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 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**

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

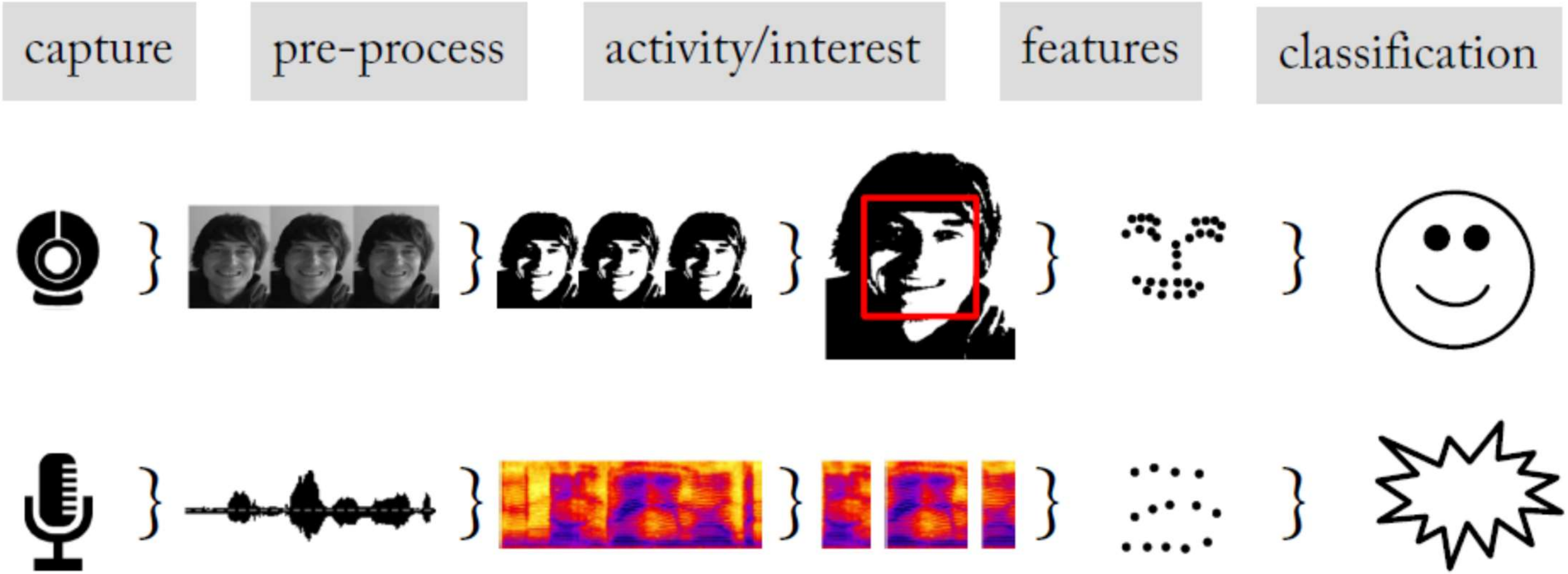
- Space and environment (proxemics)
- Physical appearance – height, body shape, skin and hair color, dress
- Facial expressions
- Gaze & head pose
- Postures and body movement
- Gestures (hand and arm)
- Vocal cues
- …

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?

Assumption:

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

Proxemics & People Detection

Thomas van Orden
joint work with the robot programming team in the Social AI lab.

Should I say hello?

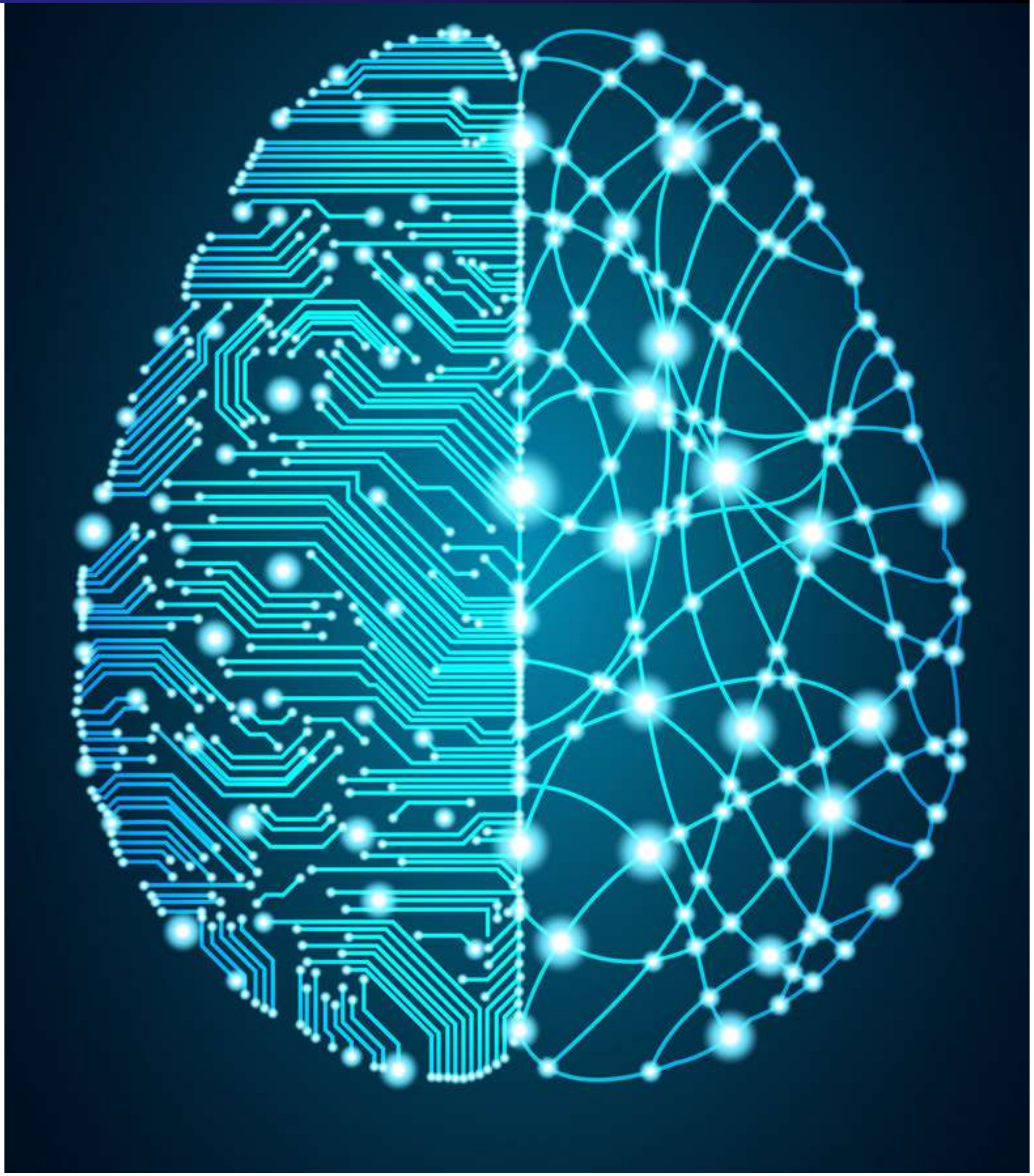
The very basics of human-robot interaction

- Who should I greet? Not all persons might be interested in a conversation.
- As humans we do not have to think about this.
- <u>The robot</u> should be able to start the interaction.
- It needs intelligence to do so!
- Social AI comes into play.

How smart am I?

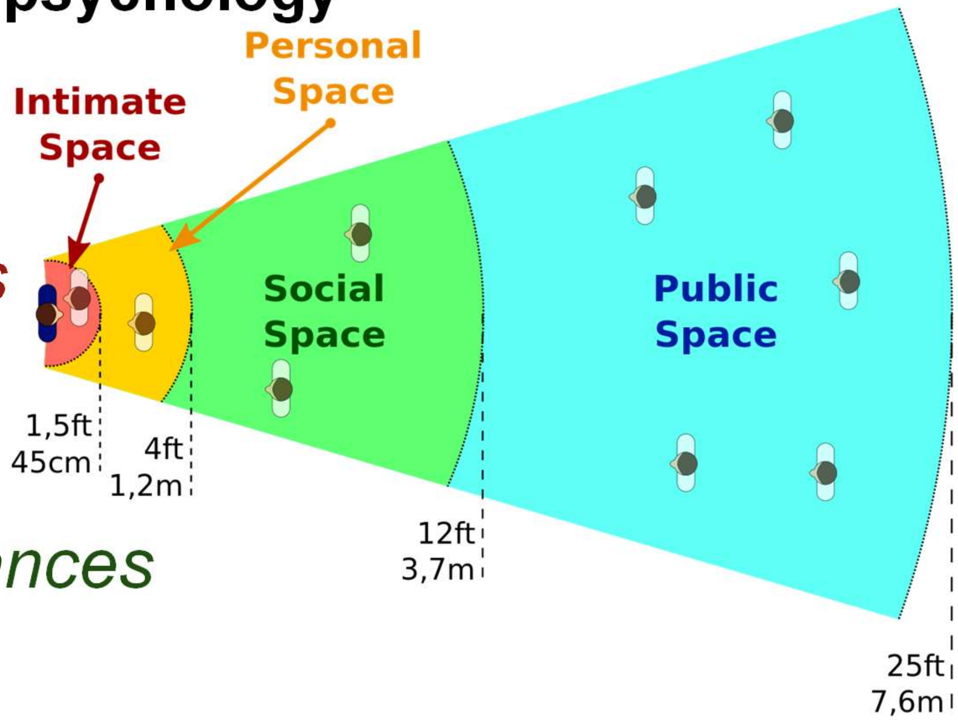
Requirements for greeting

- Detect the presence of a person.
  - How many persons are there?
- Locate <u>and</u> identify the person(s).
- Decide to greet a person or not
  - Which person do I greet?
  - When?
- Do the actual greeting action within an <u>appropriate timespan</u>.



