Università di Genova
DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

Human-Object and Human-Human Interaction through Head Pose Estimation

Christian Dagnino

Master Thesis

Università di Genova, DIBRIS Via Opera Pia, 13 16145 Genova, Italy
https://www.dibris.unige.it/

Università
di Genova

MSc Computer Science
Data Science and Engineering Curriculum

Human-Object and Human-Human Interaction through Head Pose Estimation

Christian Dagnino

Supervisor: Nicoletta Noceti

Reviewer: Manuela Chessa

March, 2026

Abstract

Modeling human interactions from 2D visual data is a key challenge in computer vision. Recent approaches often rely on large end-to-end deep architectures, but such models can be computationally costly and difficult to interpret. This thesis investigates the following perspective: whether head pose estimation, interpreted as a proxy for visual attention, can support interaction modeling through lightweight geometrical reasoning in monocular RGB settings.

Our proposed framework integrates human pose estimation, head pose regression and geometrical reasoning to construct interpretable attentional head direction vectors. Interaction predictions derive from explicit angular relationships between head orientation and scene elements.

We evaluate our approach in two complementary use cases. In Human–Object Interaction, head direction, hand and objects positions, and depth cues are used to estimate and anticipate object-directed interaction targets. In Human–Human Interaction, dyadic interaction events are detected by analyzing the relative orientation of head vectors and validated against clinical video coding and motion capture as groundtruths.

Experimental results show that head-based geometrical reasoning provides both useful signals in short-term interaction anticipation and strong event-level agreement in dyadic interaction detection. These findings indicate that interaction patterns can emerge from head orientation cues alone, highlighting the effectiveness of low-cost, interpretable models for real-world interaction analysis.

Table of Contents

List of Abbreviations

Abbreviation	Definition
1D / 2D / 3D	One-dimensional / Two-dimensional / Three-dimensional
CGU	Confidence Gated Unit
CNN	Convolutional Neural Network
COCO	Common Objects in Context ([26])
CV	Computer Vision
DE	Depth Estimation
DL	Deep Learning
DPT	Dense Prediction Transformer
ETMI	Event-level Temporal Matching Index
FC	Fully Connected
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
HHI	Human-Human Interaction
HHP-net	Heteroscedastic neural network for Head Pose estimation ([7])
HOG	Histogram of Oriented Gradients
HOI	Human-Object Interaction
HPE	Head Pose Estimation
Human PE	Human Pose Estimation
IIT	Italian Institute of Technology
LAEO	Look-At-Each-Other
MDE	Monocular Depth Estimation
MiDaS	Multiple Depth Estimation Accuracy with Single Network ([17])
MoCap	Motion Capture
NAO	Next Active Object
PAF	Part Affinity Field
ReLU	Rectified Linear Unit
RGB / RGB-D	Red, Green, Blue / Red, Green, Blue - Depth
RNN	Recurrent Neural Network
SIFT	Scale-Invariant Feature Transform
SURF	Speeded-Up Robust Features
SVM	Support Vector Machine
VFOA	Visual Focus Of Attention
WLS	Weighted Least Squares
YOLO	You Only Look Once ([34])

Introduction and Motivation

Understanding human interactions from visual data is a key challenge in computer vision (CV). Nowadays, intelligent systems are needed to interpret relational dynamics, such as who is interacting with what or with whom. Achieving this level of scene understanding is complex, particularly in unconstrained or monocular settings.

Recent advances in deep learning (DL) have produced powerful end-to-end architectures for interaction recognition and action anticipation ([25]). However, these approaches usually need large annotated datasets, huge computational resources, and are often limited in terms of interpretability. Moreover, deployment and scalability of these systems are reduced by their dependence on multi-camera setups, depth sensors or full 3D reconstructions ([35]).

Human interactions mainly rely on visual attention ([2]). Before manipulating an object or participating in social interaction, individuals usually orient their head and gaze toward their target. Psychological studies show that head orientation accounts for a large portion of gaze direction variance ([44]). Therefore, head direction plays an important role in social contexts. This observation suggests that head pose can represent a practical and geometrically interpretable proxy for visual attention, especially in third-person scenarios where gaze data is usually unavailable.

The motivation of this thesis arises from the following question:

Can we infer meaningful interaction patterns from head orientation alone, through geometrical reasoning in monocular RGB settings?

We explore whether a lightweight and modular pipeline can obtain interpretable interaction information only through head pose estimation (HPE) and spatial relations. The assumption is that social interactions emerge from geometrical relationships between head direction vectors and elements in the scene.
Therefore, the general objective of this thesis is to design and evaluate a geometrically interpretable framework for modeling human interactions from monocular HPE, both in human-object and human-human scenarios.

The main contributions of our work can be summarized as follows:

- We design a modular pipeline based on 2D head orientation and geometrical reasoning to analyze social interactions in images and videos;
- We empirically evaluate our framework in two complementary settings: Human-Object Interaction (HOI) and Human-Human Interaction (HHI). For HOI, we propose a depth-aware geometrical strategy for estimating and anticipating interaction targets in object-directed actions. For HHI, we implement an angular-based relational model for detecting dyadic interaction events, validating our approach against manual video coding and motion capture (MoCap) as groundtruths.

The aim of these contributions is demonstrating that interaction signals can be obtained through low-cost geometrical reasoning, without requiring explicit gaze tracking, multi-camera systems or complex end-to-end learning architectures.

The rest of the thesis is organized as follows.
Chapter 1 presents the theoretical background and related works, covering human pose and head pose estimation (HPE), gaze direction and depth estimation, and social interaction analysis (HOI and HHI). It builds the conceptual link between head orientation and interaction modeling.
Chapter 2 describes all the methodological components integrated in our framework, through a brief review of the works ([7], [39], [40]) that inspired and guided our approach. It highlights the HPE backbone and the geometrical reasoning strategies that characterize our approach.
Chapter 3 introduces the first experimental study, focused on Human-Object Interaction (HOI). It presents our depth-aware geometrical method for estimating and anticipating object-oriented interaction targets.
Chapter 4 presents the second experimental study, focused on Human–Human Interaction (HHI). It presents our angular-based relational method for detecting dyadic interaction events. In addition, it compares the results against clinical video coding annotations and MoCap predictions, through both a frame-level and an event-level analysis.
Finally, in "Conclusion and Future Works" we summarize the main results, discuss limitations and describe future research possibilities.

Chapter 1

Background and Related Works

This chapter introduces the theoretical foundations underlying the thesis.
We present this overview to build the argument that head orientation can be exploited as an interpretable proxy for visual attention, enabling interaction modeling in both human-object and human-human contexts.

For humans, participating in social interactions (with the environment or with others) is a natural and effortless task. Human visual perception naturally integrates body configuration and depth cues to interpret actions and anticipate intentions.
The same capability can be replicated in computer vision (CV) systems with models that can interpret visual information in images and videos similarly to human perception ([32], [33], [43]). To perform this task, we need structured pipelines that progressively transform pixels into semantic representations. High-level scene understanding requires taking into account spatial relationships in image sequences, moving beyond simple detection or recognition. In manipulative tasks the relevant visual information may be found on the geometrical relationships with the objects, while in dyadic interactions the focus may rely on the relative pose configurations of the subjects.

In the following sections, we discuss human pose, head pose and gaze direction estimation. Then, we briefly present monocular depth estimation, which we exploit in our pipeline. Finally, we connect these low-level cues to high-level social understanding in human-object (HOI) and human-human intercation (HHI).

1.1 Human Pose Estimation

Understanding interactions requires estimating the spatial configuration (position and orientation) of the human body. A body can be modeled as an articulated system composed

of rigid segments connected by rotational joints. Human Pose Estimation (Human PE) represents the task of localizing key body joints and reconstructing the skeleton of a person. It is a fundamental task in CV, at the basis of higher-level applications such as human action recognition, human-robot interaction, and video surveillance ([42]).

Traditionally, human body poses can be accurately retrieved using motion capture (MoCap) systems that rely on optical markers attached to the body ([19]). However, they require long preparation times, precise calibration and controlled environments. These constraints limit their practicality, making them unsuitable for many natural or unconstrained scenarios.
In contrast, vision-based (markerless) Human PE infers body configuration directly from monocular RGB, grayscale, or depth data, eliminating the need for physical markers ([10]). However, estimating human pose from monocular images is an under-constrained problem due to the loss of depth information when projecting 3D structures to 2D planes. Illumination changes and occlusions further increase the complexity of the task. Therefore, achieving accurate 3D understanding from a monocular view remains a desirable and active research goal ([21]).

