Using video-based pose estimation for automated analysis of interaction

UNIVERSITY LECTURER SATU SAALASTI

DEPARTMENT OF PSYCHOLOGY AND LOGOPEDICS

UNIVERSITY OF HELSINKI

HELSINGIN YLIOPISTO

Gestural communication

- Gestural, embodied (also non-verbal or visual bodily) communication is central for everyday interaction, and precedes language acquisition in childhood (Filipi, 2009).

Gestures and speech

- For example, early joint attention, imitation and use of gestures promotes language acquisition and is connected to speech and language abilities in later childhood (De Schuymer, De Groote, Bayers, Striano & Roeyers, 2011; Tomasello & Farrar, 1986).

- Gestural and oral communication are often studied together.

Image: Unsplash

Inter-individual behavioral synchrony from motion

* Inter-individual motion synchrony is related to the quality of social interactions.

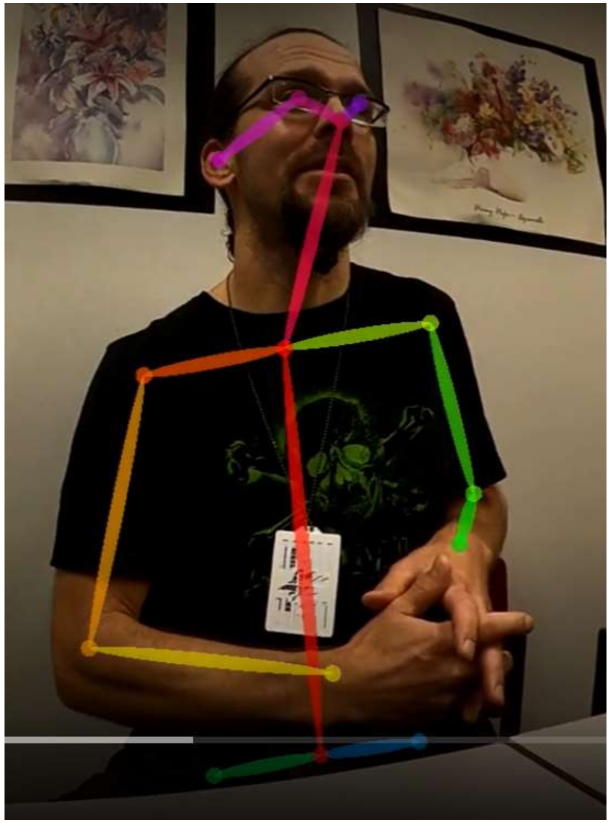
Laboratory settings with Kinect cameras:

- Leclère et al. (2016):
  - Neglect in mother–infant interaction
  - 10 extremely high-risk and 10 low-risk dyads using two-dimensional (2D) and three-dimensional (3D) sensors
- Lahnakoski et al. (2020)
  - Dyadic measures of joint orienting and distancing, synchrony and gaze behaviours during natural conversation and joint action tasks
  - Measures were sensitive for the subjective quality of social interactions

Automatic pose estimation from real-life videos

- OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. (Cao et al., 2018)
- Recognize the poses of multiple persons on a video.
- Has been used to
  - estimate the attention level of participants in a multi-person interaction scene (Komiya, Saitoh and Shimada, 2018), and
  - to study rapport in groups of 3–4 persons from nonverbal behavior, including facial expressions, hand motion, gaze, speaker turns and speech prosody (Müller, Huang & Bulling, 2018).
  - Predict autism (Kojovic, 2021)

Example video



Conversational self- and other-initiated repair

* ***Intersubjective Understanding in Atypical Interaction*** Academy of Finland Project

***** Principal investigator prof Minna Laakso, University of Helsinki

* In natural conversation, people encounter problems in hearing, speaking and understanding each other.

