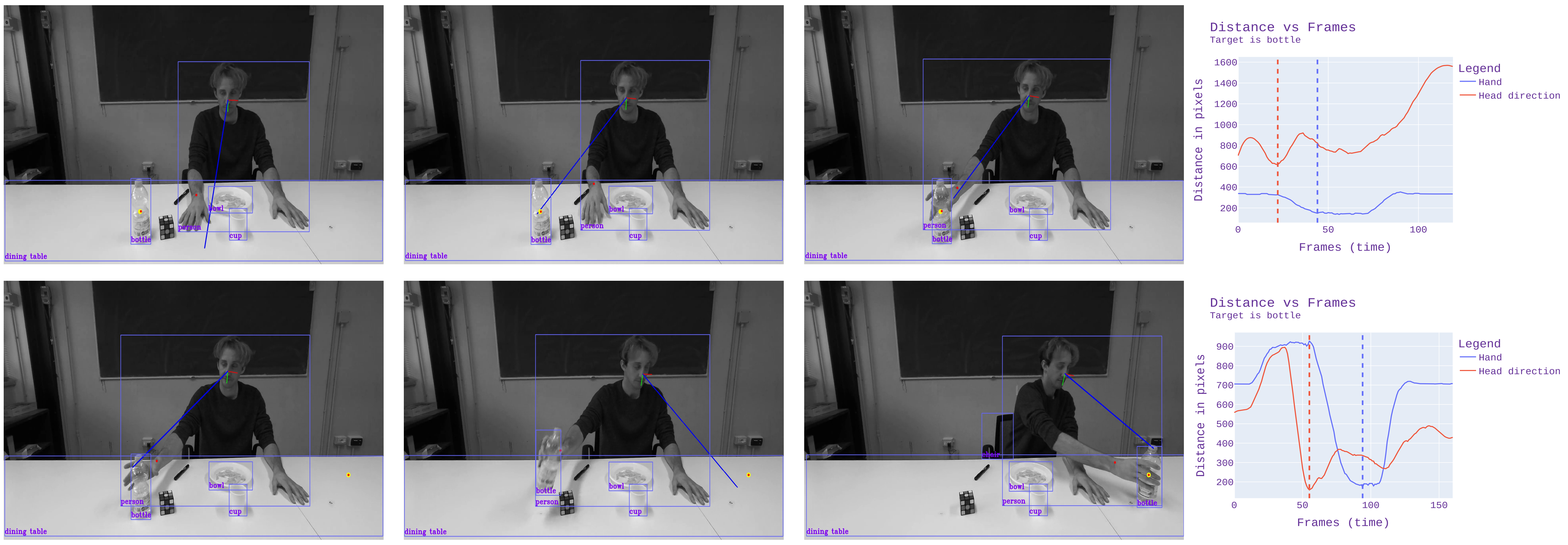
Università di Genova

DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

MaLGa

Anticipation through Head Pose Estimation: a preliminary study

Federico Figari Tomenotti, Nicoletta Noceti



Abstract & Contributions

The ability to anticipate others' goals and intentions is at the basis of human-human social interaction. Such ability, largely based on non-verbal communication, is also a key to having natural and pleasant interactions with artificial agents, like robots. In this work, we discuss **a preliminary experiment on using head pose as a visual cue to understand and anticipate action goals**, particularly reaching and transporting movements. By reasoning on the spatio-temporal connections between the head, hands and objects in the scene, we will show that short-range anticipation is possible, laying the foundations for future applications to human-robot interaction.

Method

We started from RGB frames and with a deep learning-driven approach we extracted objects (through YOLO [2]) and keypoints with Centernet [3] and then we used HHP-Net [1] for the 3D Head Pose estimation.
We then calculated distances on the pixel plane using these points as reference:

- $\mathbf{P}_O^t$ the position of the object at time t (yellow and red circle)
- $\mathbf{P}_H^t$ the position of the hand at time t (in red)
- $\mathbf{P}_G^t$ "the position" of the gaze on the table (end of the blue line)
- $\mathbf{P}_T^t$ the target position for transporting movements (yellow and red circle)

