PHUSE LAB
UNIVERSITY OF MILAN

AUDIO-VISUAL ATTENTION MODELING VIA REINFORCEMENT LEARNING

Supervisor:
Prof. Boccignone Giuseppe

Author:
Bocchino Daniele

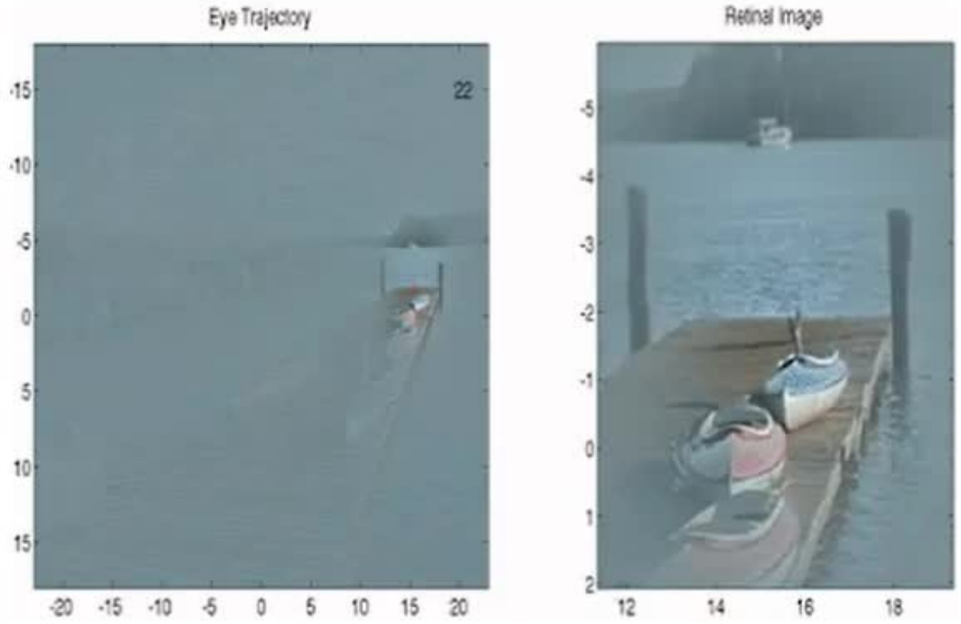
Co-Supervisor:
Dott. D'Amelio Alessandro

PHUSE Lab
Perceptual computing and HUman SEnsing

VISUAL ATTENTION

Visual attention is the ability to focus on one element and ignore irrelevant information.It consists of a continuous alternation between fixations and saccades.

Visual attention is employed in robotics to replicate human behaviors in robotic platforms such as Icub.