Shall I greet this person?

Theoretical foundation from social psychology



- Hall's Proxemic Theory[1]
- *Intimate relationship; e.g. lovers*
- *Good friends, family*
- *Impersonal business; acquaintances*
- *Public speaking*
- Social space / personal space could fit our scenario.
- Additional requirement: distance to a person.


Person 1
at 2.12m
Person
at 2.29

How to perceive depth from 2D images?

A human-inspired approach

- We as humans need the signal from both eyes to perceive depth —> use stereo images instead of mono images with robots!

- Objects closer to camera ‘move’ more between two images.
- Basic technique, cost-efficient and fast.

Pipeline for stereo depth

The devil is in the details

- Calibrate the camera (checkerboard)
- Take stereo image
- Split image into two separate images
- Align both images with calibration matrices

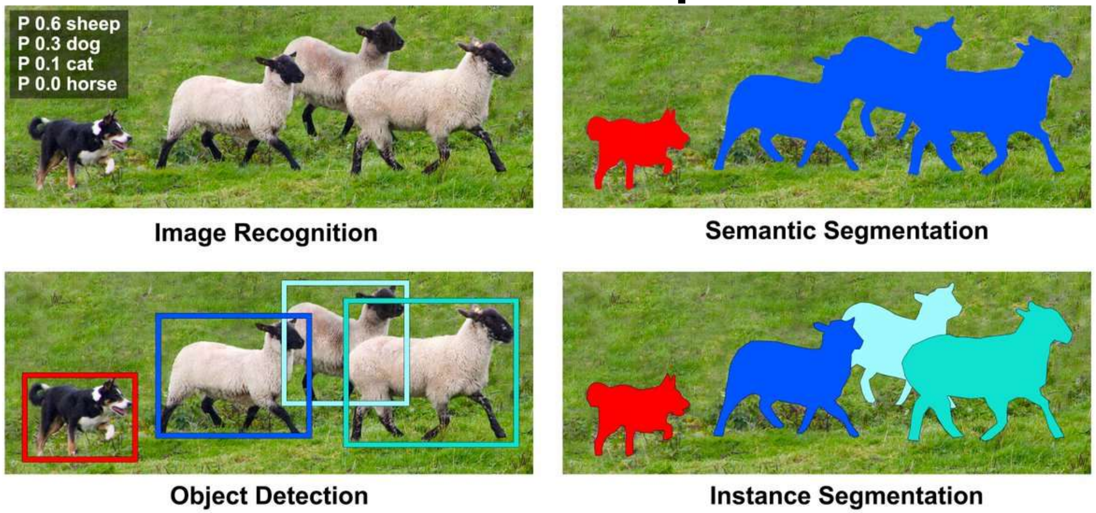
$$\boxed{z = \frac{f \times b}{x_l - x_r} = \frac{f \times b}{d}}$$

- Match every pixel in left image with a pixel in right image (occlusion is a problem)
- Create disparity map to obtain per-pixel depth information:
- Calibration and alignment are great challenges. Use more points for calibration!

Depth is cool, but what is a person?

Instance segmentation in a nutshell

- From image recognition to instance segmentation.
- We need to be able to distinguish between people.
- Object instance detection not sufficient; why not?
- Use the instance mask for depth.
- But, what if the mask is not perfect?



What is the metric used?

A: Mean
B: Median
C: Maximum

What is the metric used?

A: Mean

B: Median

C: Maximum

How to do instance segmentation?

Lots of research done in CV field

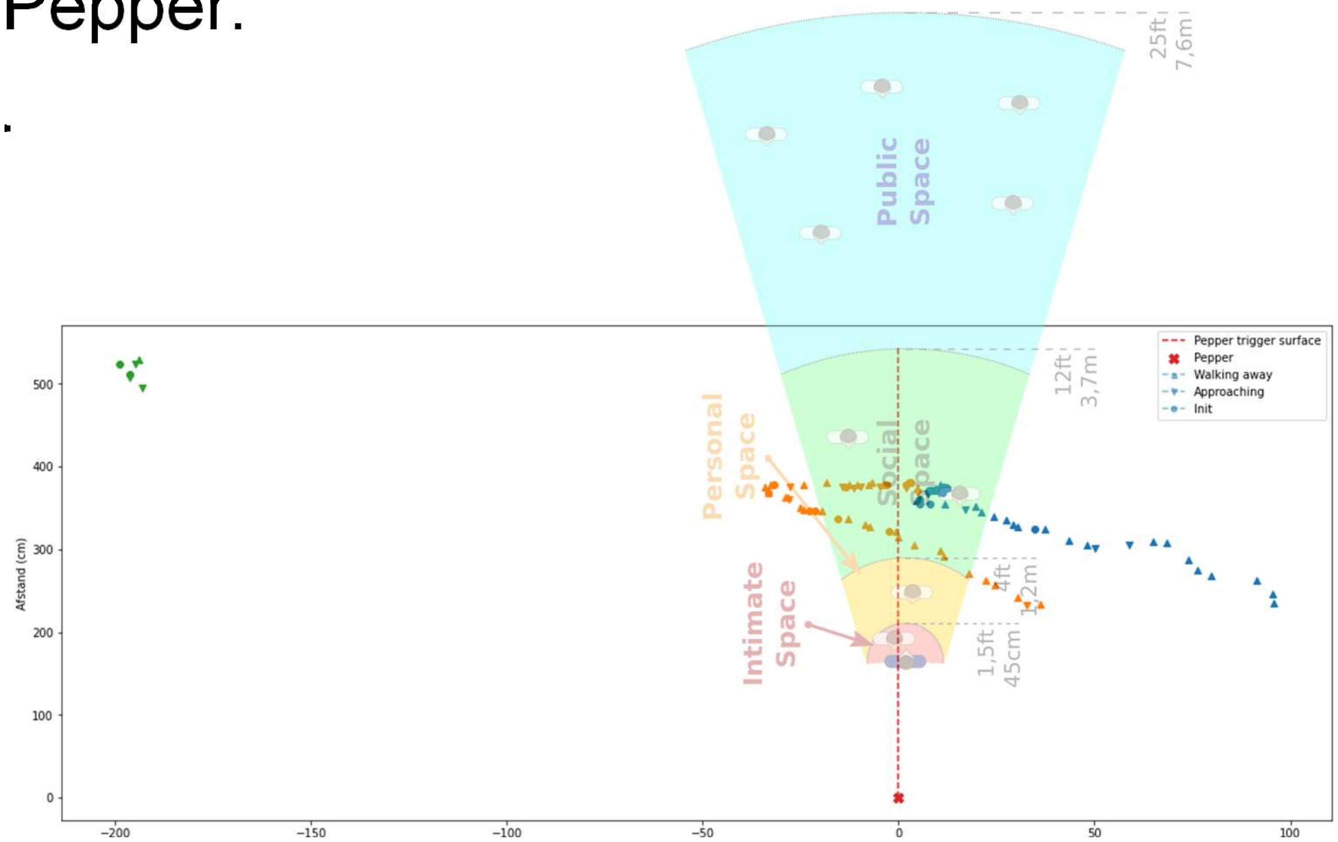
- Detecron2[3] library from Facebook Research.
- Mask-RCNN[4] structure with ResNet50[5] backbone.
- Many pertained models are available.

Detectron2

Combining it all

Depth, instance segmentation, Hall's Proxemic…

- We can now generate a top-down view of a situation.
- Distance relative to Pepper.
- Distinguish persons.
- Approaching or not.
- Hall's spaces.
- Path prediction.
- <u>Shall I say hello?</u>



Remaining challenges

We have still some work to do

- What happens when more people (e.g. 5+) enter the frame?
- How reliable is the pipeline?
- How can we <u>track</u> people better?
- How can we make the whole pipeline fast enough to run in real-time?