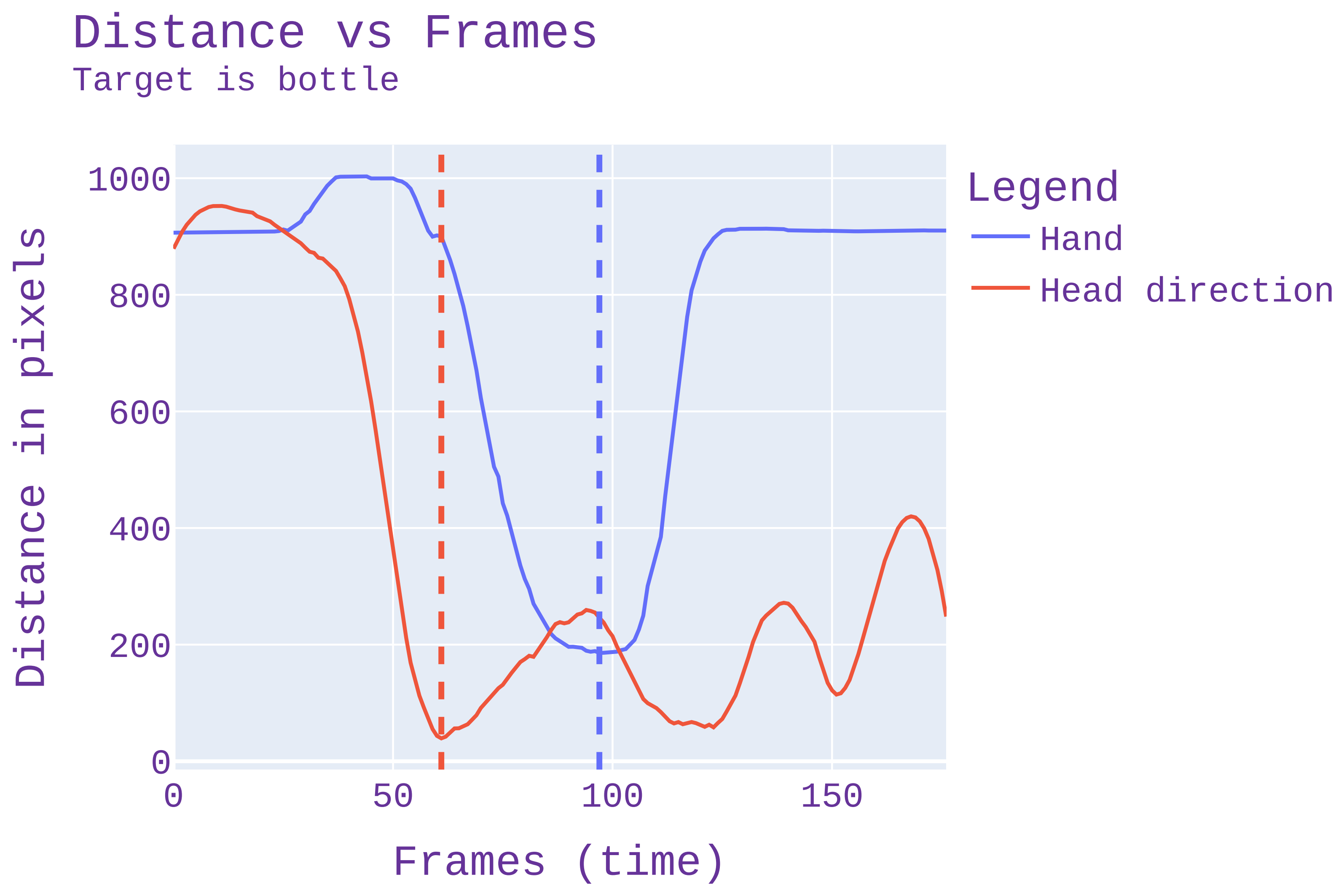
We assessed how much time in advance the head pose reaches the target with respect to the hand. We experimented with measuring two different quantities, as reported in the table at the bottom. These measures are calculated as:

- *gazing_target_time* = $\arg\min_t(\mathbf{P}_G^t - \mathbf{P}_O^t)$ or $\arg\min_t(\mathbf{P}_G^t - \mathbf{P}_T^t)$
- *touching_object_time* = $\arg\min_t(\mathbf{P}_H^t - \mathbf{P}_O^t)$
- *target_object_time* = $\arg\min_t(\mathbf{P}_O^t - \mathbf{P}_T^t)$

The *anticipation* is then estimated as:

- *touching_object_time* - *gazing_target_time* (where target is the object) for reaching movements
- *target_object_time* - *gazing_target_time* (where target is the final position).

Experimental analysis



In the image, the two curves illustrate the distances from the target for both gaze (in red) and hand (in blue). The vertical dashed lines indicate the minimum distance values used for computations: *gazing_target_time*, *touching_object_time* and *target_object_time*.

The table presents the anticipation in seconds using the head-pose projection with respect to the hand movement. The first column exposes the action class recorded in the dataset, whereas the second column shows the measured quantity for this assessment. The measures are in seconds (the minus indicates the head's anticipation with respect to the hand). All measures are averaged across 32 videos, performed by 16 subjects.

original action	measured quantity	mean [s]	std [s]	median [s]
transport bottle	reach bottle	-0.51	0.35	-0.43
touch bottle	reach bottle	-0.64	0.34	-0.63
open-close bottle	reach bottle	-0.54	0.35	-0.50
drinking	reach glass	-0.49	0.85	-0.77
transport bottle	object to target	-0.70	0.52	-0.78

CONTACTS

Federico Figari Tomenotti
federico.figaritomenotti@edu.unige.it

References

[1] G. Cantarini et al. *Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty.* In Proceedings of the IEEE WACV, 2022.

[2] J. Redmon et al. *You only look once: Unified, real-time object detection.* In Proceedings of the IEEE CVPR, 2016.

[3] X. Zhou et al. *Objects as points.* In arXiv preprint arXiv:1904.07850, 2019.