I see you, do you see me?

Socially aware robots

Koen Hindriks

Social Intelligence

A social signal processing perspective:

The ability to

recognize & **express**

social cues, signals and social behaviors

Understanding Social Signals

"The ability to understand and manage social signals of a person we are communicating with is the core of social intelligence."

Social Psychology → Signal Processing → Social Information

Facial expressions tell us something about emotion

Facial recognition

Emotion classification based on extracted face

Source: Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, *27*(12).

Social Cues and Signals

- **Social cues** are the **observable features** of an agent that are biologically and physically determined, and these are transmitted as a short, discrete set of physical/physiological activity.

- **Social signals** are **meaningful interpretations of cues** in the form of attributions of an agent's mental state or attitudes. They depend on the situational **context** and which **combinations of cues** are used

- *Example:* signal empathy towards a friend by smiling at them

Social Cues

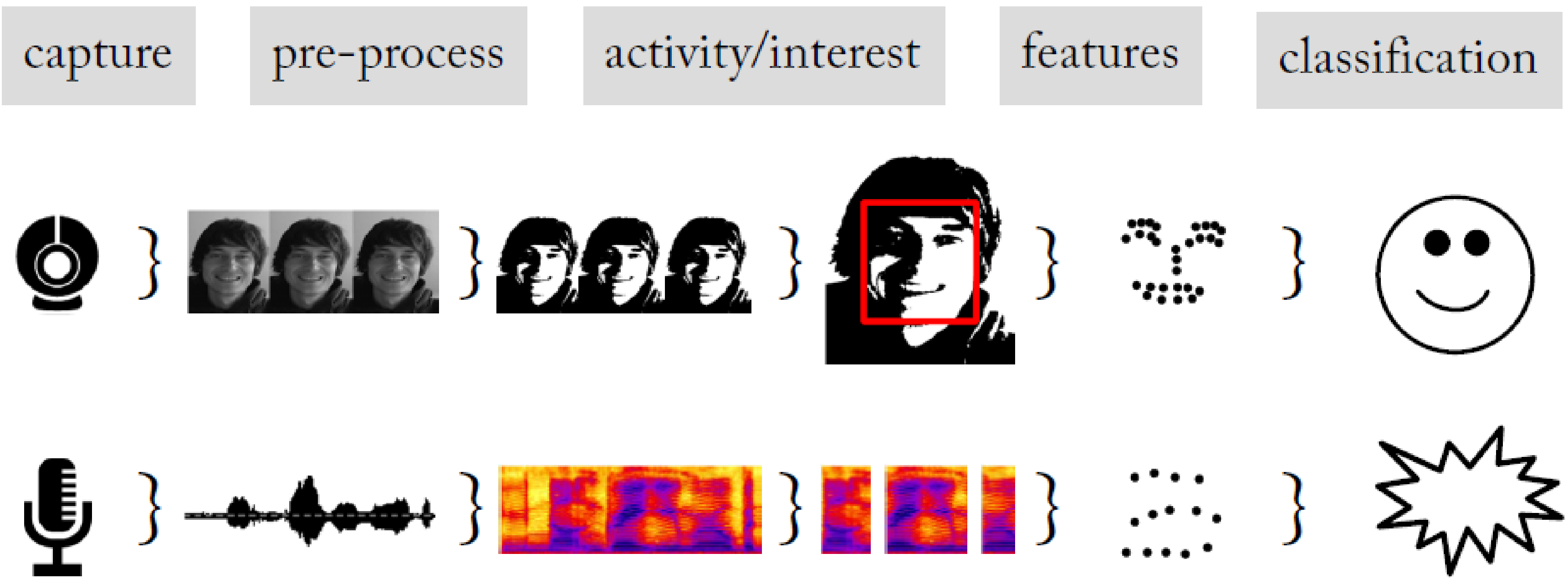
- Space and environment (proxemics)
- Physical appearance – height, body shape, skin and hair color, dress
- Facial expressions
- Gaze & head pose
- Postures and body movement
- Gestures (hand and arm)
- Vocal cues
- …

Signals: what information is conveyed?