* The problems, i.e. the trouble source, can be resolved with conversational repair practices

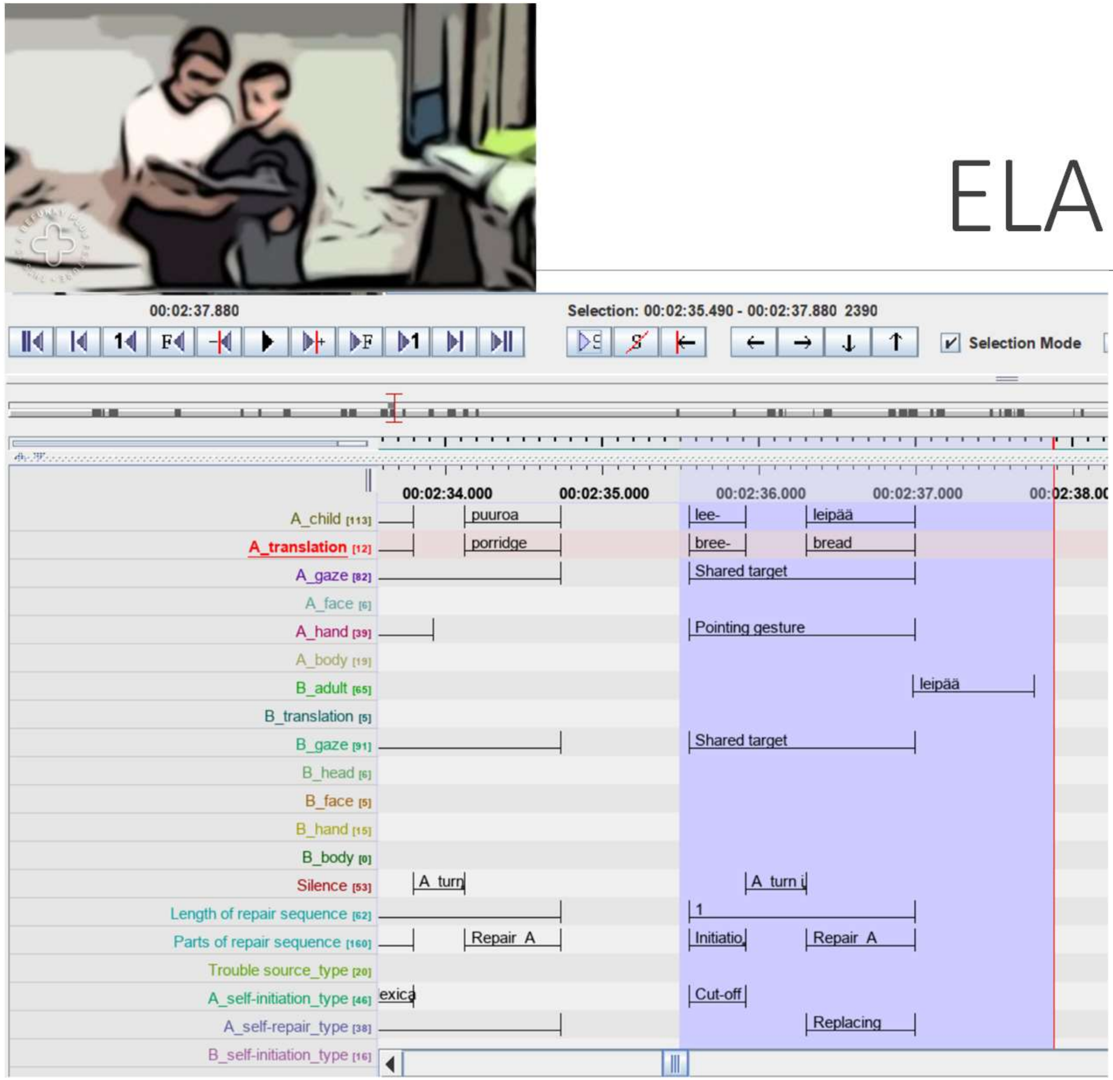
- short and transient operations used by both speakers and the recipients of trouble (Schegloff, Jefferson & Sacks, 1977).

* We study the conversational repair organization in conversational interactions involving participants with disorders on

- cognitive (dementia, autism),
- linguistic (adult aphasia and developmental language disorder),
- motor (dysarthria), and
- sensory-perceptual (hearing impairment) levels of human performance.

* Embodied practices during conversational repair – scoping review (Saalasti, Pajo, Fox, Laakso (in preparation)).

