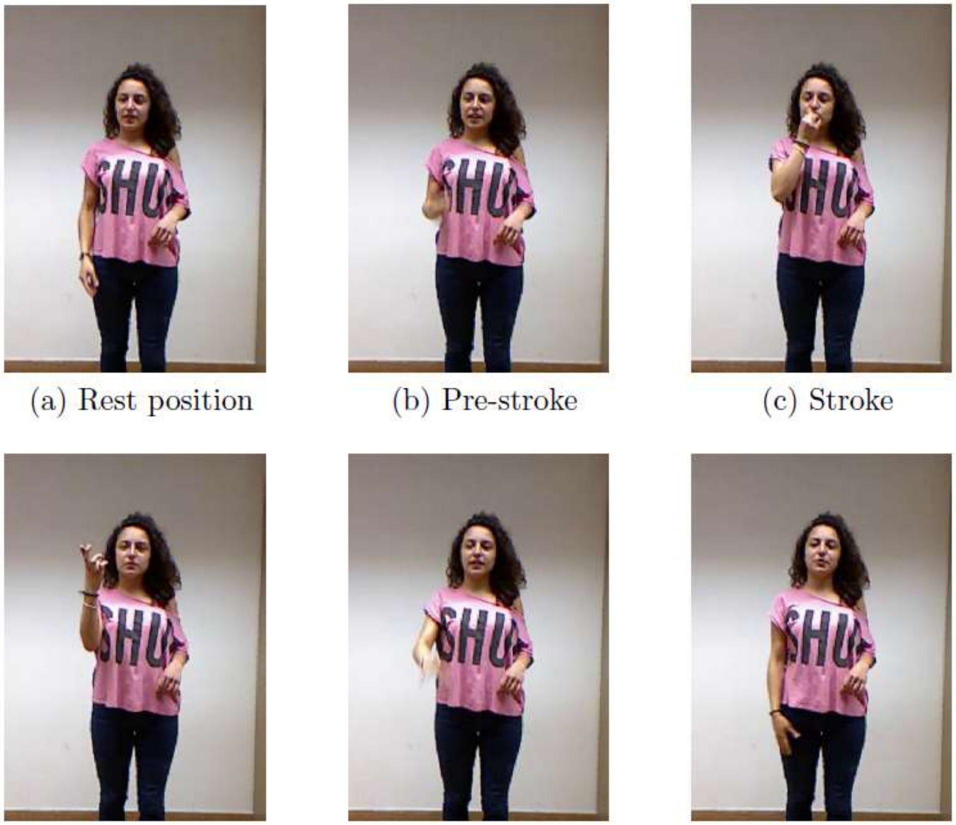


snapshot



snapshot

Limitations



(a) Rest position

(b) Pre-stroke

(c) Stroke

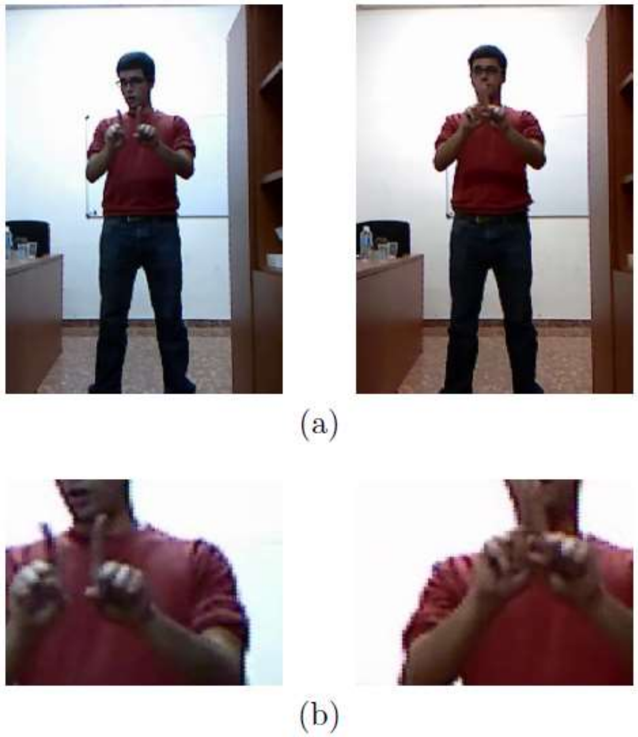
(d) Stroke

(e) Pre-stroke

(f) Rest position

Perfetto (fuzzy stroke phase)

Limitations



(a)

(b)

D’accordo (different snapshot of the same gesture)

Conclusion

- Our *Snapture* architecture achieved superior results to CNNLSTM especially in the context of co-speech gestures.

  → Similar motion patterns, missing hand details (pre-processing)

  → Independent of dominant hand

- *Snapture* $_{\textit{thold}}$ bypassed the *blurriness* issue and provided performance boost.

  → New algorithm based on SSIM for analyzing a gesture's motion/pause.

- Code + Montalbano temporal segmentations available soon on: https://github.com/sano-90/snapture

Future Work

- additional channels (facial features, speech, body pose)
  → simple (modularity of our architecture)
- Improve robustness of threshold values.
- A concrete step to support an immersive HRI scenarios without the lab restrictions.
  → gestures acted “in the wild”
  → day-to-day human environments
  → No assumption of a hand dominance

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.