Cues often accompany speech:

- **Attitudes**: emotion, cognitive attitudes, e.g., disbelief.
- **Manipulators**: towards the environment or oneself, e.g., holding a door open to signal that you should pass through it
- **Cultural emblems**: specific to cultural circle, e.g., “high five”.
- **Illustrators**: underlining information transmitted in other channels of communication, e.g., thumbs up.
- **Regulators**: affirm other communication partners or indicate turn-taking, e.g., gaze to signal someone should take turn.

Processing Pipeline



Visual and Audio Channels

Source: Wagner, 2015, Social Signal Processing (dissertation, p27)

I see you, do you see me?

Hypothesis:

When a human feels they are “being seen” by a robot, then they will perceive the robot as more *socially present*.

APPEARANCE

Physical Appearance – Clothing

- Most studies about the effects of clothing have used pictures. It has been hard to demonstrate effects of clothing in social interactions between humans.

- Is clothing only relevant for first impressions, but not for judgements over extended periods of interaction?

Clothing on Humans versus Robots

- Are the effects of clothing similar for humans and robots?

- How can we find out, i.e., establish that clothing for a particular aspect has a different effect for a human than a robot?

Differences?

Attributing sexual intent:

a lot of research on dress and sexual intent; dress on a robot such as Pepper perhaps will not lead to attributing sexual intent to it?

Clothing vs no clothing:

It is not clear how robots with and without clothing are perceived, which for robots is an interesting question to explore.

FACIAL EXPRESSIONS

Facial Expressions

Communicates:

- Affective state
- Intentions
- Personality
- Attractiveness
- Age
- Gender

Facial Expressions – FACS

- FACS provides an objective and comprehensive language for describing facial expressions
- FACS associates facial-expression changes with actions of the muscles that produce them.
- It defines:
  - 9 different action units (AUs) in the upper face,
  - 18 in the lower face,
  - 11 for head position,
  - 9 for eye position, and
  - 14 additional descriptors for miscellaneous actions

FACS – Exampe AUs

AU1 – Inner Brow Raiser
Frontalis, pars medialis

AU06 – Cheek Raiser
Orbicularisoculi, pars orbitalis

AU17 – Chin Raiser
Mentalis

AU10 – Upper Lip Raiser

Levator LabiiSuperioris, Caput infraorbitalis

AU45 – Blink
Relaxation of *Levator Palpebrae* and Contraction of *Orbicularis Oculi, Pars Palpebralis.*

From AUs to Displayed Emotions

***Displayed* Happiness = AU06 + AU12**

AU06 – Cheek Raiser

AU12 – Lip Corner Puller

AUs have intensity → can be used to derive emotion intensity

Exercise Feedback for People with Facial Paralysis

1. Raise eyebrows, holding for 5 seconds, repeating 10x.

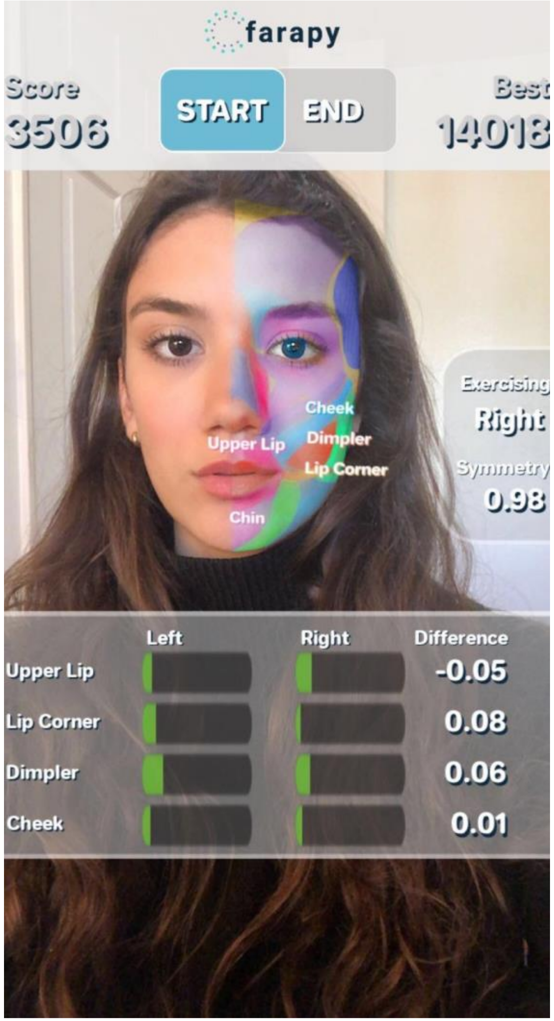
2. Wrinkle nose, holding for 5 seconds, repeating 10x.