ELAN ANNOTATIONS



00:02:37.880

Selection: 00:02:35.490 - 00:02:37.880 2390

☑ Selection Mode

00:02:34.000 00:02:35.000 00:02:36.000 00:02:37.000 00:02:38.00

A_child [113] puuroa lee- leipää

A_translation [12] porridge bree- bread

A_gaze [82] Shared target

A_face [6]

A_hand [39] Pointing gesture

A_body [19]

B_adult [65] leipää

B_translation [5]

B_gaze [91] Shared target

B_head [6]

B_face [5]

B_hand [15]

B_body [0]

Silence [53] A turn A turn i

Length of repair sequence [62] 1

Parts of repair sequence [160] Repair A Initiatio Repair A

Trouble source_type [20]

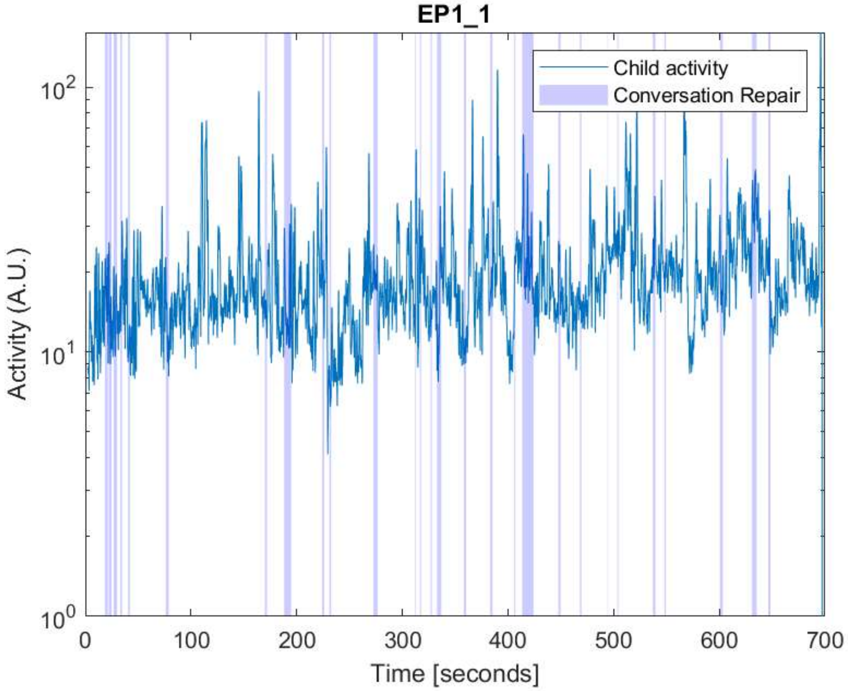
A_self-initiation_type [46] exica Cut-off

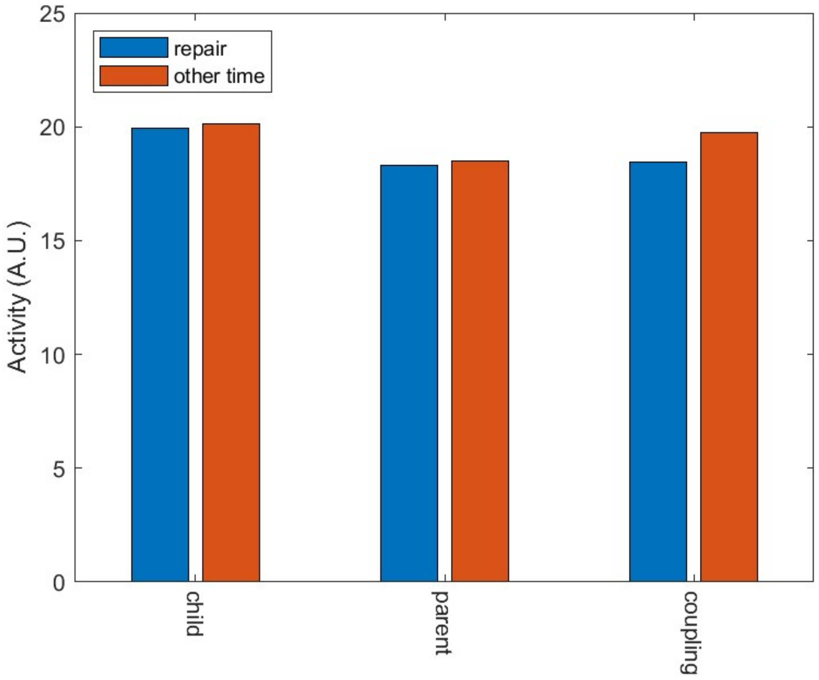
A_self-repair_type [38] Replacing

B_self-initiation_type [16]