Typical Human PE pipelines involve three main steps: feature extraction, body modeling, and an estimation method ([18]).
Feature extraction identifies visual cues corresponding to specific anatomical keypoints, called landmarks. Features can range from low-level (edges, textures, ...) to mid-level (HOG, SIFT, ...), or high-level representations (patches, geometrical descriptors, ...).
Body models can be categorized as kinematic (stick figures), planar (cardboard models) or volumetric models (meshes).

Human PE methods can be grouped into two classes: generative and discriminative ([18]).
Generative approaches start from a model and project it into the image space to compute a likelihood with respect to the observed data.
Discriminative approaches directly learn mappings from image features to pose configurations using training data.
Hybrid methods combine both approaches by initializing the pose with discriminative predictions and refining it with generative optimization.

Another similar classification distinguishes between top-down and bottom-up ([18]).
Top-down methods detect individuals and then estimate their poses: the computational cost scales with the number of people.
Bottom-up methods detect all body parts at once and then group them by individual: this offers more robustness to detection failures and constant computational cost with respect to the number of subjects.

Finally, methods can be also divided into regression-based and heatmap-based ([45]).
Regression-based approaches directly map image pixels to the coordinates of the joints.

Heatmap-based ones predict likelihood maps where peaks correspond to keypoint positions.

With the rise of deep learning (DL), classical approaches have been outperformed by convolutional neural networks (CNNs) and, more recently, transformer-based models ([45]). These architectures automatically learn features, offering robustness to occlusion, lighting, and body shape variation. However, challenges such as limited annotated data, depth ambiguity, and occlusion remain active research problems.

Among DL–based HPE models, we briefly present two state-of-the-art frameworks: OpenPose ([8]) and YOLO-Pose ([28]).

OpenPose is a bottom-up 2D multi-person pose estimator based on CNNs ([8]). It can work in real-time. It extracts features from images and detects body joints locations by predicting both confidence maps for keypoint detection and part affinity fields (PAFs). PAFs are 2D vector fields that associate detected body parts to individuals. All keypoints belonging to the same subject are grouped during a post-processing step.
Differently from top-down approaches, OpenPose computational cost is not affected by the number of people. Therefore, the method is efficient and robust to detection errors. However, its performance can be sensitive to input quality and lighting conditions.

YOLO-Pose is a heatmap-free, end-to-end framework that integrates into a single network both object detection and 2D Human PE ([28]). It is built upon the YOLO architecture ([34]). Therefore, it jointly predicts both bounding boxes (of people) and keypoint coordinates. This eliminates the need for complex post-processing, resulting in a constant runtime.
Recent variants, such as YOLOv8-Pose ([13]), show high accuracy and speed, making them suitable for real-time applications and scalable deployment.

The 2D task can be extended to three dimensions (3D Human PE), recovering the 3D spatial positions of joints. According to [38], such 3D methods are further classified depending on input modalities (monocular RGB, depth sensors, or multi-view setups) and number of subjects (single-person, or multi-person).
While 2D Human PE has reached high accuracy with CNNs and transformers, 3D Human PE remains challenging due to the implicit loss of depth information in monocular images.

In summary, Human PE has evolved from model-based, marker-dependent methods to deep, markerless, data-driven systems suited for real-time multi-person analysis. This capability represents the starting point for higher-level CV tasks such as head pose estimation (HPE) and gaze direction inference.
In this thesis, we exploit 2D markerless Human PE as a fundamental step for extracting from monocular RGB data the facial keypoints required for HPE. Our vision-based approach is therefore less costly and more scalable with respect to marker-based systems.

1.2 Head Pose Estimation (HPE)

Head Pose Estimation (HPE) consists in determining the position and orientation of a person's head with respect to a reference coordinate system. It provides fundamental information about the direction in which a person is looking and focusing the attention. Therefore, HPE plays a central role in understanding interactions (HOI and HHI) in many cases of application, including attention monitoring and human-robot interaction.

In general, head pose can be described in terms of its translational (x, y, z) or rotational components ([3]). Rotational orientation is commonly represented by three angles: yaw, pitch, and roll. They respectively correspond to the rotations around the vertical, lateral, and longitudinal axes. The origin of the rotation in 3D space usually corresponds to the head center or the nose.
Several mathematical representations of 3D rotations exist, such as Euler angles (Tait-Bryan angles convention), quaternions and rotation matrices. Euler angles are the most intuitive, but may suffer from problems like ambiguity or gimbal lock, a phenomenon in which two axes become parallel, resulting in the loss of one degree of freedom.

Depending on the available input data, HPE can be performed using 2D (RGB) images, depth (RGB-D) data, or videos.
In monocular RGB settings, the task is non-trivial due to the lack of explicit depth information. Another challenge is the data variability introduced by different facial expressions, illumination, and occlusions.
Pre-processing typically involves face detection, facial landmark localization or head modeling ([3]). Landmark detection aims to identify specific keypoints on the face (nose, eyes, ears, mouth), which can be used as visual cues for estimating head orientation. As an alternative, some approaches use 3D head models to determine geometrical correspondences between generical head representations and image features.

From a representational point of view, HPE can be performed both in 2D and in 3D.
2D HPE computes the projection of head orientation onto the image plane. Such estimate is often sufficient for fast and approximated orientation estimation.
3D HPE recovers the full 3D rotation (and sometimes translation) of the head, expressed as a rotation matrix, or in terms of Euler angles. Although 3D HPE provides more information, it is also more challenging, especially when inferred from monocular images where self-occlusion and depth ambiguity must be resolved.

From a methodological point of view, HPE approaches can be categorized into two groups. They can be classified into landmark-based and landmark-free ([3]).
Landmark-based techniques rely on the detection and localization of specific facial keypoints (usually 5 to 7) to estimate the head's orientation in 3D space. These methods offer precise localization and interpretable results. However, they are sensitive to land-

mark detection quality and may fail under partial occlusions, extreme head rotations, or poor lighting conditions. They can also be computationally costly when tracking many landmarks or multiple people in real time.
In contrast, landmark-free approaches infer head orientation directly from image or feature representations. They usually provide better robustness to occlusion and reduced computational cost.

Alternatively, methods can be classified into training-free or training-based ([1]).
Training-free techniques rely on handcrafted features and geometrical constraints. For this reason, they often require time-consuming manual design and domain knowledge. Although slower to develop, they are usually more generalizable and do not require time-costly retraining once deployed.
Training-based (DL) approaches automatically learn discriminative features from data, achieving state-of-the-art accuracy.
The choice between these two categories depends on the application environment and on the available computational resources.

Recent advances in DL have significantly improved the performances of HPE techniques. For example, convolutional and transformer-based models can now directly estimate head pose angles from images, as a single task or jointly with facial keypoint detection ([23]). These methods enable robust real-time HPE in both controlled and unconstrained settings.

In summary, HPE connects low-level facial analysis and high-level cognitive tasks by providing directional cues about human attention. It also serves as a fundamental intermediate step for more advanced and refined analysis such as gaze direction inference.
In this thesis, head pose plays a central role: it provides a geometrically interpretable representation of human visual orientation, bridging low-level perception and high-level interaction modeling. We retrieve 3D HPE from monocular images through a landmark-based approach, projecting it back to 2D to obtain head direction in image space.

1.3 Gaze Direction Estimation

Gaze direction indicates where a person is visually focusing: it refers to the eyes orientation toward a specific point in the visual field. Therefore, it represents a fundamental element for understanding how humans interact with the environment.
Estimating gaze direction is an essential task in both CV and cognitive sciences, as it provides information about visual attention and social interactions. In CV, gaze estimation is based on visual cues extracted from images or videos. Combined with object detection, it enables the analysis of the relationships between people and objects, as well as between individuals. Therefore, understanding gaze behavior allows machines to anticipate human intentions and analyze social interactions.