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

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

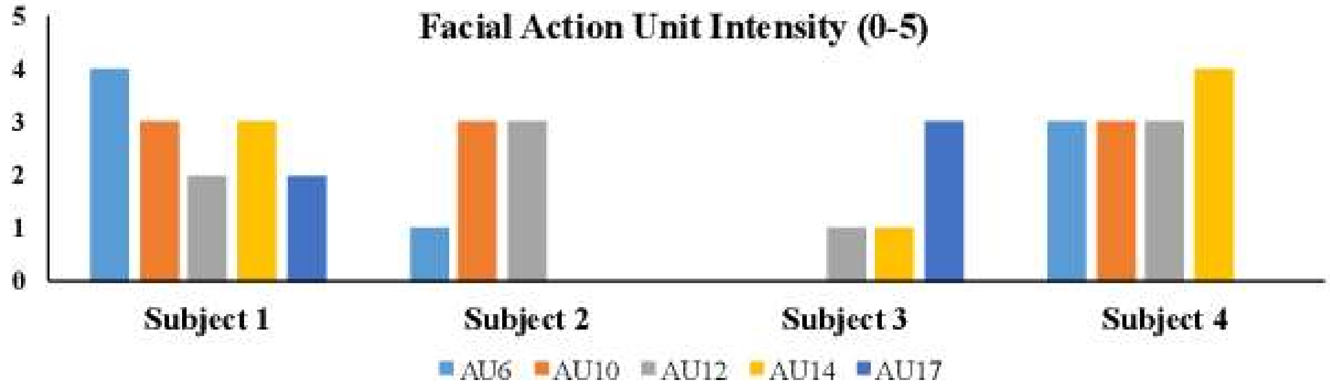
Displayed **Happiness = AU06 + AU12**

AU06 – Cheek Raiser

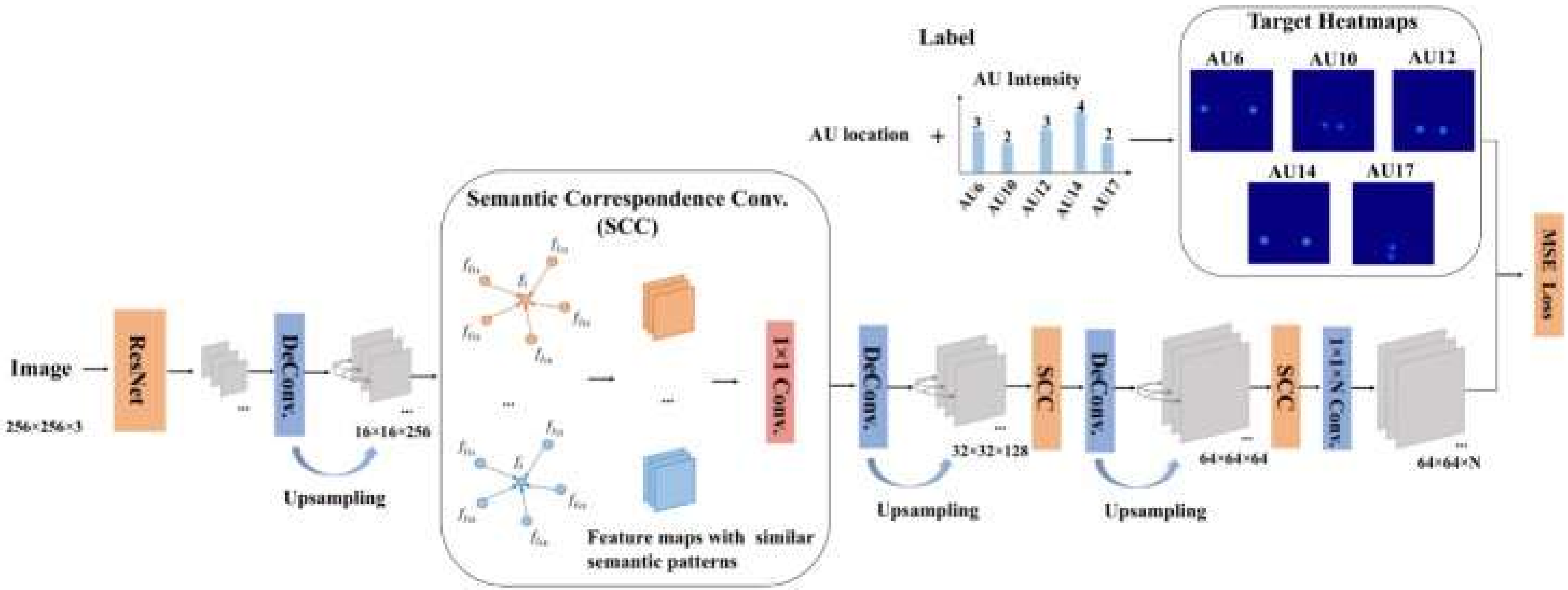
AU12 – Lip Corner Puller

AUs have intensity → can be used to derive emotion intensity

Facial Action Unit Intensity Estimation



Facial Action Unit Intensity Estimation via Semantic Correspondence Learning with Dynamic Graph Convolution. Yingruo Fan, Jacqueline C.K. Lam and Victor O.K. Li. *AAAI 2020*



Exercise Feedback for People with Facial Paralysis

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

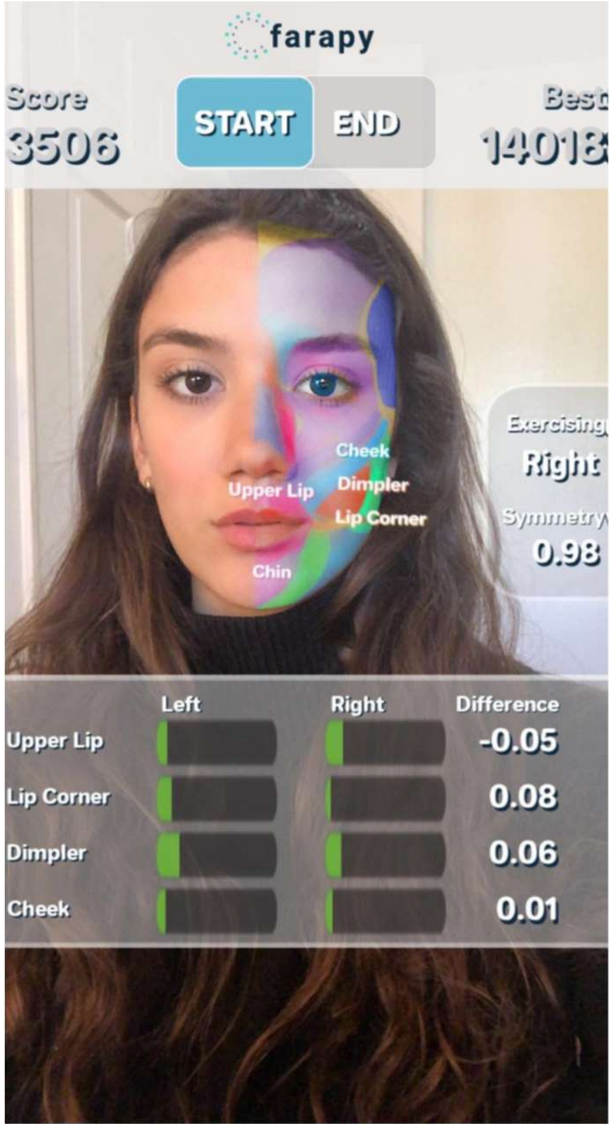
2. Wrinkle nose, holding for 5 seconds, repeating 10x.

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.