Brugman, H. & A. Russell (2004)
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A
Sloetjes, H. (2006).

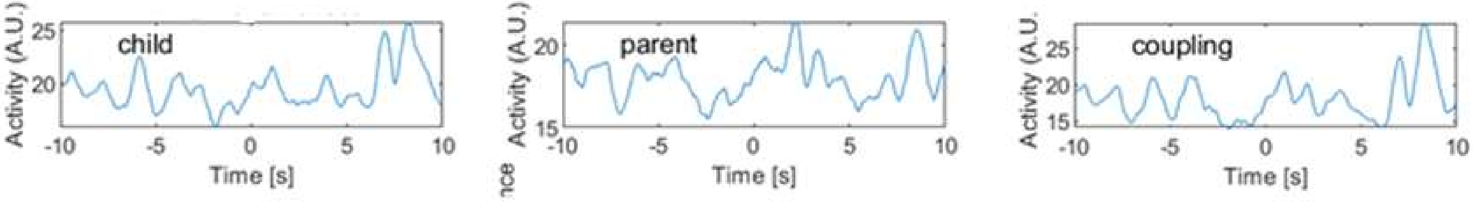
ELAN AND MOTION TRACKING DATA COMBINED



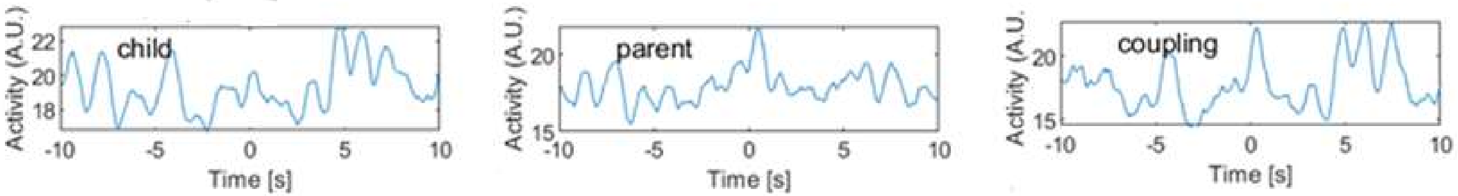


Total movement activity

REPAIR ONSET



REPAIR OFFSET



Sharing data?

* Hundreds of hours of video data

* New data could be collected with new consent forms where data sharing specified.

* Old data → only ELAN annotations and motion time courses can be shared?

Key references:

Brugman, H. & A. Russell (2004). Annotating Multimedia / Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation

Cao, Z., Hidalgo, G, Simon, T., Wei, S.-E. & Sheikh, Y. (2018). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. arXiv:1812.08008

Filipi, A. (2009). *Toddler and parent interaction: The organisation of gaze, pointing and vocalisation* (Vol. 192). John Benjamins Publishing.

Kojovic, N., Natraj, S., Mohanty, S. P., Maillart, T., & Schaer, M. (2021). Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Scientific Reports*, *11*(1), 1-10.

Lahnakoski, J. M., Forbes, P. A., McCall, C., & Schilbach, L. (2020). Unobtrusive tracking of interpersonal orienting and distance predicts the subjective quality of social interactions. *Royal Society open science*, *7*(8), 191815.

Müller, P., Huang, M. X., & Bulling, A. (2018, March). Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces* (pp. 153-164).

Saalasti, Kauramäki, & Glerean (in preparation): Automatic computational tracking of motion in dyadic interaction.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In 5th international conference on language resources and evaluation (LREC 2006) (pp. 1556-1559).