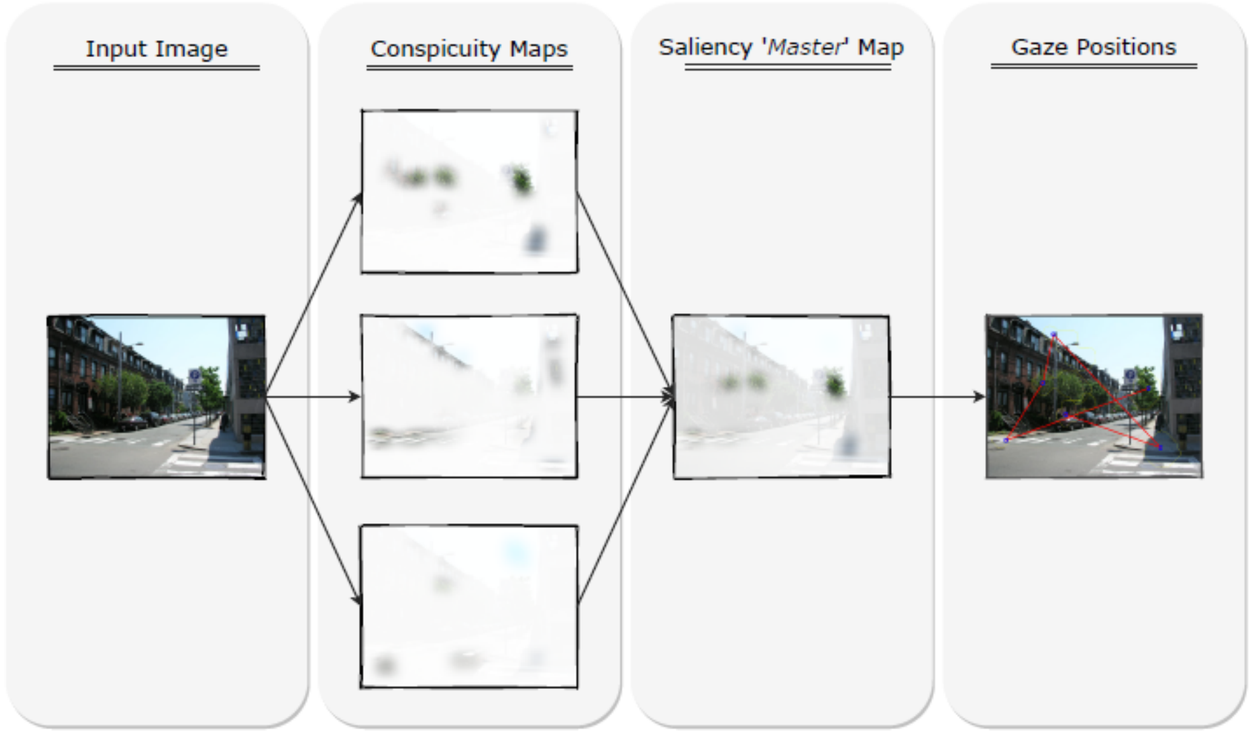


Bocchino Daniele

13 luglio **2023**

PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing

COMPUTATIONAL MODELS OF VISUAL ATTENTION

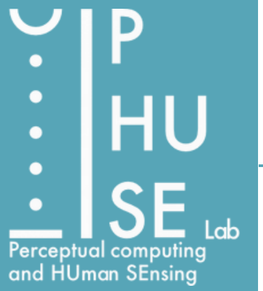


Two Main Stages:

- Perceptual Representation
  - Saliency Model
- Gaze Shift Model
  - Mechanics of Oculomotion
  - Time to spend in a location
  - Choose where to look next

PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing

PHUSE LAB
UNIVERSITY OF MILAN



PERCEPTUAL REPRESENTATION
SALIENCY MODEL

Perceptual representation constructs an image of what the observer perceives, highlighting objects of interest such as speakers, non-speakers, text, and salient aspects like color and contrast.
Perceptual representation is essential in computer vision tasks such as object detection and face detection

MECHANICS OF OCULOMOTION
ORNSTEIN UHLENBECK PROCESS

The Ornstein-Uhlenbeck process is a stochastic process that describes the dynamics of variables returning to a mean value over time.

Attective Position

Current Observer Position

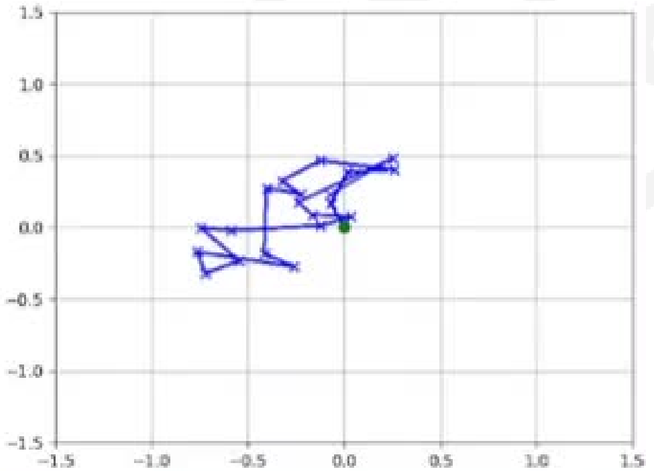
Random Walks

$$d\mathbf{r}_F(t) = \mathbf{B}_{p^*}^{(z_t)}\left[\boldsymbol{\mu}_{p^*}^{(z_t)} - \mathbf{r}_F(t)\right]dt + \mathbf{D}_{p^*}^{(z_t)}\left(\mathbf{r}_F(t)\right)d\mathbf{W}^{(z_t)}(t)$$

Fixation/ Saccades

The matrix B regulates the attraction strength towards the center.