6. Show lower teeth, holding for 5 seconds, repeating 10x.

3. Snarl, holding for 5 seconds, repeating 10x.

4. Smile, holding for 5 seconds, repeating 10x.

5. Pucker lips, holding for 5 seconds, repeating 10x.



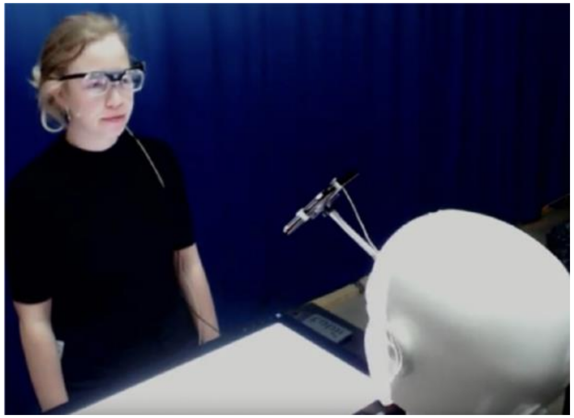
GAZE

MIT's Gaze 360

Background

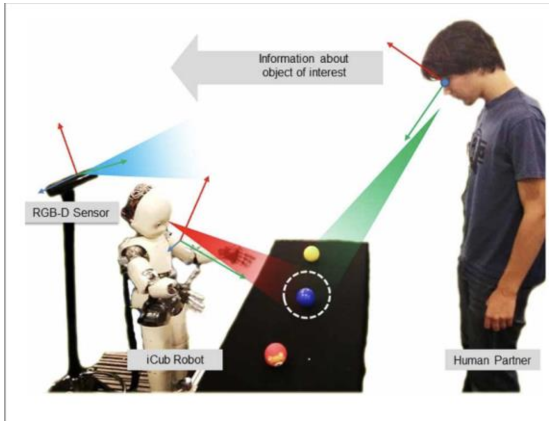
Social AI Group
Vrije Universiteit Amsterdam