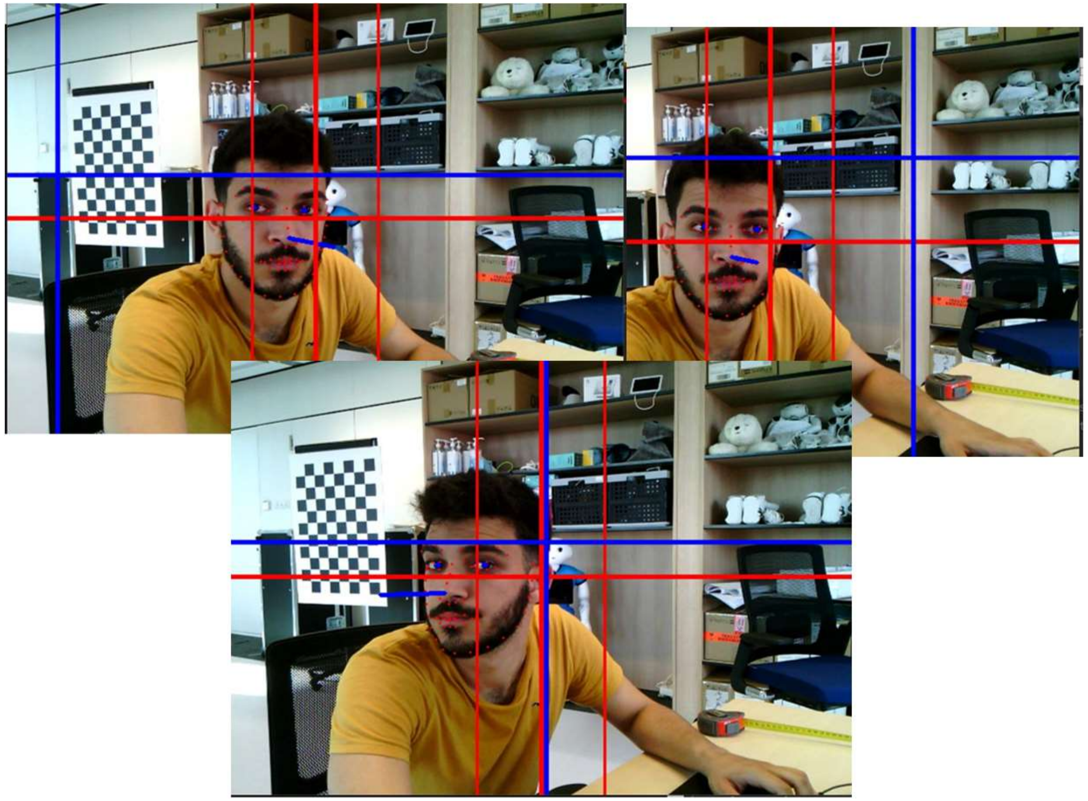


Gaze & Head Pose

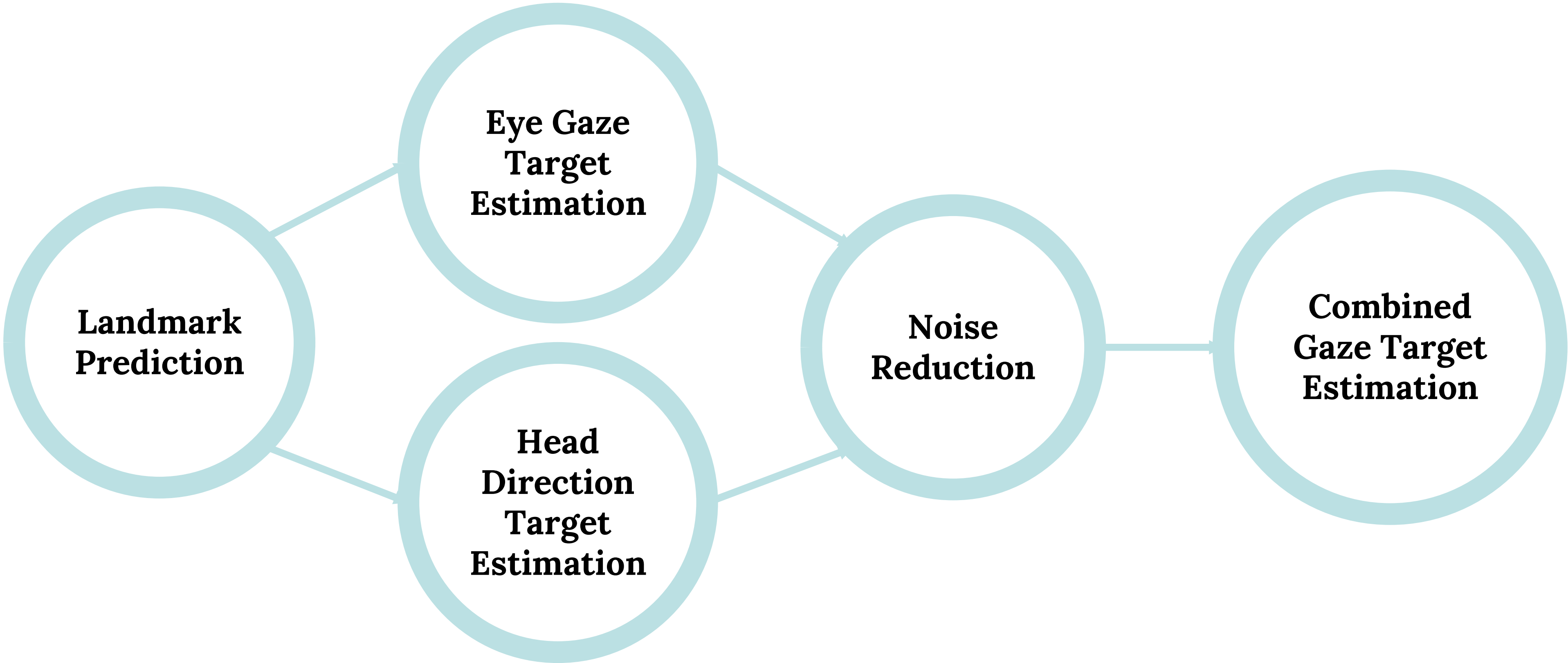
Gaze Direction Estimation under Varying Head Positions using a Pepper Robot's In-Built Camera

Thesis of Marinos Savva

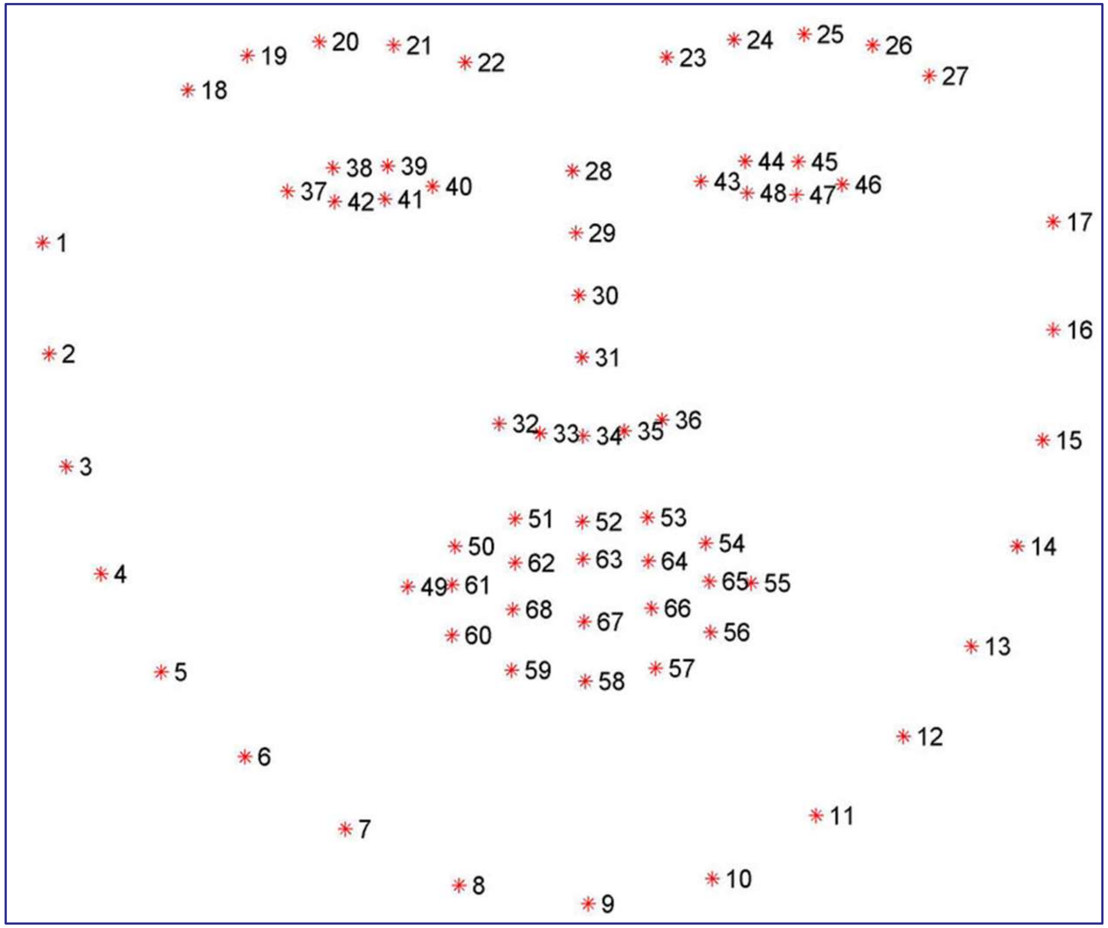
How can we estimate human eye gaze direction from camera input?