The matrix D represents the diffusion of the random walks

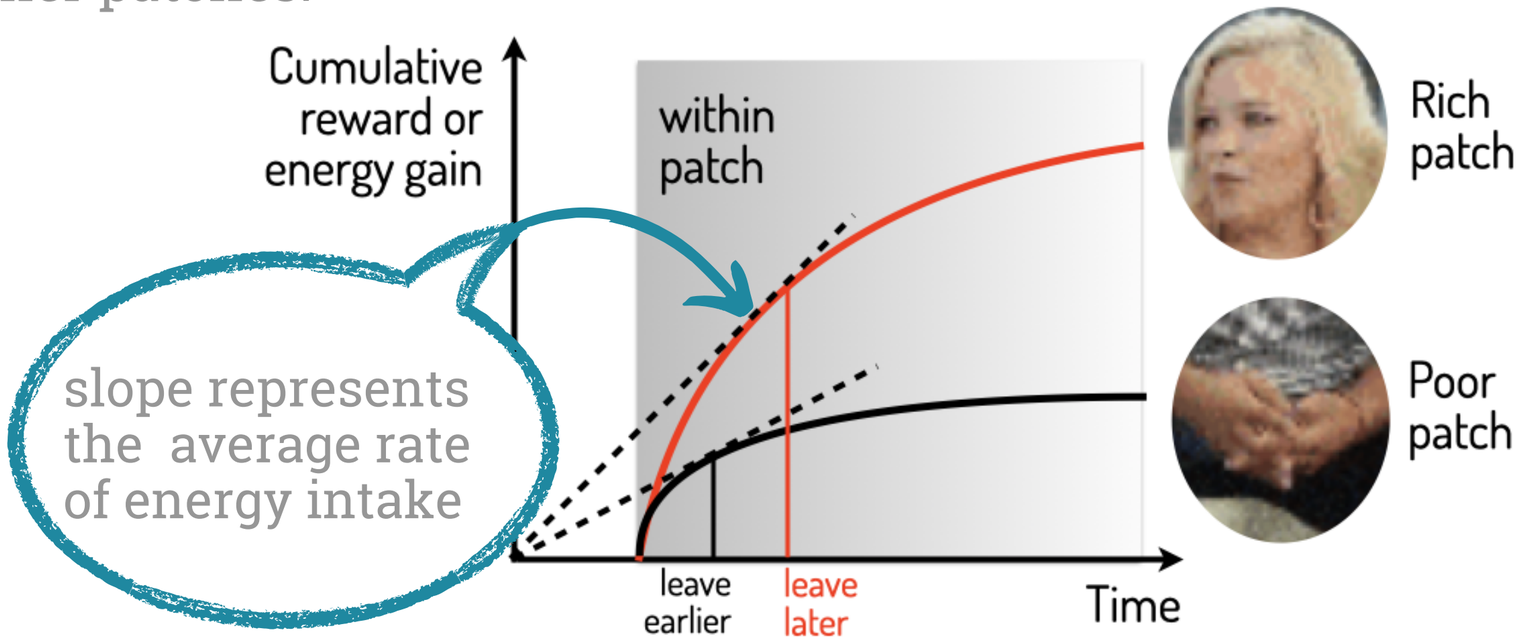


PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing
Bocchino Daniele

13 luglio **2023**

TIME SPENT IN A LOCATION
MARGINAL VALUE THEOREM

The observer should leave the current patch when the reward from that patch is lower than the average reward rate from the other patches.



The observer should leave the current patch when the reward from that patch is lower than the average reward rate from the other patches.

Bocchino Daniele

13 luglio **2023**

PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing

WHERE TO LOOK NEXT?
CONTEXTUAL MULTI ARMED BANDITS

- **Context:** At each time step t, the agent receives a vector denoted as x_t.
- **Action:** Consists of K arms or actions, where K is the total number of arms. At each time step the agent selects an arm k_t based on the observed context x_t.
- **Reward:** The reward r_t obtained by pulling arm k_t at time step t is a random variable. The reward depends on the context x_t and the chosen arm k_t.