Hardware



G. Perugia 2021

Head gaze



Serena Ivaldi 2014

model-based

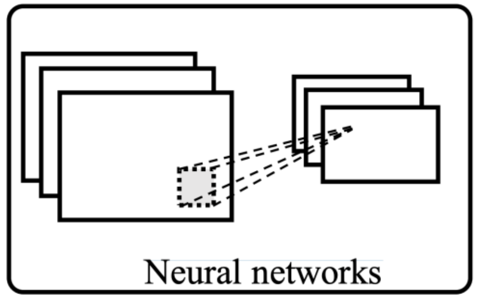
Oskar Palinko 2015

Background

Social AI Group
Vrije Universiteit Amsterdam

Appearance-based methods

pictures



gaze

Dataset

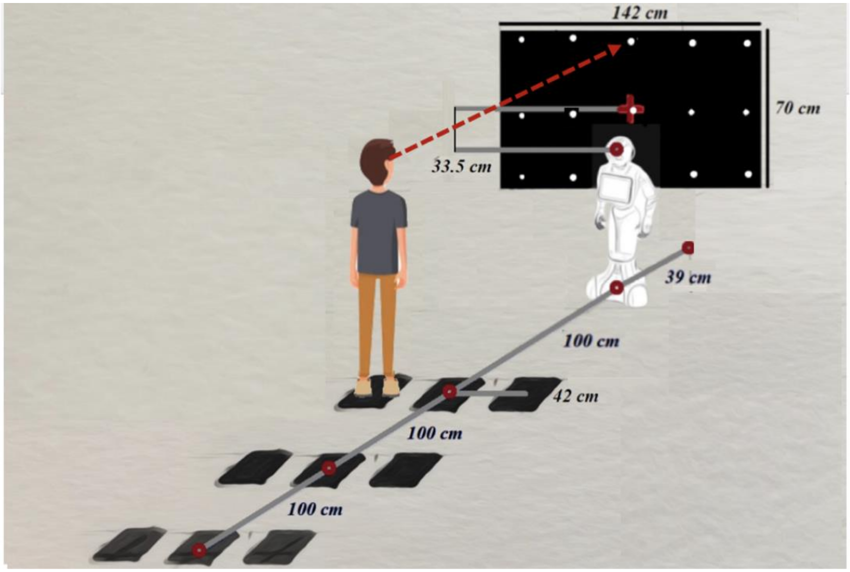
40°

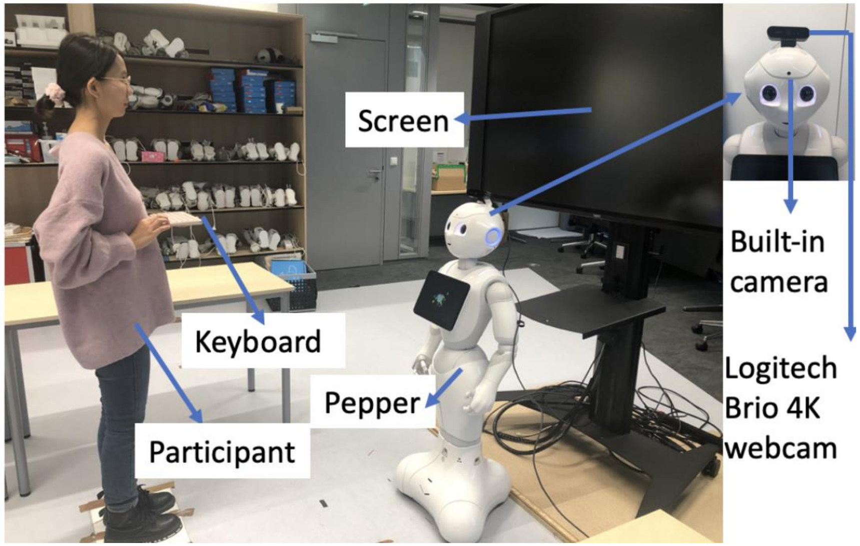
120°

Dataset	People	Head Pose(yaw and pitch)	Gaze(yaw and pitch)	Data	Resolution	Distance	Outdoor
Columbia	56	0°, ±30°	±15°,±10°	5880	5185*3856	200cm	No
UTMV	50	±36°, ±36°	±50°,±36°	64000	1280*1024	60cm	No
EYEDIAP	16	±15°, 30°	±25°,20°	237min	HD and VGA	80-120cm	No
MPIIGaze	15	±15°, 30°	±20°,±20°	213659	1280*720	40-60cm	No
RT-GENE	15	±40°,±40°	±40°,-40°	122531	1920*1080	80-280cm	No
Gaze360*	238	±90°,unknown	±180°,-50°	172000	4096*3382	100-300cm	Yes
ETH-XGaze*	110	±80°,±80°	±120°,±70°	1083492	6000*4000	100cm	No