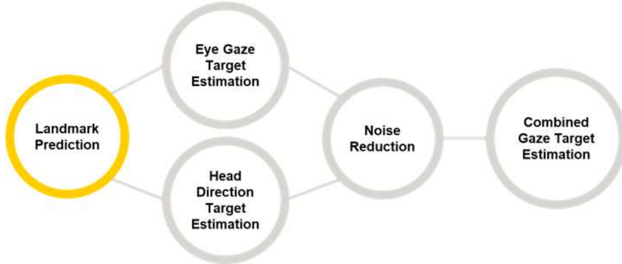


A First Principles Approach



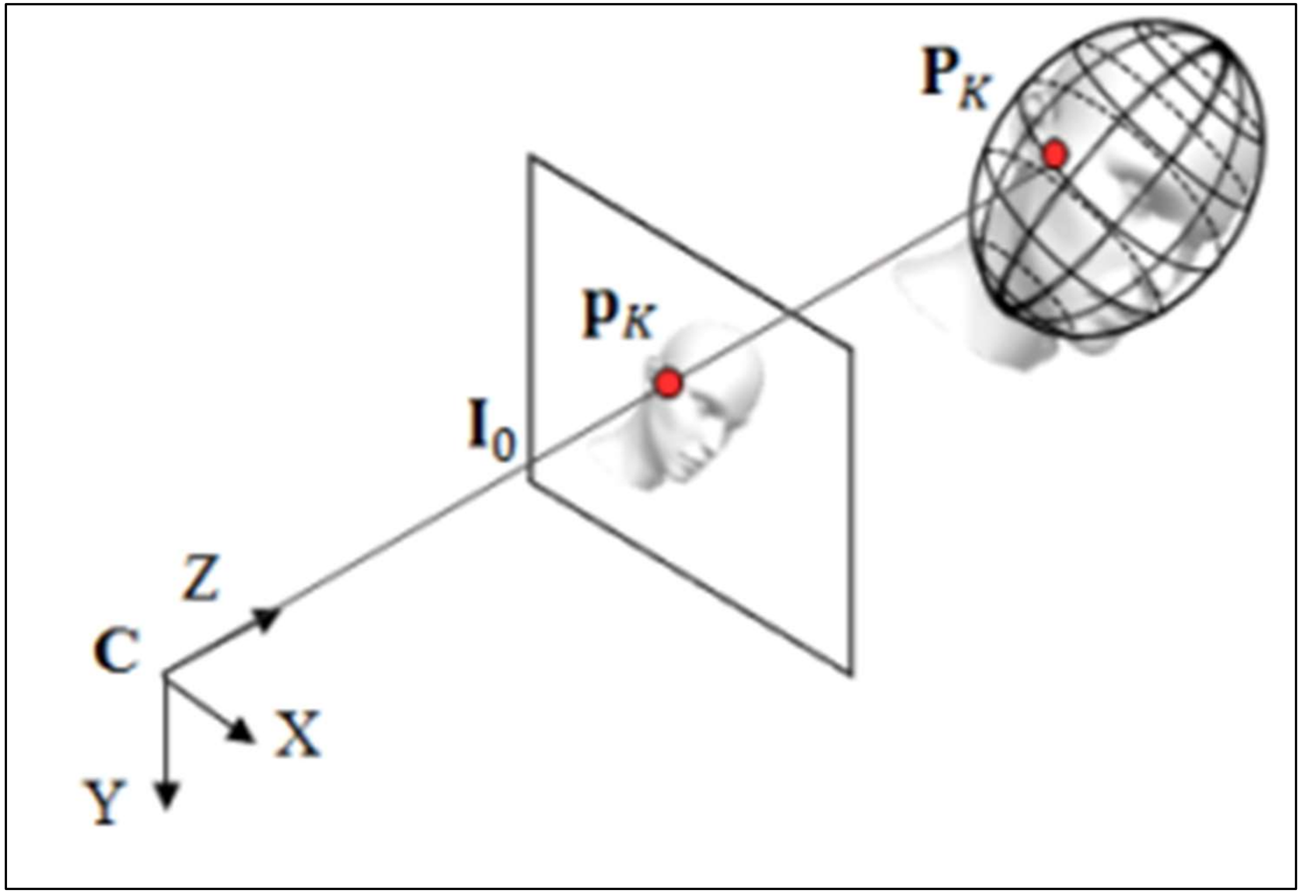
Step 1: Landmark Prediction

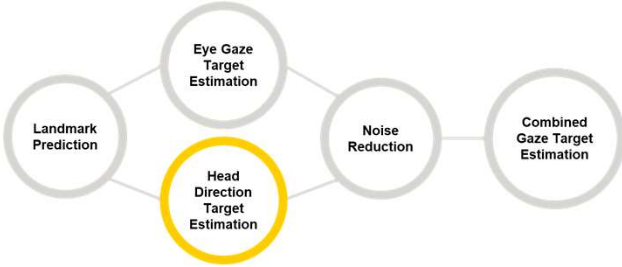




Source: V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees," in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867-1874

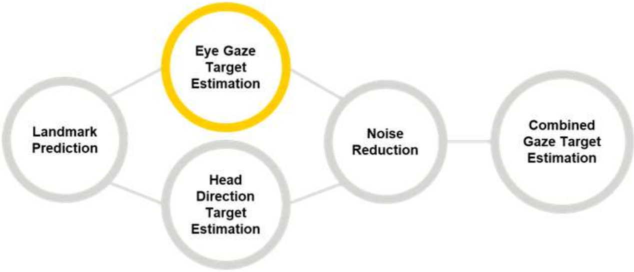
Head Pose Estimation





Source: J. M. D. Barros, B. Mirbach, F. Garcia, K. Varanasi, and D. Stricker, "Real-time head pose estimation by tracking and detection of keypoints and facial landmarks," in VISIGRAPP, 2018.

Eye Gaze Target Estimation



Region of Interest Extraction

The landmarks acquired are utilized to create a cut-out of the eye while omitting the area around the eyeball

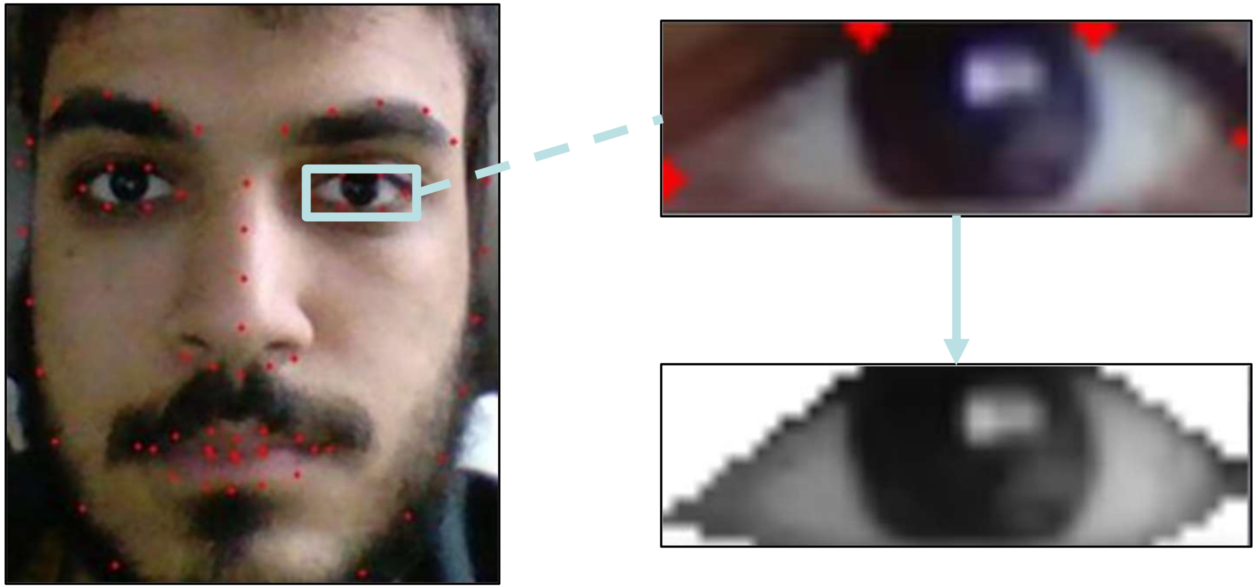
Image Processing

The extracted image is processed to remove the effects of surface light reflections and to localize the pupil