Vector Context

Transforms linear reward into a probability.

$$\theta_k(x) = \frac{1}{1+exp^{-f(x)}}$$

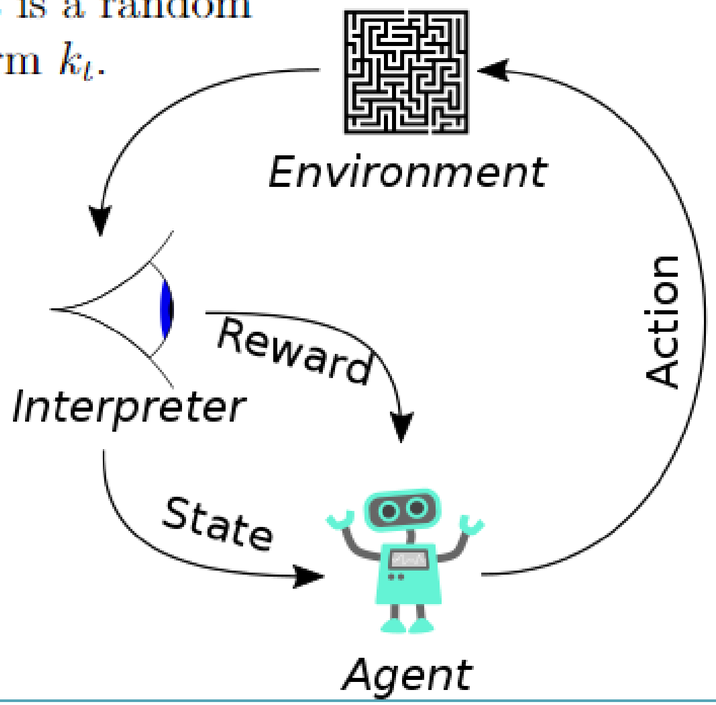
Expected Value

$$f(x) = \beta_0 + \beta_1 \cdot x + \epsilon$$

Linear Function Params.

$$\epsilon \sim N(0, \sigma^2).$$

Randomness



PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing
Bocchino Daniele

13 luglio **2023**

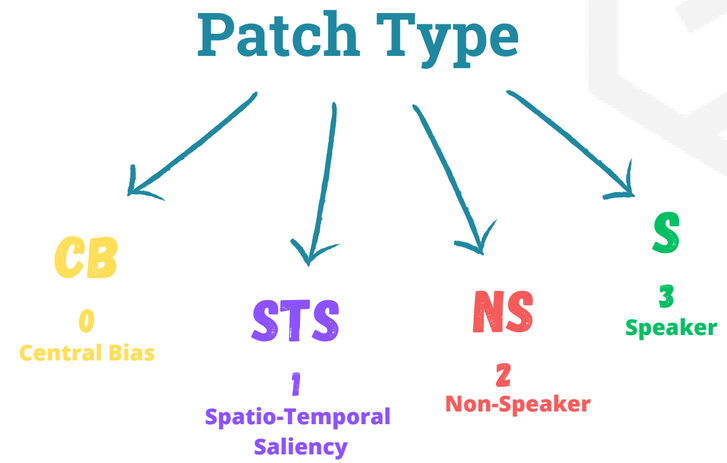
HOW TO LOOK NEXT ?

Each patch $p \in 1 \cdots N_p$ at time t is characterized as a vector $\mathbf{c}_{p,t} \in \mathcal{R}^3$:

$$\mathbf{c}_{p,t} = (\ell_p, d_{p,p^*}, \phi_{p,p*})$$

- $p \in 1 \cdots N_p$: the Patch ID
- $\ell_p \in (1, \cdots, N_\ell)$ identifies the priority map from which the patch p is generated.
- Euclidean Distance $d_{p,p*} = \|\boldsymbol{\mu}_p - \boldsymbol{\mu}_{p^*}\|$
- angle $\phi_{p,p*}$: the degree of deviation

$$x_t = [\mathbf{c}_{1,t}| \cdots |\mathbf{c}_{p,t}| \cdots |\mathbf{c}_{N_p,t}]$$