Methodology

Social AI Group
Vrije Universiteit Amsterdam





Camera	Participants	Resolution	Images	Valid images
no4k	11	640*480	14850	13804
no4k	10	640*480	13500	12604
4k	10	3840 * 2160	13500	12706

What's our target

Social AI Group
Vrije Universiteit Amsterdam

In order to study **if the models based on GAZE 360 and ETH dataset can be used in HRI**. We design an experiment to

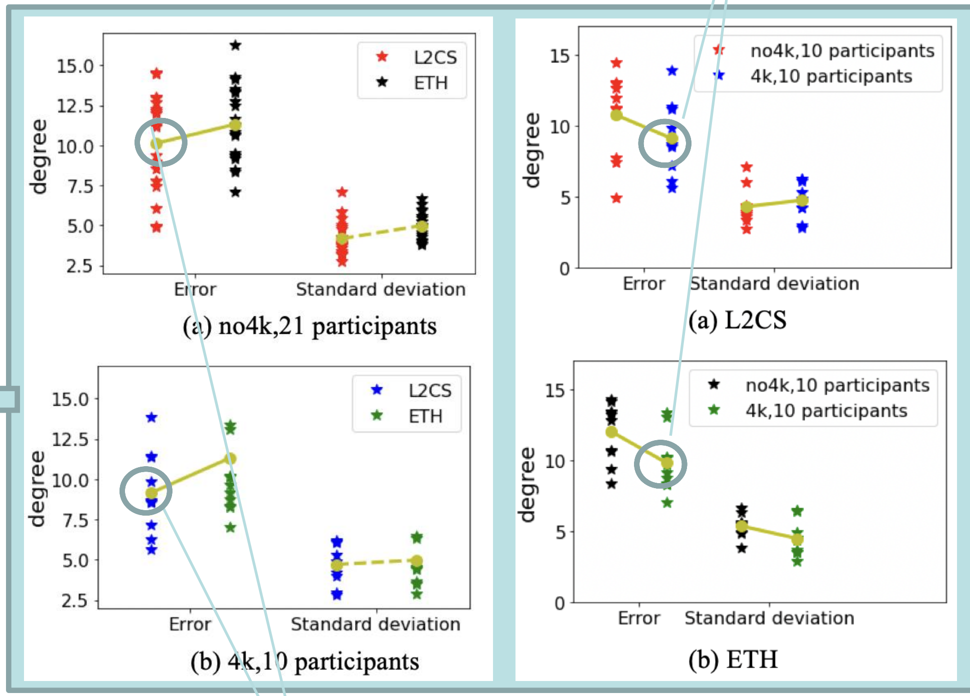
1. **systematically evaluate their quality**.
2. Explore **which factor influence the quality**(e.g. resolution, distance between human and robot,etc.)
3. Get to know **how these factor work**

Result

Social AI Group
Vrije Universiteit Amsterdam

A. Gaze estimation for different models and camera resolutions

The mean error of 4k is less than no4k for L2CS and ETH



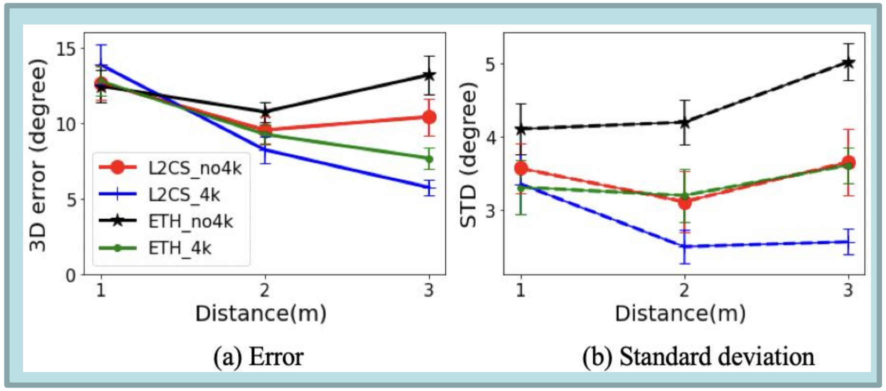
L2CS perform better than ETH in accuracy and precision under both resolutions(no 4k and 4k)