From a psychological point of view, gaze plays a dual role in social communication.
According to early studies ([2]), looking behavior has both monitoring and communicative functions within social interactions. During dyadic communication, individuals continuously observe and respond to each other's gazes. Such behaviors can represent valuable nonverbal cues even in interactions with non-human agents.
Following the classification presented in [9], gaze patterns can be categorized into distinct states: shared gaze (focus on same object), mutual gaze (looking-at-each-other or LAEO), single gaze (looking at), miss (being looked at), and void (no interaction).
In psychology, atypical gaze behaviors are also important diagnostic indicators, as they can reflect developmental difficulties.

From a computational point of view, the goal of gaze estimation is determining the visual axis (line connecting eyes and gazed point) and its geometrical relationship with the optical axis of the camera. However, accurate gaze estimation is still challenging due to factors such as head pose variance, occlusions (eyeglasses or hair) and illumination changes.

Psychological studies suggest that head orientation accounts for a large portion (70%) of visual gaze direction, highlighting its strong correlation with head pose ([44]). Therefore, in real-world third-person scenarios where high-resolution eye images or infrared sensors are unavailable, head pose represents a robust and practical proxy for gaze direction. In addition, the assumption that most people look and orient their head where they are interacting forms the basis for applications in social scene understanding.
As presented in [30], landmark-free HPE techniques based on RGB or RGB-D data can provide robust approximations of gaze direction, even under low-resolution or partially occluded conditions.

Moreover, since eyes orientation can differ from head direction by only $\pm 35^\circ$ ([37]), head direction represents a robust gaze approximation. This is especially useful for third-person settings where eyes are not clearly visible, and their rotation cannot be precisely inferred without direct eye tracking.
Therefore, all head-based gaze estimation methods intrinsically involve uncertainty.

To estimate apparent gaze direction and its uncertainty, in [12] and [22] it is proposed a simple and lightweight architecture (for a multi-camera assisted living scenario). The method relies on the extraction of specific facial keypoints (eyes, ears, and nose) collected from a generic pose estimation model (OpenPose [8]). Facial centroid and direction vector are computed exploiting the position of such keypoints. Finally, gaze vector is the projection onto the image plane of the unit vector centered at the facial centroid.
Similarly to the Gated Recurrent Units (GRUs) of the recurrent neural networks (RNNs), the architecture involves Confidence Gated Units (CGUs) with a rectified linear unit (ReLU) and a sigmoid unit. The full network consists in 10 CGUs as input layer, 2 fully connected (FC) hidden layers with 10 units each, and an output layer with 3 units.
The 3D output vector represents the gaze direction and its associated uncertainty, strongly

correlated with actual angular error in gaze estimation. Such uncertainty depends on the input, and in particular on keypoint visibility: as in human perception, a prediction is more reliable when the entire face is visible.

In summary, head-based gaze direction estimation and its intrinsical uncertainty provide a sufficient indicator of human visual interactions in third-person scenarios.
In this thesis, we exploit HPE as a reliable approximation of gaze direction. We base our methodological work on the assumption that interaction targets (objects in HOI or humans in HHI) can be inferred by reasoning geometrically only on head orientation vectors.

1.4 Depth Estimation

Depth estimation represents another significant element in scene understanding.
Among vision-based (markerless) systems, a clear distinction can be made whether a depth map is produced or not ([10]). A depth map is an image in which each pixel is associated with the distance of the corresponding scene point from the camera.
Depth-sensing (RGB-D) systems can be implemented using binocular stereo configurations or active sensors that infer depth by projecting light patterns into the scene.

Depth perception enables humans to interpret the 3D structure of their surroundings (objects and environment) by exploiting binocular disparity and other visual cues ([21]). Analogously, in CV, depth estimation (DE) provides spatial information that allows machines to perform tasks such as object localization, 3D reconstruction, and scene understanding.

DE methods can be categorized into three main types: stereo, monocular, and multi-view approaches ([35]).
Early methods relied on vanishing points or focus-defocus relations, while other traditional techniques rely on handcrafted features (SIFT, SURF, . . .) based on stereo correspondence. In contrast, modern DL-based approaches automatically learn features from large datasets, achieving improved robustness and generalization.
Stereo-based methods compute depth by measuring binocular disparity between images taken by two synchronized cameras placed at slightly different points of view.
Multiview approaches combine information from multiple perspectives, improving depth accuracy in complex scenes, with the drawback of higher computational cost and setup complexity.
Monocular depth estimation (MDE) techniques, instead, infer depth from a single image. This is an underconstrained (ill-posed) problem, as a single 2D projection can correspond to multiple 3D configurations. However, DL techniques have significantly improved MDE performances by exploiting sophisticated neural network architectures and rich datasets.

MDE is appealing because it only requires a single camera, making it cost-effective and

suitable for real-time applications.
Among state-of-the-art MDE methods, MiDaS (Multiple Depth Estimation Accuracy with Single Network) is one of the most accurate and widely adopted ([17]). It is based on CNNs and produces relative depth maps from single frontal-view 2D images, where brighter pixels correspond to closer objects. Its accuracy depends on image quality, training data, and preprocessing conditions. MiDaS offers the advantages of speed, simplicity, and independence from specialized hardware, making it suitable for real-time or resource-limited applications. The models released with MiDaS v3.1 achieved high accuracy and strong generalization across different environments ([6]).
More recent and advanced transformer-based architectures, such as DPT (Dense Prediction Transformer), further improve depth estimation accuracy and scene consistency. However, they are usually more computationally costly.

In summary, depth estimation enables CV systems to exploit the 3D structure of the environment emulating human depth perception.
In this thesis, we use monocular depth estimation as a complementary cue to support geometrical reasoning in the HOI scenario presented in Chapter 3.

1.5 Social Interaction Understanding

In the previous sections, we presented the low-level and mid-level vision-based concepts that are at the basis of scene understanding. Higher-level social understanding emerges from relational reasoning. Therefore, it requires moving beyond the simple estimation of physical aspects such as pose and gaze.

The aim is moving toward the concrete interpretation of object-centric or human-centric interaction dynamics. We illustrate how head pose can serve as a powerful tool to study social behavior in HOI and HHI.

1.5.1 Human-Object Interaction (HOI)

Humans heavily rely on nonverbal visual cues, such as gaze and head motion, to infer others' intentions ([15]). During manipulative tasks, gaze shifts toward the target object before the hand moves, providing predictive information about the future action. Analyzing anticipatory patterns is fundamental for building systems that can interpret human intentions in real time.

In the context of HOI, action anticipation concerns the prediction of goal-directed behaviors. In CV, this means predicting what a person is likely to do based on visual observations of their current motion, pose and visual attention. This ability mirrors human cognition,

where people continuously process visual signals to infer intentions and prepare responses. Anticipation can be formulated from both egocentric (first-person) and exocentric (third-person) perspectives. Third-person point of view is more common and has a longer studying history, because of the early availability of exocentric video datasets.

The concept of visual focus of attention (VFOA) represents a bridge between low-level perceptual cues and high-level action prediction. The VFOA, defined in [11] as the behavioral and cognitive process indicating where and at what a person is looking, can be inferred from head pose and gaze dynamics. It often precedes the start of an action.
This relationship between visual and motor behavior is essential for anticipating manipulative actions in HOI contexts.

A recent survey ([25]) categorizes anticipation tasks according to the considered temporal window and the prediction objective. This classification includes: next-action anticipation, goal and final action anticipation, short-term and long-term anticipation.
Action anticipation methods range from classical machine learning models, such as SVMs, to more recent architectures based on RNNs, transformers, and large language models.

In summary, in Human-Object Interaction (HOI), attention usually precedes manipulation. Therefore, head orientation strongly contributes in the analysis of anticipation patterns in such scenarios.
In this thesis, we analyze the geometrical relationships between head orientation, hands and objects positions, and depth cues to explore temporal anticipation between gaze and motion in HOI (Chapter 3).

1.5.2 Human-Human Interaction (HHI)

Nonverbal cues represent the most trivial but at the same time complex manifestations of human social behavior. This includes how individuals use head orientation, gaze and gestures to express intention and attention in nonverbal communication.
Among the various visual cues, eye contact is considered a fundamental aspect of social engagement and dyadic (face-to-face) interaction. In literature, eye contact has been studied across very different contexts: from conversational and group dynamics to developmental and clinical studies. The presence (or absence) of gaze contact provides strong signals about interpersonal connection and attention.