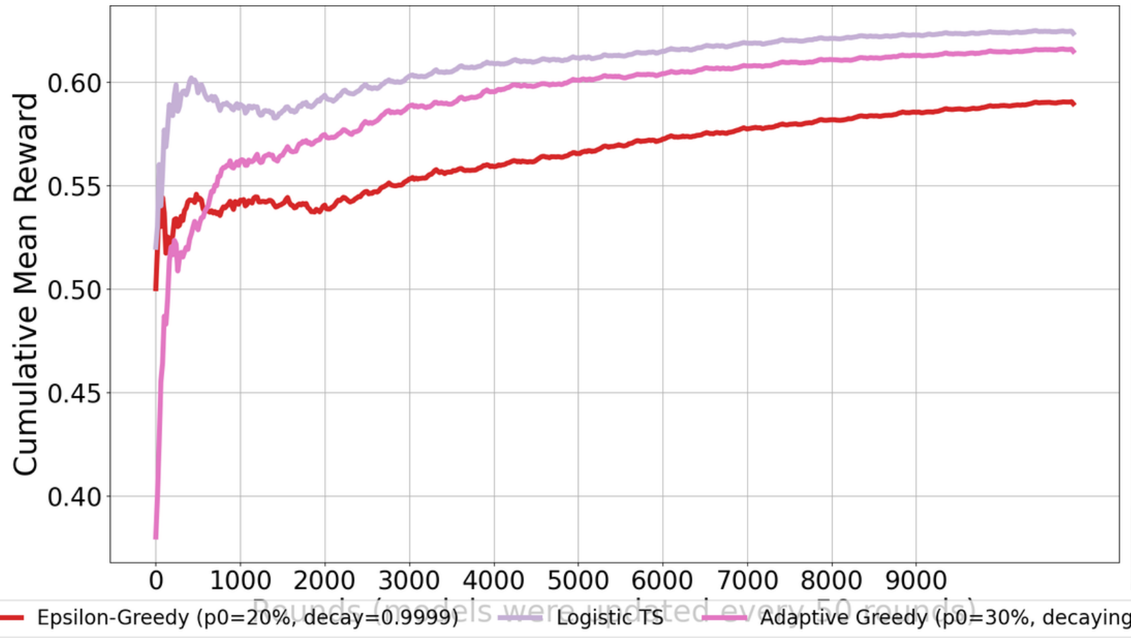


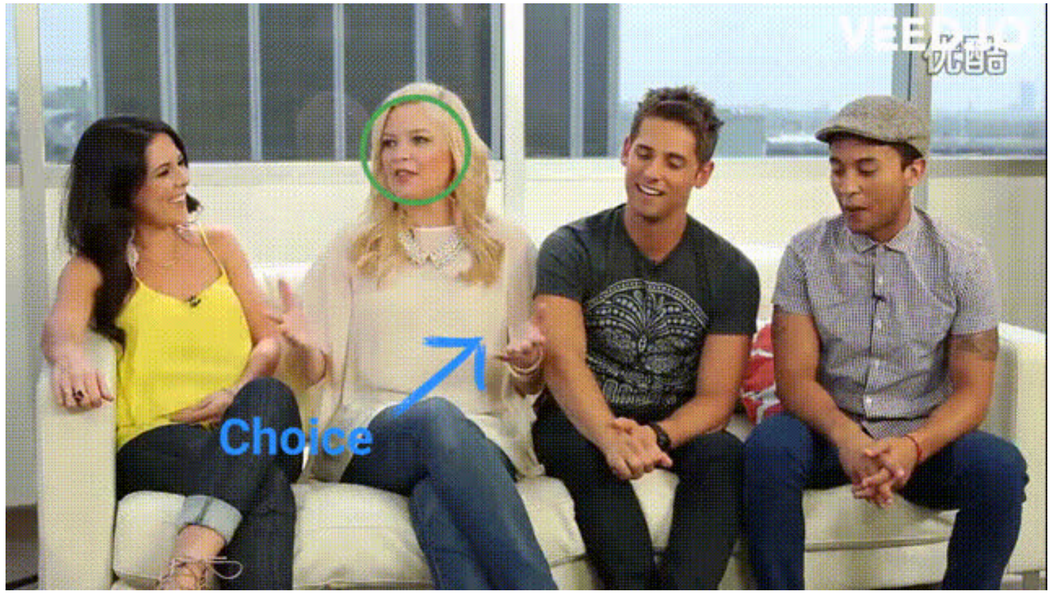
PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing
Bocchino Daniele

PHUSE LAB
UNIVERSITY OF MILAN

THOMPSON SAMPLING
CONTEXTUAL MULTI ARMED BANDITS

The Thompson Sampling algorithm enhances decision-making by effectively balancing exploration and exploitation. It selects values based on their probabilities of being the highest.