4k perform better than no4k in accuracy for both models

The mean error of L2CS is less than ETH under no4k and 4k

Result

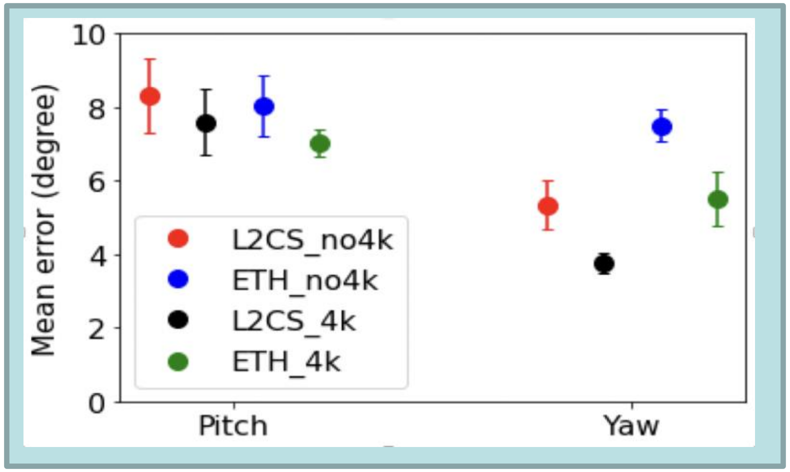
B. The role of distance from the camera



(a) Error (b) Standard deviation

L2CS model outperforms the ETH model in terms of accuracy and precision at 2 and 3 meters, regardless of resolution. L2CS combined with 4k is the best one

C. Offset Correction



offsets for both yaw and pitch, especially on pitch (the mean error reaches 7°). offset correction can significantly improve the performance on both accuracy and precision.

Application example

Social AI Group
Vrije Universiteit Amsterdam

1m(human&robot) ,experimenter view

Vocal cues

- **Prosody** (how something is said): pitch, tempo, and energy
- **Back-channeling** (express attention, agreement, wonder, etc.) and **disfluencies** (non-words, or fillers): ehm, ah-ah, uhm, etc.
- **Non-linguistic vocalizations**, e.g., coughing, laughing, sobbing, crying, whispering, groaning, etc.
- **Silences**: hesitation & psycholinguistic (difficulty), and interactive (convey messages about the interactions taking place)