Adopted approaches can be grouped into direct (real-time) and indirect (video-based) methods ([24]). Traditional analysis usually relies on manual annotation, which is time-costly and subjective. In contrast, automated video-based approaches offer faster and more objective alternatives.

Head orientation plays a central role in encoding and decoding these social signals. As

shown in [41], fast or small changes in head direction can indicate engagement (or disengagement) during dyadic interactions.
Automated systems based on 3D MoCap or video-based HPE enable continuous analysis of such dynamics, overcoming the limitations of manual coding. Such approaches must still balance precision with interpretability, as not every micro-movement corresponds to a meaningful event.

In behavioral analysis, the task of detecting whether two people are looking at each other (LAEO) has become a standard benchmark for dyadic interaction recognition. As presented in [29], estimating mutual gaze involves the combination of head detection, HPE and geometrical reasoning in both 2D and 3D spaces to infer when two gaze directions intersect. Similarly, the object-driven concept of shared or joint attention, which happens when two individuals focus on a common target, represents a powerful basis for understanding social and collaborative behaviors. The problem of inferring joint attention in third-person videos is addressed in [16], where the authors propose a spatio-temporal neural network that exploits both gaze directions and potential target boxes.

Studying social interactions is also useful in developmental and clinical contexts. For example, individuals with autism spectrum disorder usually show aypical head orientation and gaze patterns ([4]). In such settings, assistive technologies have been explored to support diagnosis and intervention.

In summary, Human-Human Interaction is characterized by episodes of social engagement, such as looking-at-each-other (LAEO) or shared attention. Therefore, reciprocal dyadic head orientation patterns indicate the type of social interaction.
In this thesis, we explore the relative head orientations of individuals to understand the type of HHI (Chapter 4).

Chapter 2

Proposed Pipeline

This chapter describes the methodological components that we use to build the pipelines proposed in Chapters 3 and 4, for HOI and HHI respectively.
Section 2.1 introduces such components through a brief review of the literature works that inspired and guided our approach.
Section 2.2 explains how we integrate these methodological components into our general pipeline. The proposed framework is then specifically refined in Chapters 3 and 4 for both HOI and HHI scenarios to differently deal with the two presented use cases.

2.1 Methods

In this section, we introduce the main works that inspired and guided us in our methodological and experimental development. With this brief overview, we technically describe all the components that we integrate into our framework.

The considered studies ([7], [39], [40]) cover research topics and application domains closely related to head pose estimation in human-object and human-human interaction scenarios. Specifically, they span from 3D head pose estimation and 2D head direction projection, to the anticipation of manipulative actions and the analysis of dyadic interactions.

HHP-net: A Lightweight Method for Head Pose Estimation

In [7] it is introduced HHP-net, a lightweight neural network for estimating head pose Euler angles (yaw, pitch, and roll) from single RGB images. As in [12] (presented in Section 1.3), the approach relies on the extraction of a small set of facial keypoints (eyes, ears, and

nose) detected by a generic 2D pose estimator.
The architecture is compact (0.5 Megabytes), efficient, and designed for real-time applications. The input consists in a set of n keypoints located on the image plane: $\{(x_1^i, x_2^i, c^i)\}_{i=1}^n$, where x_1^i, x_2^i and c^i are respectively horizontal and vertical coordinate, and confidence of the i-th keypoint. Coordinates are centered (with respect to their centroid) and normalized (with respect to their maximum value). The confidence value c^i is in the range $[0, 1]$.
The full HHP-net architecture is presented in Figure 2.1.

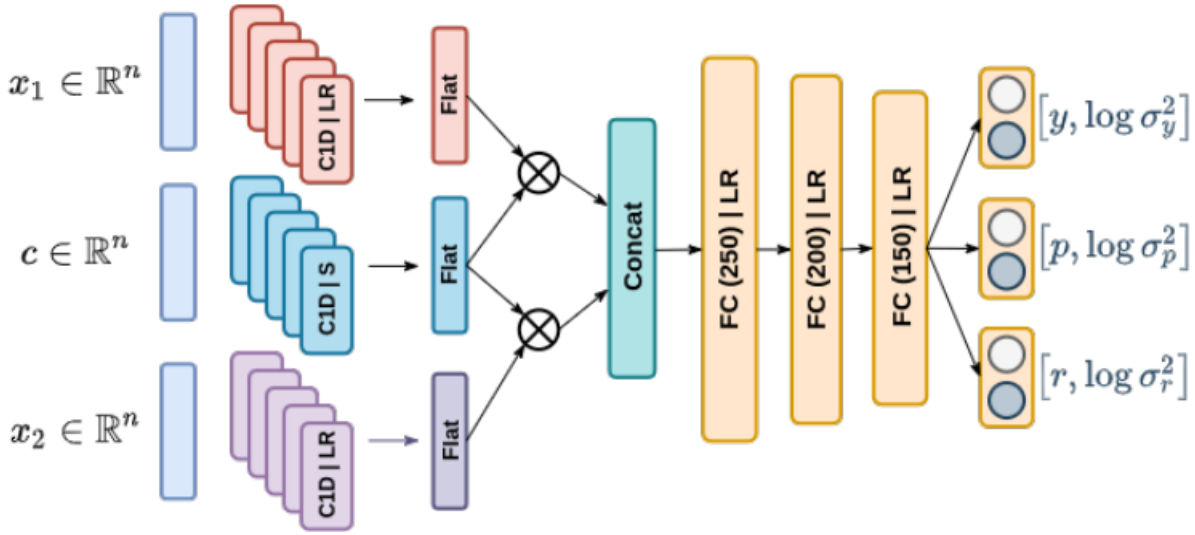


Figure 2.1: The figure is taken from [7]. It shows the end-to-end HHP-net architecture. C1D = 1D convolution, S = sigmoid, LR = leakyReLU, $\bigotimes$ = element-wise multiplication.

The input vectors are $\mathbf{x}_1 = [x_1^1, \ldots, x_1^n]$, $\mathbf{x}_2 = [x_2^1, \ldots, x_2^n]$ and $\mathbf{c} = [c^1, \ldots, c^n]$. They are processed independently with a 1D convolutional layer, followed by non-linear activations: Leaky ReLU for $\mathbf{x}_1$ and $\mathbf{x}_2$, sigmoid for $\mathbf{c}$. These initial results ($\mathbf{x}_1^*$, $\mathbf{x}_2^*$ and $\mathbf{c}^*$) are flattened and combined using element-wise multiplication, following the CGU logic proposed in [12] (and presented in Section 1.3). The two gated outputs, $\mathbf{v}_1 = \mathbf{x}_1^* \otimes \mathbf{c}^*$ and $\mathbf{v}_2 = \mathbf{x}_2^* \otimes \mathbf{c}^*$, are concatenated. The resulting vector is provided to a sequence of three fully connected layers (of respectively 250, 200 and 150 neurons), each one including a Leaky ReLU to avoid vanishing gradients. The three output layers returns the estimated head pose Euler angles (yaw, pitch, and roll), each one accompanied by an uncertainty value.

This pipeline enables a lightweight, accurate and interpretable estimation of head orientation in unconstrained visual conditions.

Projecting 3D Head Pose Estimation in 2D Image Space

To visualize the output of the pipeline presented in [7], the 3D vector of angles $\mathbf{q} = [q_y, q_p, q_r]$ can be projected onto the 2D image plane according to Tait-Bryan angles convention.

As better formalized in [39], the projections are computed as follows:

$$\begin{aligned} x_r &= \cos q_y \cdot \cos q_r + \Delta_x \\ y_r &= \cos q_p \cdot \sin q_r + \cos q_r \cdot \sin q_y \cdot \sin q_p + \Delta_y \\ x_g &= -\cos q_y \cdot \sin q_r + \Delta_x \\ y_g &= \cos q_p \cdot \cos q_r - \sin q_y \cdot \sin q_p \cdot \sin q_r + \Delta_y \\ x_b &= \sin q_y + \Delta_x \\ y_b &= -\cos q_y \cdot \sin q_p + \Delta_y \end{aligned} \tag{2.1}$$

In Equation 2.1, (x_r, y_r), (x_g, y_g) and (x_b, y_b) are the coordinates of the endpoints of three vectors (red, green and blue vectors, as called in [39]), while (Δ_x, Δ_y) is their common application point. These vectors represent the 2D projected orientation axes of the head's 3D local reference frame, corresponding respectively to the rotated $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ directions. Among all three vectors, the one pointing forward from the head (with endpoint (x_b, y_b) in Equation 2.1) is considered as the 2D head direction.