Bocchino Daniele

13 luglio **2023**

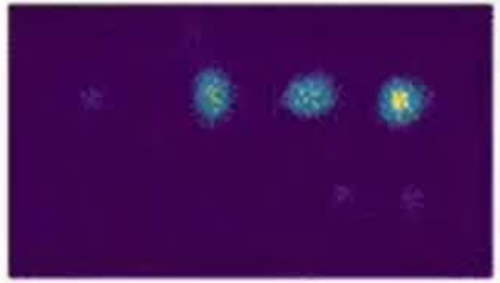
SIMULATION

Simulation was conducted involving 75 videos and 39 observers. The training phase utilized data from 10 observers, while the testing phase involved data from 29 observers.

ORIGINAL FRAME

PERCEPTUAL RAPRESENTATION

REAL FIXATION MAPS

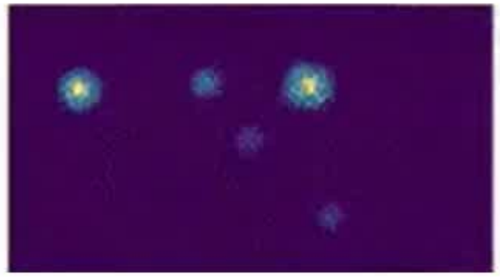


REAL

FOCUS of ATTENTION (FOA)

GENERATED GAZE DATA

GENERATED FIXATION MAP



GENERATED

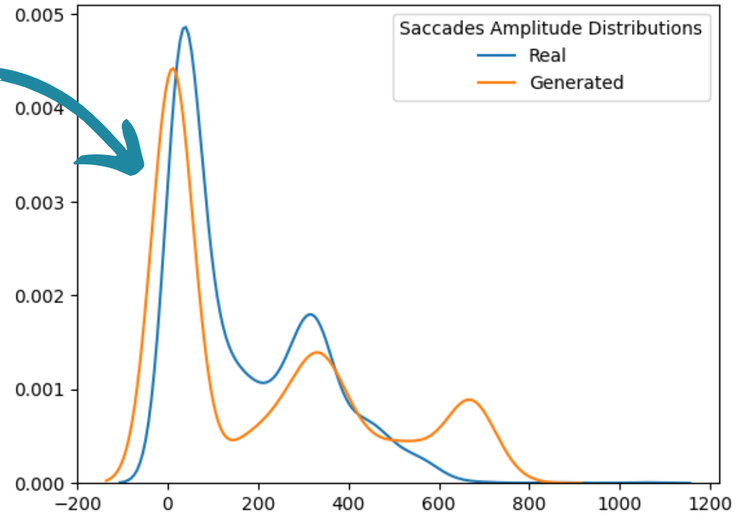
Bocchino Daniele

13 luglio **2023**

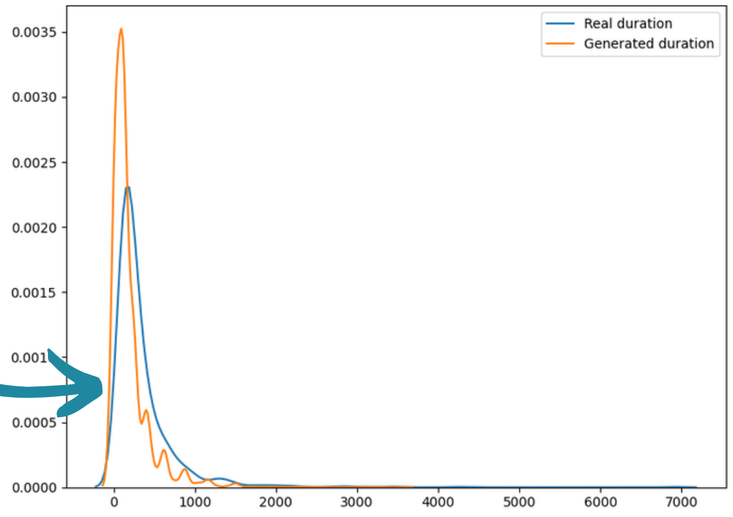
PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing

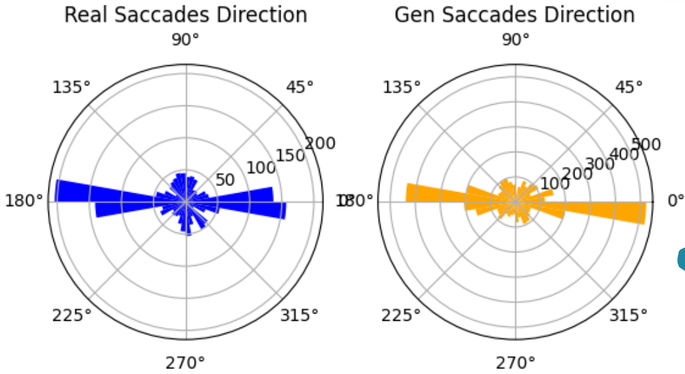
ANALYSIS OF RESULTS

Saccades Amplitude

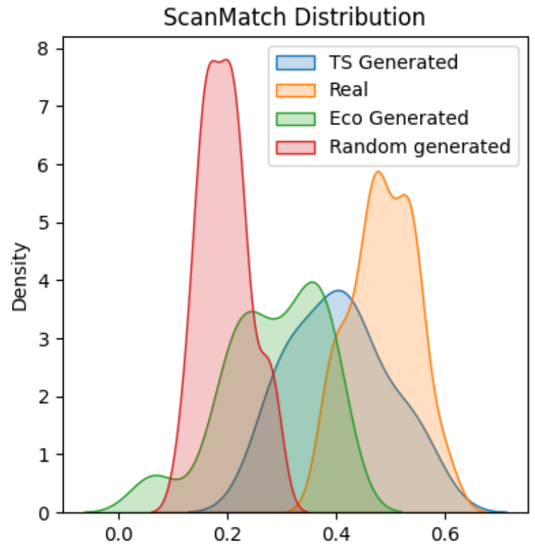


Fixation Duration





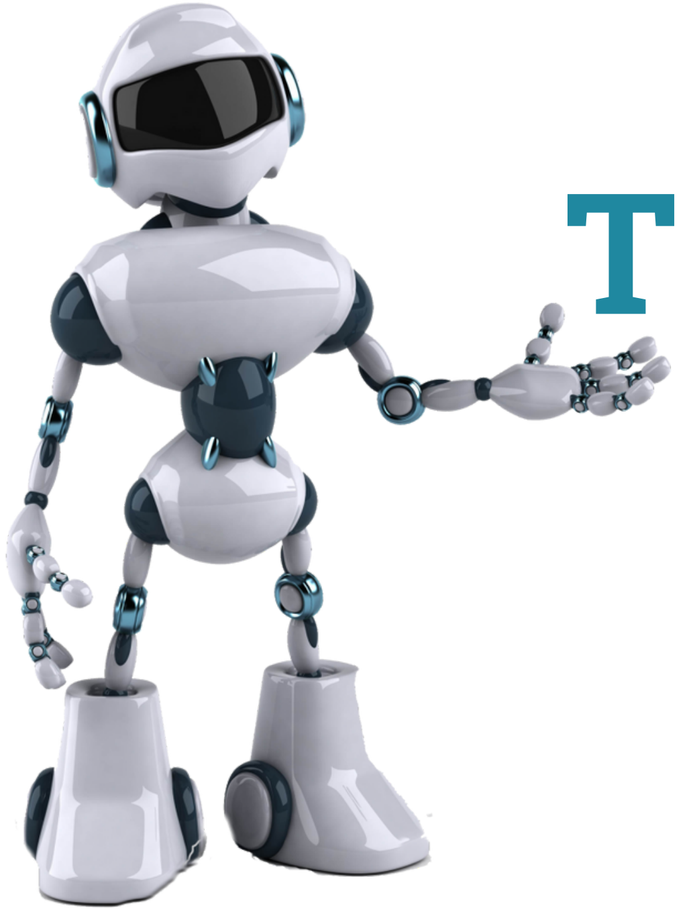
Saccades Direction



ScanMatch

Bocchino Daniele

13 luglio **2023**



THANK YOU

Bocchino Daniele

13 luglio **2023**

PHUSE LAB
UNIVERSITY OF MILAN
PHUSE Lab
Perceptual computing and HUman SEnsing