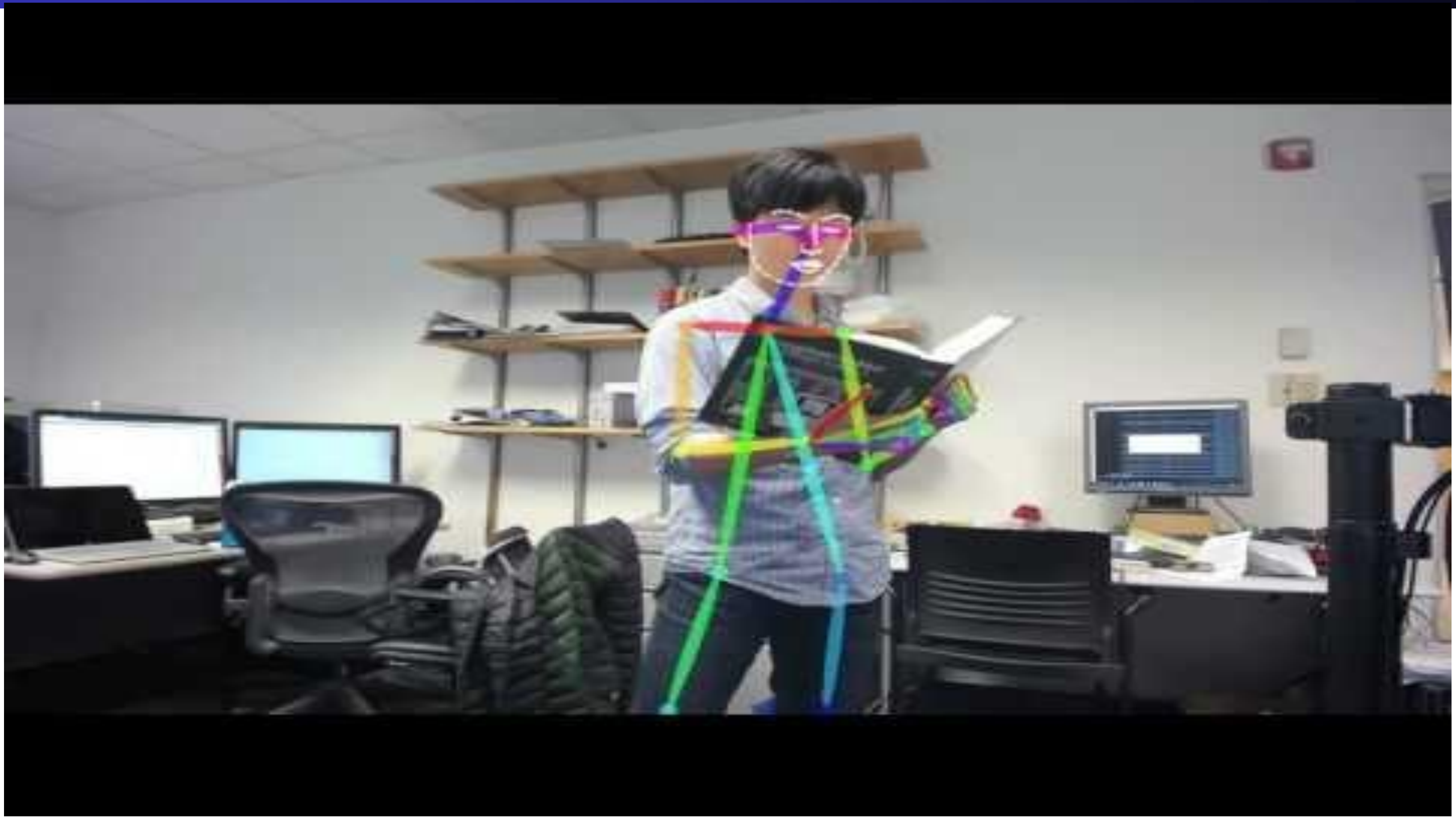
Postures and body movement

- Inclusive vs non-inclusive: looking at vs looking away

- f2f or parallel: more active (monitoring) vs less attentive

- Congruence vs incongruence: mirroring in interactive setting

Openpose & Gestures

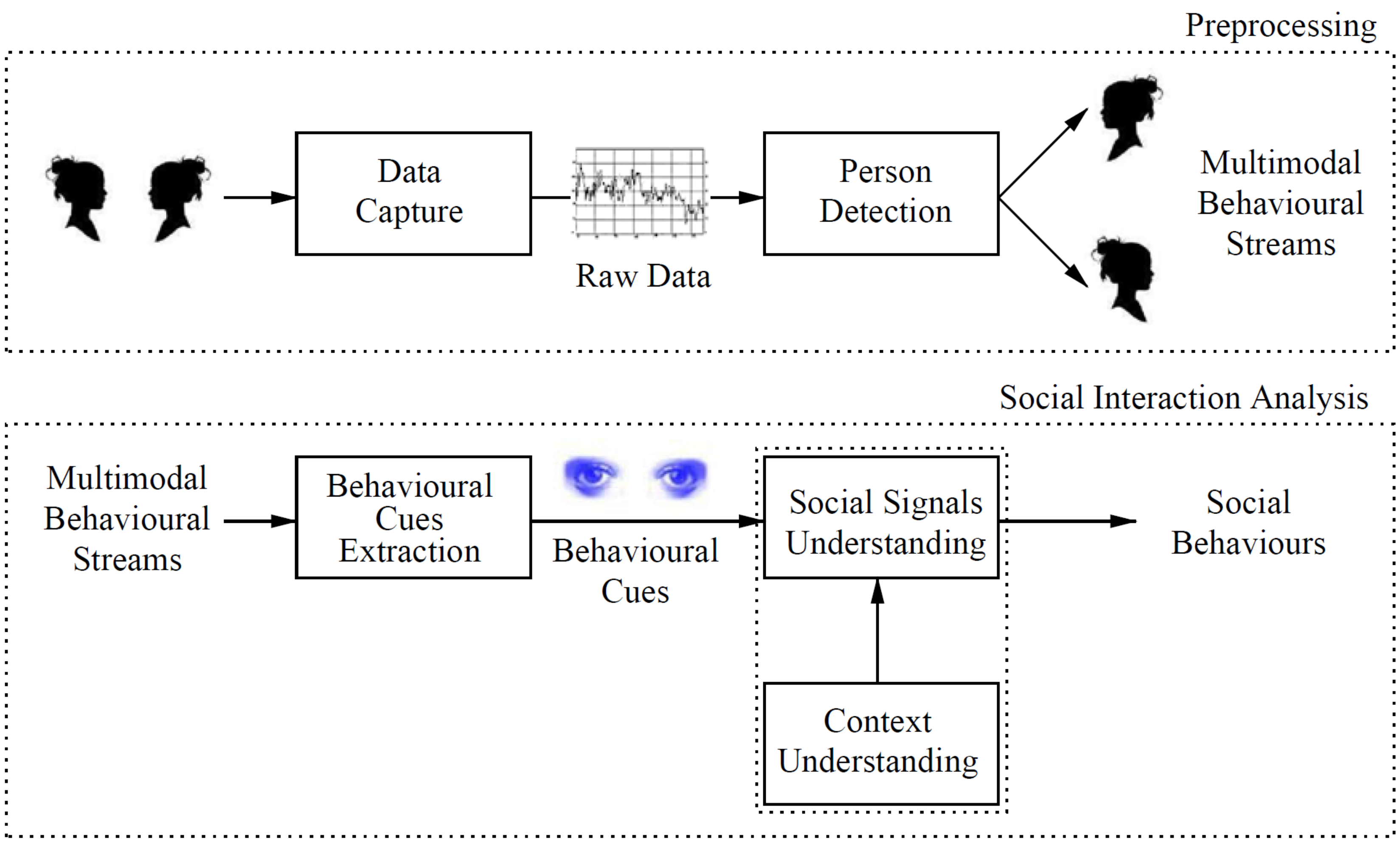


Two challenges:

- detecting the body parts in the gesture (e.g., hands)
- modeling the temporal dynamic of the gesture

CONTEXT

Is Social-Aware also Context-aware?



Source: Figure 6 in Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, *27*(12), 1743-1759.

How to interpret a smile?

A smile can be a display of:

- politeness,
- contentedness,
- joy,
- irony,
- empathy,
- greeting,
- …

How to interpret a smile?

To identify a smile **as a social signal** we need to know:

- *Where*: the location of the subject is (outside, at a reception, etc.),
- *What*: current task
- *When*: timing of the signal
- *Who*: the expresser is (identity, age, …)

This is the **W4 model** (where, what, when, who)

How to interpret a smile?

But comprehensive human behavior understanding requires the W5+ model (where, what, when, who, why, how):

- Why and how:
  - Identify the stimulus that caused the social signal (e.g., funny video)
  - Identify how the information is passed on (e.g., by means of facial expression intensity).

Addressing W5+ is key challenge of data-driven SSP.

Future work

Important but not discussed today:

- context-dependent multimodal fusion
- multimodal temporal fusion
- multiparty
- are social signals natural or cultural?