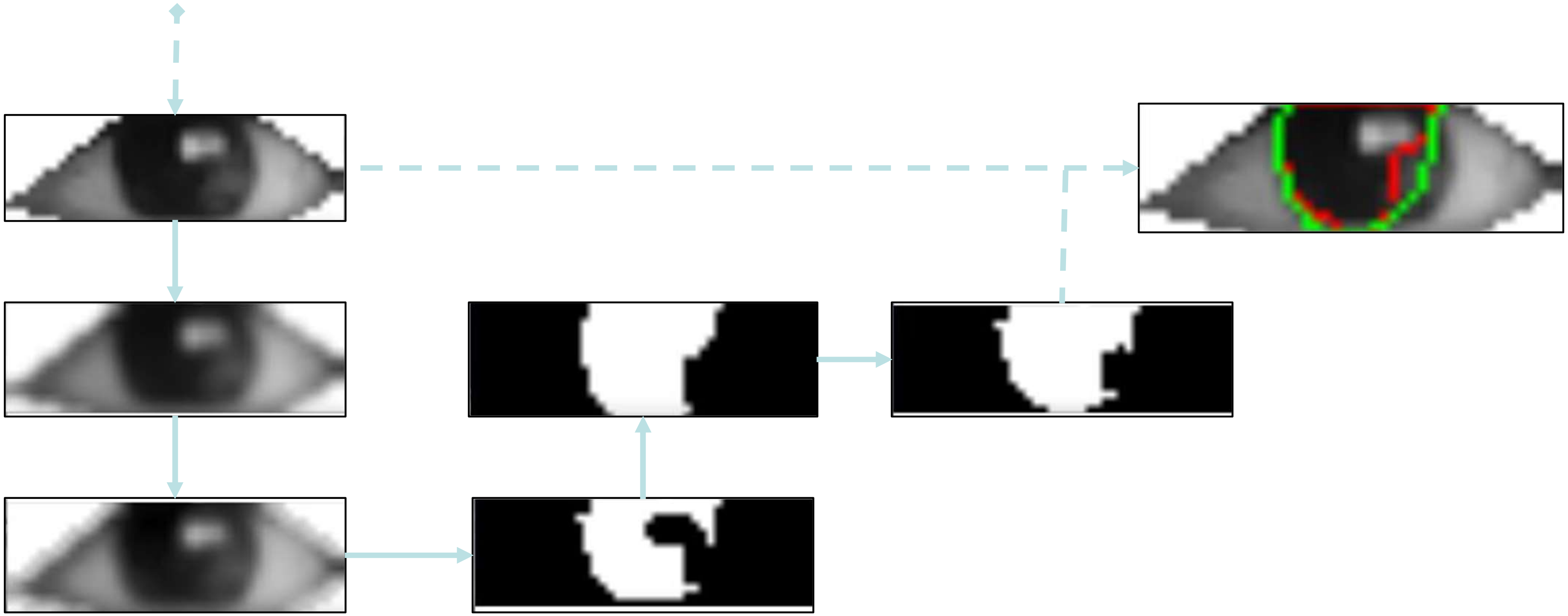
Gaze Angle Calculation

The angle of gaze is calculated using a simplified eyeball model

Region of Interest Extraction



Localizing the Pupil

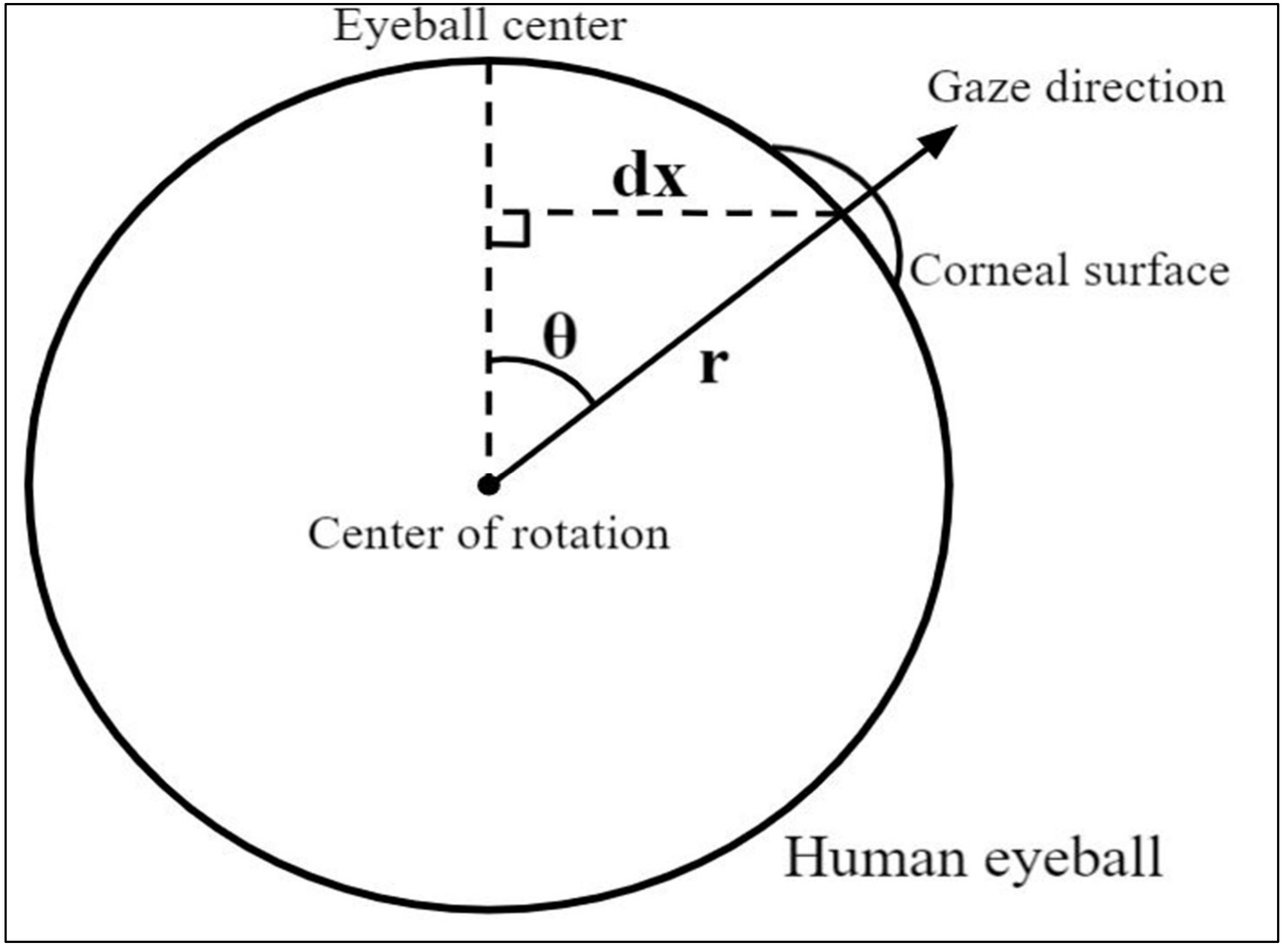


Gaze Angle Calculation

A simplified model of the eyeball is used, where some assumptions are made

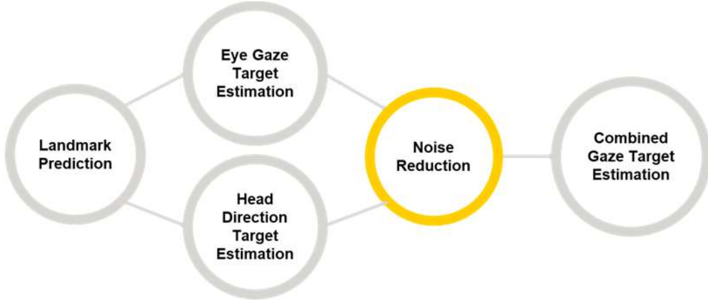
- Eyeball shape approximated to that of a sphere of constant radius
- Eyeball radius approximated to human average of 10.94mm
- Assumed no difference between pupillary and visual axes