Using Head Pose Estimation for Estimating Action Anticipation

Building upon HHP-net ([7]), in [40] it is explored the anticipatory value of head pose information in manipulative object-directed actions.
The paper discusses a preliminary experiment on short-term action anticipation (in reaching and transporting actions) by exploiting head pose as the only visual cue.

The proposed pipeline integrates a human pose detector (CenterNet [14]), an object detector (YOLOv8 [34]), and HHP-net for head pose estimation.
As in [12] (presented in Section 1.3), it is hypothesized consistency between gaze and head direction. The 2D head direction consists in the projection (according to Tait-Bryan angles convention) on the image plane of the three Euler angles obtained from the HHP-net.

The anticipatory analysis is based on three temporal instants in HOI involving reaching and transporting actions: the "gazing_target_time", the "touching_object_time" and the "target_object_time". These respectively correspond to the moments where the subject looks at the target (object for reaching, final position for transporting), touches an object (for reaching) and places the object in target position (for transporting).
The hypothesis is that "gazing_target_time" always anticipates the other two instants.

From a practical point of view, at a generic time t:

- Object position ($\mathbf{P}_O^t$) is represented by the center of its bounding box;
- Hand position ($\mathbf{P}_H^t$) is approximated by the wrist keypoint;
- Gaze "position" ($\mathbf{P}_G^t$) is identified by the endpoint (x_b, y_b), in Equation 2.1, of the head direction vector (for reaching);
- Target position ($\mathbf{P}_T^t$) is fixed and available as a prior knowledge (for transporting).

Therefore, the three relevant instants are defined as follows:

- "gazing_target_time" is the time t where $\mathbf{P}_G^t - \mathbf{P}_O^t$ (for reaching) or $\mathbf{P}_G^t - \mathbf{P}_T^t$ (for transporting) are minimal;
- "touching_object_time" is the time t where $\mathbf{P}_H^t - \mathbf{P}_O^t$ is minimal (for reaching);
- "target_object_time" is the time t where $\mathbf{P}_O^t - \mathbf{P}_T^t$ is minimal (for transporting).

Anticipation is then calculated as the difference between "touching_object_time" and "gazing_target_time" (for reaching), or between "target_object_time" and "gazing_target_time" (for transporting).

By reasoning on spatial and temporal relationships between head, hands, and objects, the study shows that short-range action anticipation from head direction is possible.
In this thesis, we integrate the HHP-net method in our pipeline to extend the scope of this preliminary study ([40]). Our experimental HOI use case is presented in Chapter 3.

Using Head Pose Estimation for Detecting Dyadic Interactions

Building upon HHP-net ([7]), in [39] head pose estimation is exploited as a cue for understanding social interactions.
The paper discusses an experimental analysis of the HHP-net method ([7]), presented as a plug-in to any pose estimator, as well as an application to dyadic interaction detection. The analysis is based on different benchmarks, with comparisons with other approaches.

The proposed pipeline integrates generic human pose detectors (OpenPose [8], CenterNet [14], and MediaPipe [27]) and HHP-net for head pose estimation.
Once again, gaze is approximated with head direction, which consists in the Tait-Bryan projection on the 2D image plane of the angles obtained from the HHP-net.

The authors empirically show that, in terms of accuracy and computational cost, the method performs equally well or better than other approaches. Moreover, an application to social interactions is presented to prove that the pipeline can be exploited to automatically

detect dyadic events (LAEO) in images and videos.
The lightweight HHP-net approach is shown to be generally accurate and robust under different challenging input conditions.

The study demonstrates that combining reliable head pose estimation with geometrical reasoning can allow the inference of simple social interactions directly from visual data. Therefore, it provides a bridge between low-level perception and high-level social understanding.
In this thesis, we integrate the HHP-net method in our pipeline to analyze dyadic interactions (between mother and child) through head pose estimation. Our experimental HHI use case is presented in Chapter 4.

2.2 Adopted Methodological Components

This thesis integrates lightweight and interpretable modules into a coherent geometrical reasoning pipeline. Our design choices prioritize monocular RGB input, computational efficiency, modular structure and explicit geometrical interpretability. Here, we explain how we integrate into our pipeline all the different methodological components presented in Section 2.1.

The studies considered in Section 2.1 demonstrated that it is possible to use head pose alone for both short-term action anticipation ([39]) and dyadic interaction detection ([40]). Unlike end-to-end learning-based interaction classifiers, our proposed approach preserves interpretability, as predictions directly depend on measurable geometrical quantities.

The central component of our framework is the HHP-net ([7]).
As we already presented in Section 2.1, it is a compact neural network designed to estimate head pose angles from facial keypoints extracted by a generic 2D pose detector. In summary, it processes normalized keypoints coordinates, incorporates confidence gating and produces angle estimates with their associated uncertainty.
The architecture enables robust head orientation estimation in monocular scenarios, making it suitable for the experimental interaction analysis that we present in this thesis.

The concept of confidence gating introduced in [12] (presented in Section 1.3) highlights the importance of uncertainty when estimating head direction as a proxy for gaze.
In this thesis, we consider uncertainty values as informative signals that reflect occlusion (in the input) and confidence in geometrical reasoning (in the output).

Head pose angles are projected onto the image plane using Tait-Bryan angles convention (Equation 2.1), enabling the construction of 2D head direction vectors.
In this thesis, after the projection we perform geometrical reasoning by evaluating angular relationships between:

- Head direction vector and object centroids (HOI);
- Head direction vectors of two individuals (HHI).

The minimal version of our pipeline (common for both HOI and HHI scenarios) is presented in Figure 2.2, while Figures 3.1 and 4.1 (in Chapters 3 and 4) show the specific refinements of our framework for HOI and HHI respectively. The steps in red are based on the literature, while the steps in green represent the main novelty of this thesis.

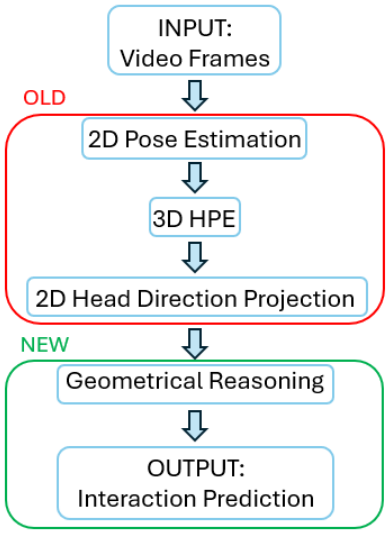


Figure 2.2: Our general (minimal) end-to-end pipeline. The steps highlighted in red are based on the literature ([7], [39], [40]), while the ones in green are the main novelty.

Specifically, starting from each video frame, 2D head direction is computed as follows:

- 2D Pose Estimation: Five facial keypoints (nose, left/right eyes and left/right ears) are extracted with a generic pose estimator. In this thesis, we adopt YOLOv8-Pose [28] as it provides a good trade-off between accuracy and time consumption;
- 3D Head Pose Estimation: Keypoints and confidence values are provided as inputs to the HHP-net, that outputs three Euler angles with their associated uncertainties;
- 2D Head Direction Projection: Angles are projected using Equation 2.1. Among the three orthogonal axes of the head reference frame, only the direction pointing forward from the head (with endpoint (x_b, y_b)) is retained. Such unit vector represents the 2D head direction in image coordinates.

After these components, this thesis extends the framework with geometrical reasoning:

- In HOI, geometrical reasoning is applied to estimate and anticipate interaction targets from head-hand dynamics and spatial (depth-aware) cues;
- In HHI, geometrical reasoning is applied to detect and validate dyadic interactions from relative head orientation between subjects.

As output of the pipeline, we obtain information on the occurring type of interaction.

Both use cases, detailed in Chapters 3 and 4, share the same common assumption that social interactions emerge from geometrical relationships between head direction vectors and elements in the scene.