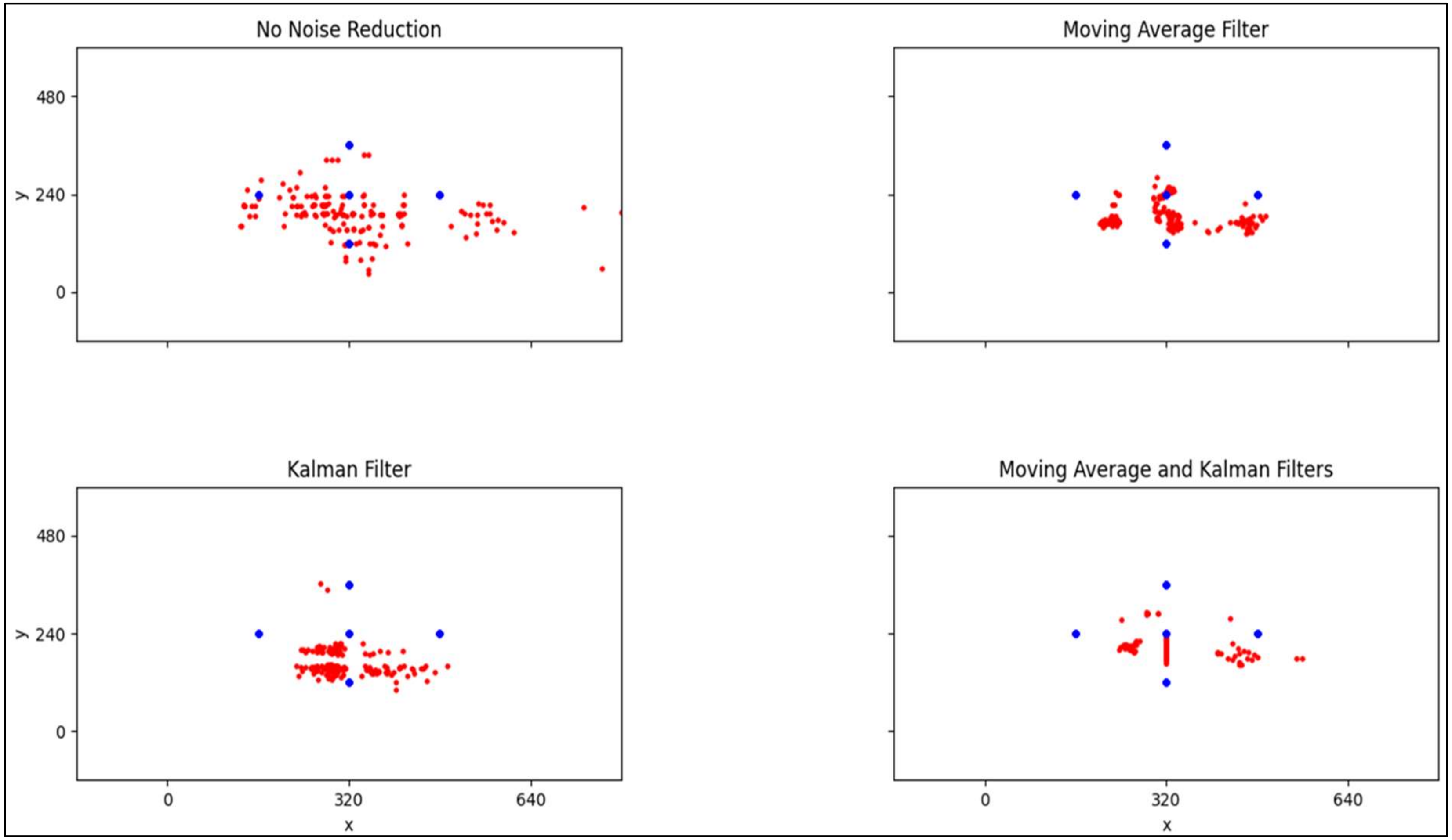
Gaze Angle Calculation



Source: N. M. Scoville, R. Y. Lu, and H. Jung, "Optical axes and angle kappa," url: https://eyewiki .aao.org/Optical_Axes_and_Angle_Kappa.

Noise reduction





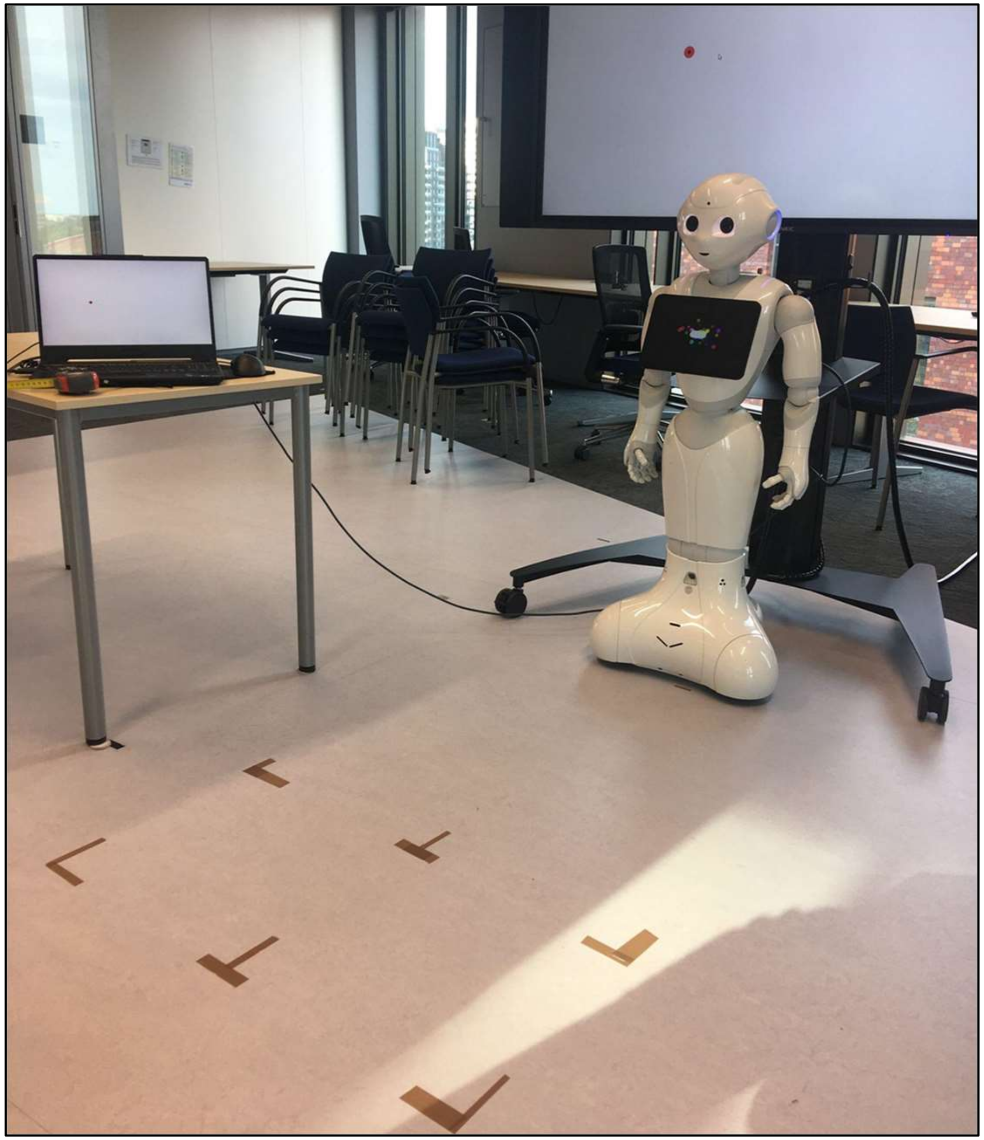
Evaluation: Setup

Small experiment:

- 12 participants

- 6 standing positions
  - 2 distances
  - 3 offset positions

- 5 projected points per position



Evaluation: Results

Participant	80cm		120cm	
	Error(px)	Accuracy	Error(px)	Accuracy
1	122.41	61%	159.95	55%
2	151.41	54%	225.54	43%
3	159.29	53%	266.36	33%
4	167.66	50%	298.63	23%
5	165.25	50%	268.66	30%
6	170.58	49%	295.40	20%
7	164.10	51%	288.00	23%
8	162.35	51%	270.35	28%
9	162.59	51%	258.83	30%
10	159.3	52%	247.76	33%
Mean	158.49	52%	257.65	32%

Participant	80cm		120cm	
	Accuracy X-Coordinate	Accuracy Y-Coordinate	Accuracy X-Coordinate	Accuracy Y-Coordinate
1	78%	44%	74%	37%
2	69%	40%	63%	24%
3	66%	41%	65%	1%
4	65%	34%	47%	0%
5	65%	36%	51%	9%
6	62%	38%	38%	3%
7	65%	39%	38%	8%
8	67%	36%	42%	14%
9	68%	35%	45%	16%
10	69%	36%	48%	19%
Mean	67%	38%	51%	12%