Chapter 3

Use Case 1: Human-Object Interaction

This chapter presents the first use case of application of the thesis, focusing on Human–Object Interaction (HOI).
We investigate how 2D head direction, obtained through a lightweight Head Pose Estimation (HPE) pipeline, can be exploited as a proxy for gaze direction in simple goal-directed and object-centric actions. In particular, the relationship between head direction (visual attention) and hand motion (physical attention) is examined in short manipulative (reaching and transporting) tasks.

The use case is based on the experimental setting introduced in [40], but extends its scope. Such study focused mainly on quantifying the anticipation between gaze and hand. Here, we also introduce a new goal: we aim to identify in each video, primarily at frame-level, the object which is going to be acted upon, and follow the interaction by combining head pose, hand proximity, and depth-aware spatial reasoning.
From now on, we refer to the identified interaction object of each video as the Next Active Object (NAO) of the scene.

Although we extend the scope of [40], our work should still be considered preliminary, since we only test the pipeline on a small set of data.

3.1 Setting

The videos we base our experiment upon are a small subset of the Stereo-HUM dataset which is considered and partially used in [40]. Stereo-HUM is an action classification dataset of 10 actions (drink, eat crisps, open and close a bottle, play with a Rubik's cube,

sanitise hands, touch a bottle, touch a Rubik's cube, transport a bottle, transport a pen, transport a Rubik's cube) acquired in-house and introduced in [31]. In such dataset, each action (elementary interaction with an object) is performed twice by each one of the 16 participants, for a total of 320 short RGB videos, with an approximate length of 5–10 seconds each.

In our work, we select 5 actions (one more than [40]) characterized by the presence in the scene of a well-defined target object which is also present in the COCO classification labels ([26]). The selected actions are: drink from a cup (reaching and transporting back), eat crisps from a bowl (reaching), open and close a bottle (reaching and transporting back), touch a bottle (reaching) and transport a bottle (reaching and transporting to a target location).
For our experimental analysis, we consider only the actions (both trials) performed by one random participant, for a total of 10 videos.

The scene setup is controlled and substantially the same across different videos:

- The subject always sits behind a desk;
- Different (unique) objects are placed on the desk (bottle, cup, bowl, ...);
- All objects remain visible throughout the entire video;
- During each video, only one object is actually involved in the interaction.

This configuration allows to unambiguously define a posteriori the interaction object (NAO) as a groundtruth. Moreover, the simplicity of each scene enables the evaluation of whether such object can be correctly inferred using only visual cues, such as head direction.

3.2 Method

We propose a method that aims to identify throughout each video (at a frame-level) the interaction object (NAO) by integrating information from head pose, hand position, and scene geometry. Such information is computed frame-by-frame from monocular RGB input videos.

The full pipeline of the proposed approach is presented in Figure 3.1.
As a first step, 2D head direction is obtained as presented in Section 2.2.

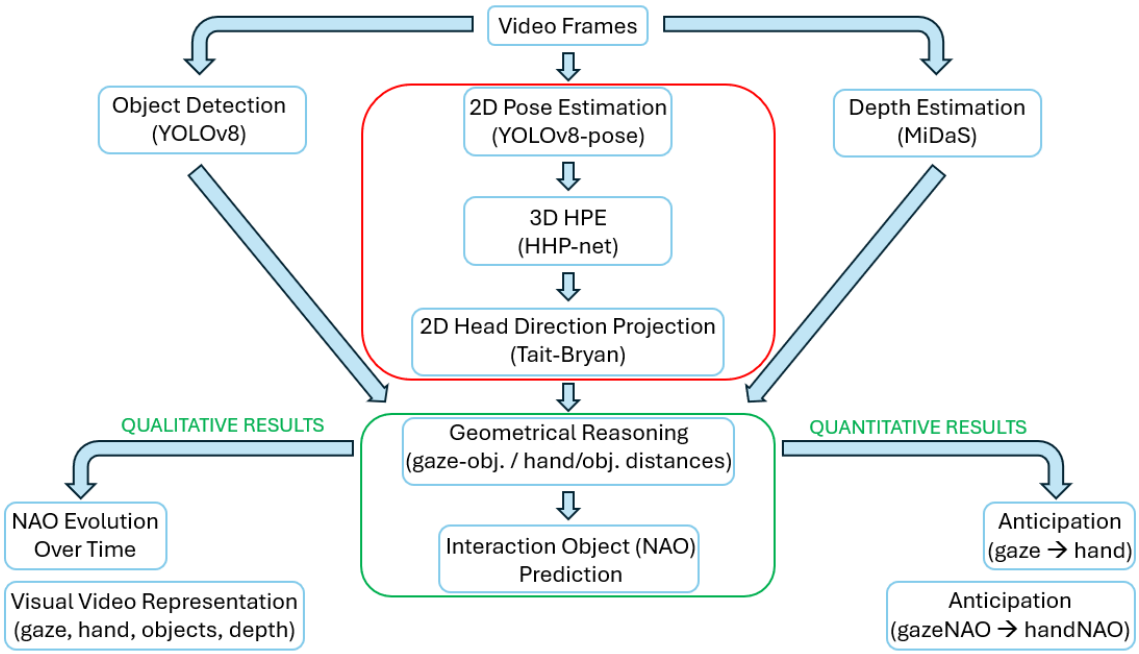


Figure 3.1: End-to-end pipeline for our Use Case 1. The steps highlighted in red are based on the literature ([7], [39], [40]), and are in common for both Use Case 1 and Use Case 2. The steps highlighted in green are the main novelty of our Use Case 1 method. Results we obtain are both qualitative and quantitative.

Object Detection and Depth Estimation

In parallel with head pose estimation, a generic object detector (in this work, YOLOv8 [34]) is applied to each frame to detect and classify all the objects appearing on the desk. For each detected object, the following pieces of information are retained: class label, object center and bounding boxes in image coordinates.

Moreover, to enrich the spatial representation of the scene, a relative monocular depth map is estimated for each frame using MiDaS (presented in Section 1.4). Unlike metric depth estimation, MiDaS predicts a quantity that is proportional to inverse depth.
Formally, the predicted depth $\hat{D}(x)$ at pixel x can be interpreted as $\hat{D}(x) \propto \frac{1}{Z(x)}$, where $Z(x)$ denotes the true scene depth. Therefore, predictions preserve relative depth ordering, but do not provide absolute distances in metric units.
To ensure numerical consistency within each frame, the predicted depth map is min-max

normalized:

$$D_{\text{norm}}(x) = \frac{\hat{D}(x) - \min_{x'} \hat{D}(x')}{\max_{x'} \hat{D}(x') - \min_{x'} \hat{D}(x')},$$

mapping the values to the interval $[0, 1]$. After normalization, pixels with values close to 1 correspond to regions near the camera, while those with values close to 0 correspond to far regions. Although this representation does not produce metric 3D reconstruction, it enables approximate spatial reasoning from 2D images and supports more reliable relative distance computations between subject and objects.

Geometrical Reasoning: 2D Distance Computation

After the first three steps of head direction projection, object detection and depth estimation, geometrical reasoning in the form of hand-to-object and gaze-to-object distance computation (and a weighted version of both) is applied.
In addition to the five facial keypoints necessary for the HPE, left and right wrist keypoints are extracted from the pose estimator (YOLOv8-Pose [28]) to approximate the position of both hands.
Hand-to-object distances are computed as the Euclidean distance (in pixels) between each object center and the position of each wrist. Similarly, gaze-to-object distances are computes as the angular distances (in degrees) between the head direction vector and the vector connecting the head center to each object's center.
A "fused" version of both distances is obtained, after normalization, through a weighted linear fusion. Formally, let $d_g(c)$ and $d_h(c)$ denote respectively the gaze-to-object and hand-to-object distances computed for the object of class c, for each $c \in \mathcal{C}$ (set of detected classes). Since gaze-to-object (angular) and hand-to-object (linear) distances lie on different numerical scales, they are independently normalized using max-normalization:

$$\tilde{d}_g(c) = \frac{d_g(c)}{\max_{c' \in \mathcal{C}} d_g(c')}, \qquad \tilde{d}_h(c) = \frac{d_h(c)}{\max_{c' \in \mathcal{C}} d_h(c')}.$$

With such normalization, both distances are mapped into the interval $[0, 1]$, ensuring scale comparability. Finally, we can define the "fused" distance score for class c as the weighted average of the normalized distances:

$$d_f(c) = \frac{w_g \, \tilde{d}_g(c) + w_h \, \tilde{d}_h(c)}{w_g + w_h},$$

where $w_g, w_h > 0$ are two weights. In our implementation, we set $w_g = 0.7$ and $w_h = 0.3$, assigning a greater importance (weight) to gaze-based than to hand-based distance. Our choice reflects the expected anticipatory role of head direction with respect to hand motion.

Geometrical Reasoning: Depth-Aware Distance Computation

In this phase of the pipeline, depth information is (optionally) exploited to enhance and correct the computation of hand-based and gaze-based distances.

To obtain depth-aware hand distances, we incorporate depth into the hand-to-object distance estimation. Formally, we define the depth-aware hand distance as:

$$d_{h,o} = \sqrt{(x_h - x_o)^2 + (y_h - y_o)^2 + \lambda(D_h - D_o)^2} = d_h + \sqrt{\lambda(D_h - D_o)^2},$$

where d_h is the hand-to-object distance (as previously defined), (x_h, y_h) and (x_o, y_o) are the image coordinates of hand and object centers, D_h and D_o are their corresponding normalized (in $[0, 1]$) monocular depth values (obtained from MiDaS), and $\lambda > 0$ is a scaling factor. This factor balances the contribution of the depth term with respect to the spatial (pixel) coordinates. This formulation corresponds to a scaled 3D Euclidean distance in the augmented space $(x, y, \sqrt{\lambda}\, D)$.
Image coordinates are measured in pixels, while depth values lie in $[0, 1]$. Therefore, the depth term would be negligible without a scaling factor λ. To ensure comparability between magnitudes, the scaling factor is defined as $\lambda = \frac{(W/2)^2+(H/2)^2}{(\Delta_D)^2}$, where $W = 1920$ and $H = 1080$ are respectively image width and height, and Δ_D represents the expected depth range (equal to 1 after the previously defined max-normalization). The result of this choice is that a large depth variation contributes approximately as much as a large spatial displacement in the image. Therefore, the added term $\lambda(D_h - D_o)^2$ in $d_{h,o}$ captures relative separation in depth, potentially improving robustness of hand-to-object distance estimation.

Similarly, to obtain depth-aware gaze distances, we incorporate depth into the gaze-to-object distance estimation. Formally, we define the depth-aware gaze distance as:

$$d_{g,o} = \alpha\, d_g + \beta\, |D_g - D_o|,$$

where d_g is the gaze-to-object distance (as previously defined), $|D_g - D_o|$ is the difference between the normalized (in $[0, 1]$) monocular depth values (obtained from MiDaS) at the head center (D_g) and at the object center (D_o), and $\alpha, \beta > 0$ are weighting coefficients. In our implementation, we set $\alpha = 0.7$ and $\beta = 0.3$, assigning greater importance to directional alignment while using depth to resolve possible geometrical ambiguities. Therefore, the added terms $\beta|D_g - D_o|$ and α in $d_{g,o}$ capture relative separation in depth, penalizing objects that are angularly aligned with head but located at inconsistent distances, potentially improving robustness of gaze-to-object distance estimation.

Interaction Object (NAO) Prediction

Finally, as the last step of the pipeline, a frame-wise interaction object (NAO) prediction is obtained as follows:

- The object with the smallest Euclidean distance (hand-to-object distance d_h^*) is selected as the "hand-based NAO" for that frame;
- The object with the smallest angular distance (gaze-to-object distance d_g^*) is selected as the "gaze-based NAO" for that frame;
- The object with the smallest weighted normalized distance ("fused" distance d_f^*) is selected as the hand-gaze weighted NAO ("fused NAO") for that frame.

For what concerns the hand-based NAO predictions, we only consider the distances with respect to the right wrist, since the subject of the videos is right-handed and only acts with that hand. Although we consider the possibility to use the minimum distance between left and right wrist, we decide to only use the distances from the most dynamic hand for the NAO predictions, since in the experimental videos the left hand always lies on the table, close to non-important objects.

Novelty of Our Approach

We end this section highlighting the main differences and additions with respect to [40], which we use as reference and starting point.
For what concerns the quantitative exploration of gaze-to-hand anticipation, our work mainly differs for the following methodological aspects:

- We use YOLOv8-Pose [28] as pose detector, instead of CenterNet [14] (mainly due to time consumption reasons);
- Since our work still represents an experimental testing of the method, we apply our pipeline to the videos obtained from a single subject, instead of exploiting all 16 subjects. Anyway, we are able to consider for our analysis one more action (eat crisps) that was discarded in [40];
- While in [40] the "gazed point" is obtained arbitrarily elongating the head vector (to approximately reach the table), in our work we only consider the head direction to obtain the gaze-based angular distances. When we need a proper comparison between gaze-based (in degrees) and hand-based (in pixels) distances, we apply a different reasoning to convert angular gaze distances into Euclidean distances, with pixels as a common metric. We define the "gazed point" in the image plane as the intersection of the subject's head direction with the interaction object's vertical level (the horizontal line passing through its center). The Euclidean distance between "gazed point" and object center is then computed analogously to the hand-object distance. Such formulation enables a proper comparison between gaze and hand

dynamics over time, allowing us to quantitatively assess the temporal anticipation of gaze with respect to hand movements;

- While in [40] the end target position for transporting movements is fixed and available as prior knowledge, in our work it is automatically retrieved as the position of the interaction object center once it stops being transported (when position differs "significantly" with respect to previous frame but is "the same" as the following). Similarly, the starting target position for transporting movements is automatically retrieved as the position of the interaction object center before it starts being transported (when position differs "significantly" with respect to following frame but is "the same" as the previous);

- Anticipation is estimated as the temporal difference between the two instants in which hand and "gazed point" are respectively the closest to the interaction object center (for reaching) or to the end target position (for transporting). In [40], such instants correspond to the first local minimum in an arbitrarily specified time window, calculated to consider only the initial (for reaching) or a subsequent (for transporting) part of the motion. In our work, we keep the logic of the first minimum in an arbitrarily specified time window only for reaching actions (without object motion). In case of transporting actions (with object motion), the windows are automatically identified as the frames before the object starts moving (from the starting target position) or ends its movement (in the end target position).

In addition to what obtained in [40], we present a qualitative analysis based on a frame-wise classification of the interaction object (NAO). Moreover, we also use the NAO frame-wise computation to add a different logic, based on temporal differences between gaze-based and hand-based NAO, to the quantitative gaze-to-hand anticipation analysis.
In Section 3.3, we present both the qualitative and the quantitative results we obtain applying our pipeline to this specific HOI scenario.

3.3 Results

The results we obtain with our method are both qualitative (Section 3.3.1) and quantitative (Section 3.3.2).

For what concerns qualitative results, we produce visual representations of each video frame, where all the HOI key aspects that we consider are highlighted: gaze direction, hand position, object detections (centers and bounding boxes), relative depth estimation, computed distances and frame-wise NAO predictions with respect to one specific distance metric (gaze-based, hand-based or "fused", with or without depth-awereness).

We also generate temporal NAO evolution plots throughout each video to visualize how the frame-wise identification of the interaction object (NAO) evolves over time.