MIT's Gaze 360

wearable eye tracker glasses

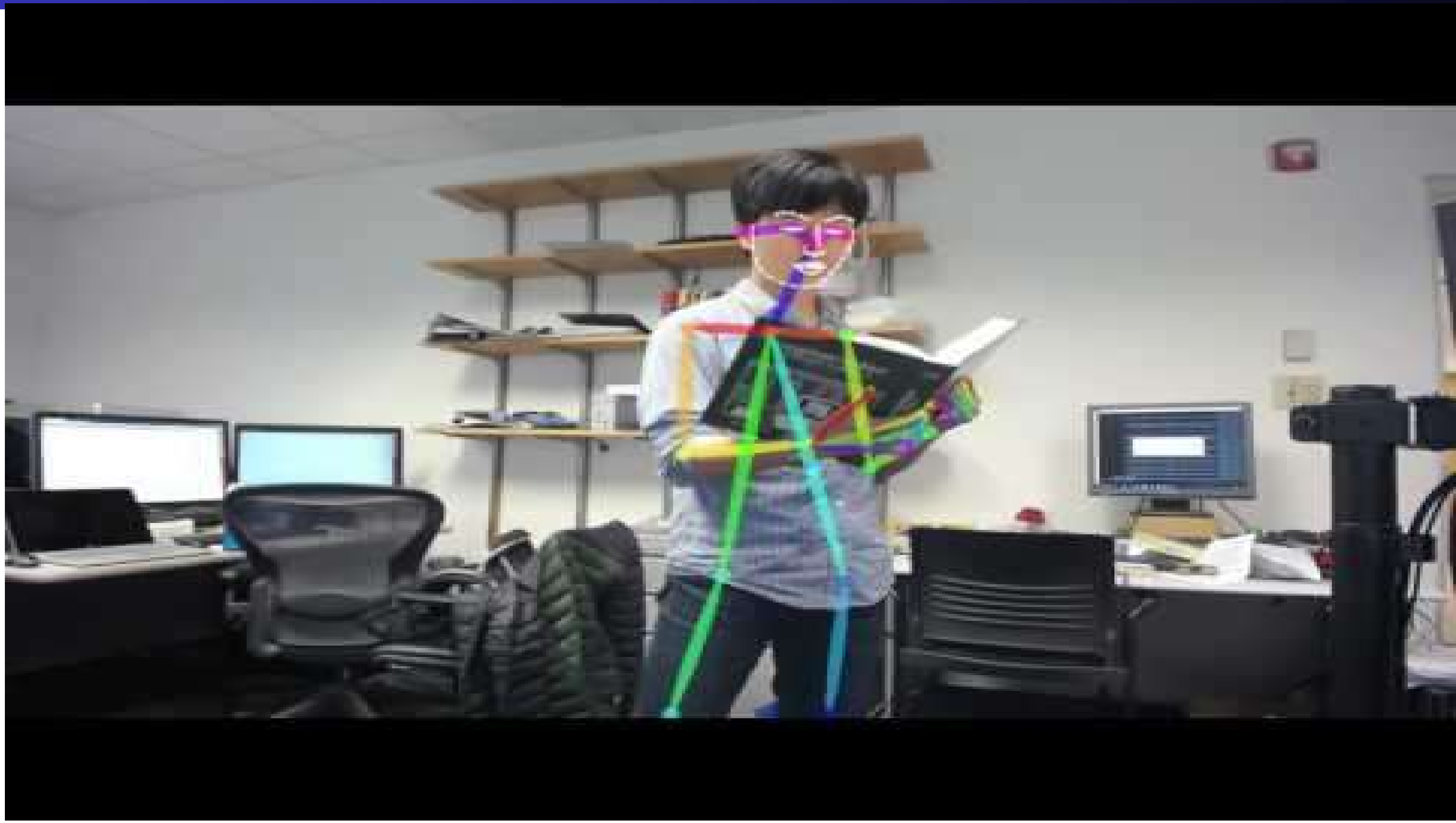
Vocal cues

- **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

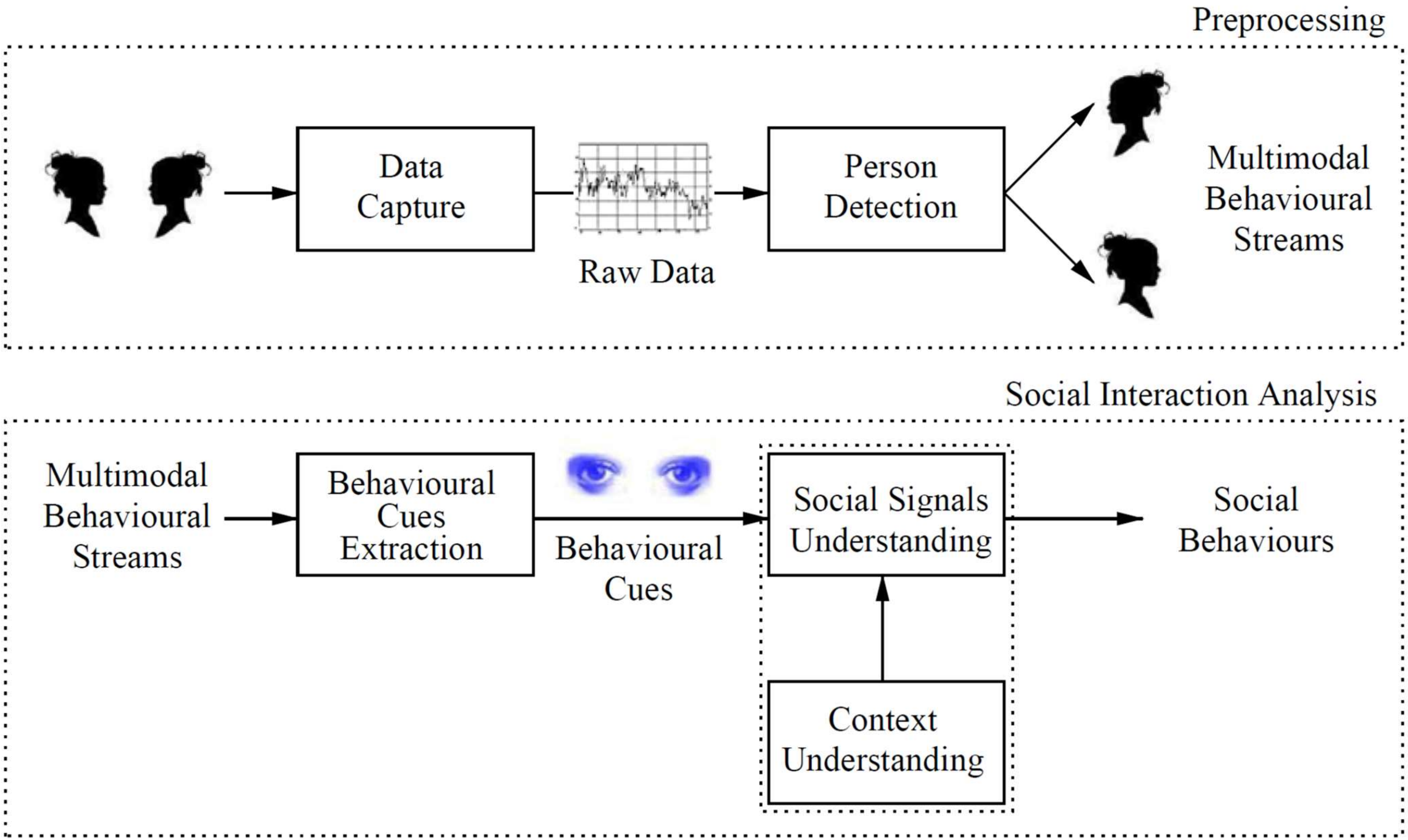
Openpose & Gestures



Two challenges:

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

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)

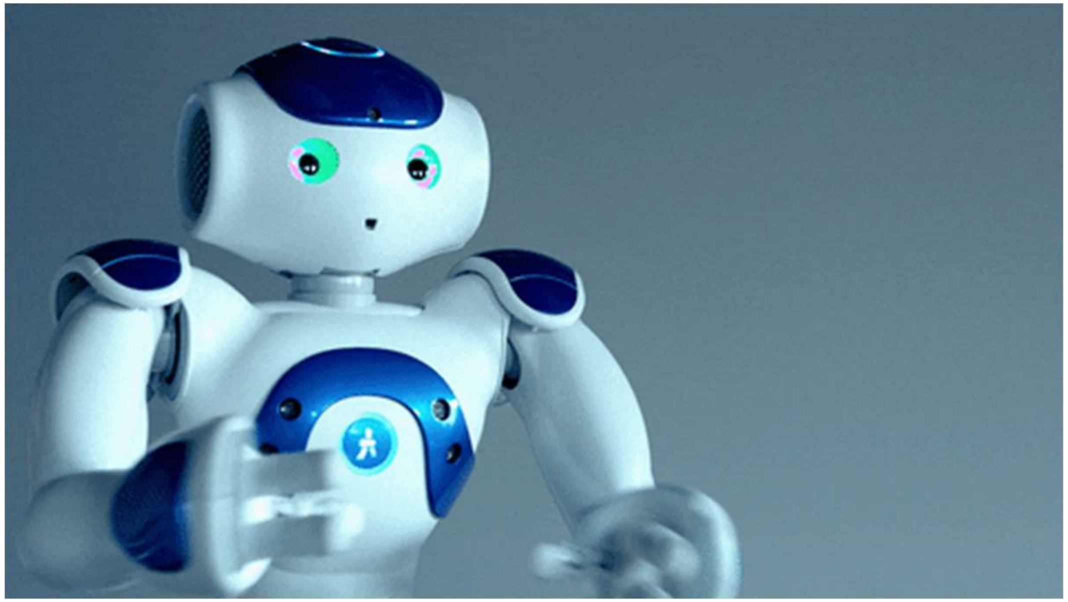
How to interpret a smile?

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?