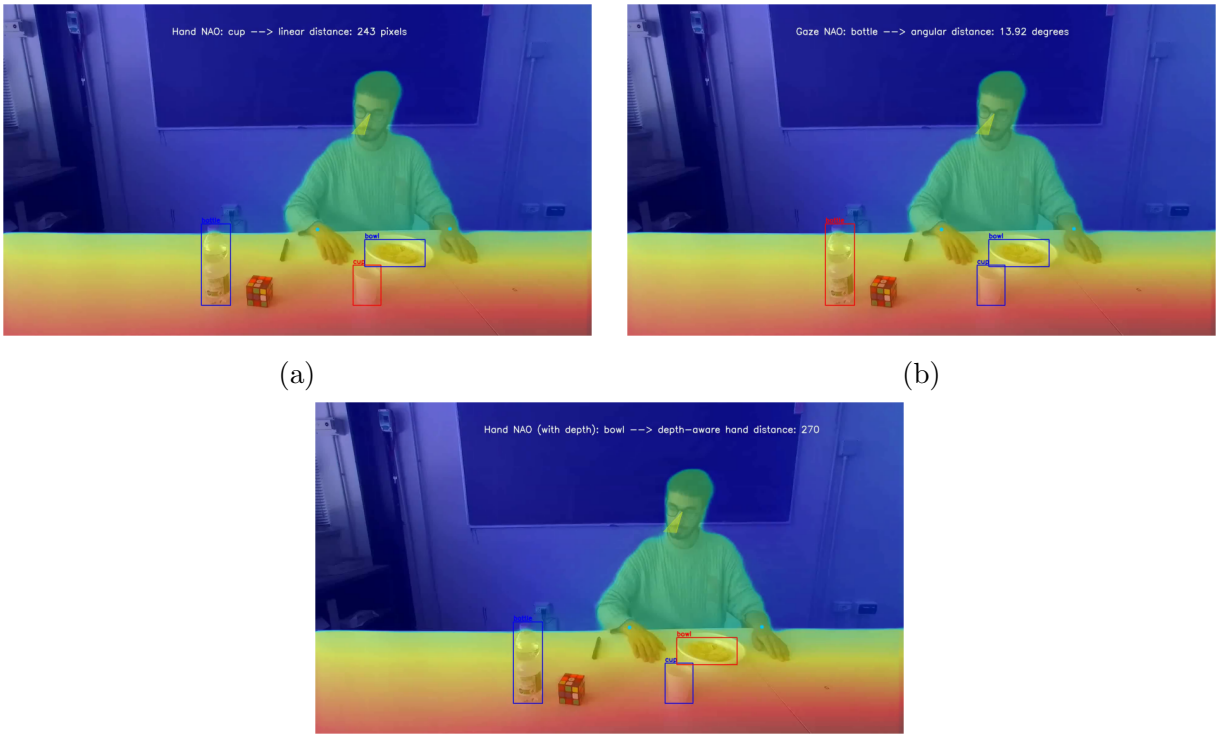
In addition to the NAO temporal evolution qualitative analysis, the temporal evolution over time of both gaze-based and hand-based distances is explicitly examined to quantify gaze-to-hand anticipatory behavior. Similarly to [40], we obtain gaze-to-hand anticipation times for both reaching and transporting actions, determining how early head direction (as a proxy for gaze) precedes hand motion toward the interaction object.
Moreover, we also obtain a slightly different anticipation measure for reaching actions, which is represented by the temporal anticipation between the instants when gaze-based NAO and hand-based NAO predictions corresponds to the correct interaction object.

Both qualitative and quantitative results allow us to highlight the importance of visual cues, and in particular HPE as a proxy for gaze, in HOI scenarios.

3.3.1 Qualitative Analysis

To qualitatively assess our method, we overlay useful visual representations (head direction, hands and objects positions, relative depth, ...) on top of original video frames. In Figure 3.2, we can observe the same frame with the following visual representations on top:

- Two light-blue dots, which identify the position of both hands (left and right wrist);
- A yellow cone (symmetrically built around the head direction), which identifies the subject's gaze direction. The visual cone is built exploiting the HPE uncertainty (obtained as output of the HHP-net), which is otherwise not considered in calculations: the higher the uncertainty, the wider the cone. It would also be possible to visualize the head direction as a vector instead of a visual cone;
- One bounding box for each object, which identifies its position in the image. The predicted frame-wise interaction object (NAO) is identified with a red bounding box, while the other objects are identified with a blue one. The bounding boxes corresponding to non-interesting object classes, such as "person" and "dining table", are removed from the visualization;
- A color heatmap of the relative depth, which distinguishes near and far objects. Close pixels appear more red, while far pixels appear more blue;
- One line of text, which highlights the predicted frame-wise NAO and its distance according to the selected metric. As we presented in Section 3.2, six different metrics can be selected to predict the frame-wise NAO: hand-to-object distance, gaze-to-object distance, "fused" (hand-gaze weighted) distance, depth-aware hand distance, depth-aware gaze distance, and depth-aware "fused" distance.



(a) (b)

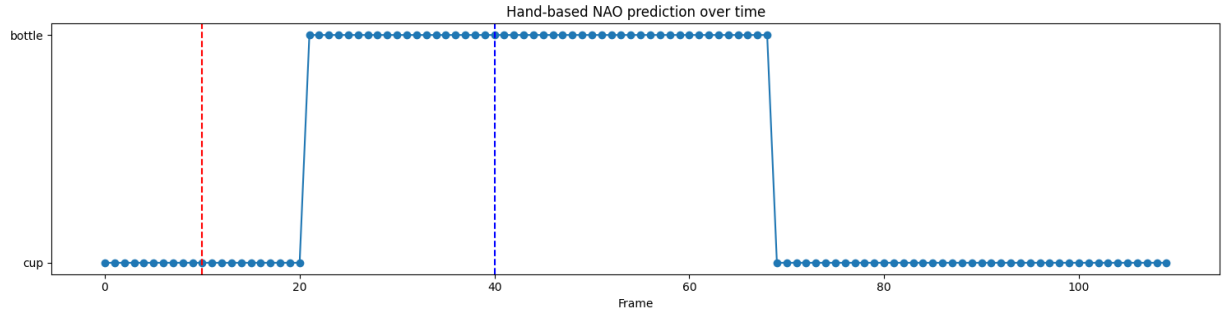
(c)

Figure 3.2: First frame from a transporting action video (with bottle as target object) with useful visual representations overlaid on top: gaze (visual cone), hands (wrists), objects (bounding boxes), depth (color heatmap), NAO prediction (red bounding box) with respect to a specific distance computation metric. It is presented a qualitative comparison between hand-based (a), gaze-based (b), and depth-aware hand-based (c) NAO estimations.

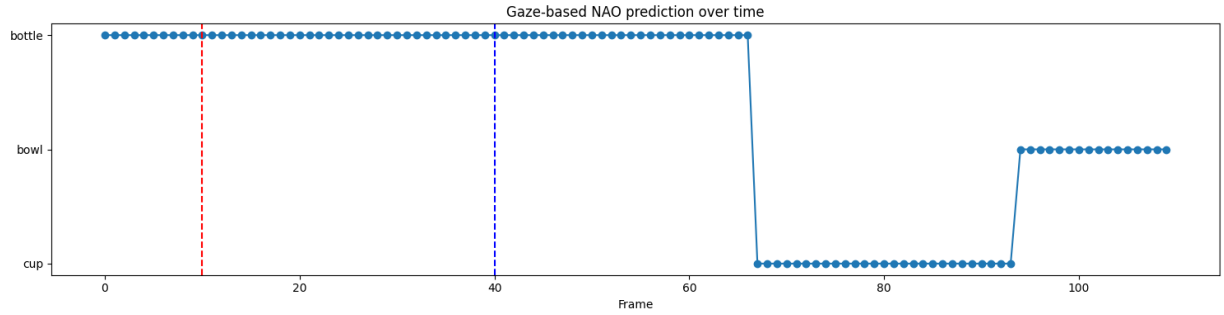
We can observe how the frame-wise NAO prediction is affected by the metric considered. In Figure 3.2, it is clear that the NAO prediction for the same frame can vary from "cup", considering hand-to-object distances (Figure 3.2a), to "bottle", considering gaze-to-object distances (Figure 3.2b), and "bowl", considering depth-aware hand distances (Figure 3.2c). The choice of the most meaningful distance metric to predict the correct interaction object (NAO) may depend on the specific action that is happening or going to happen: gaze-only distance can be more useful to anticipate the interaction with the object, while hand-only distance can better identify the interaction object when the action is occurring; weighted distance can represent a more flexible way to adapt to both situations, while depth-aware distances can correct the computation in cases where relevant depth disparities (between head, hands, and objects) are present.

Although depth-aware distances can sometimes lead to different NAO predictions with

respect to 2D distances (as in the case of Figures 3.2a and 3.2c), we observe that generally depth does not provide any significant additional information in most scenarios. For example, in those cases where the position of the object or its depth is not-at-all or barely modified, like in reaching actions or in trasporting actions that occur on a horizontal plane, 2D distance computation (without depth) can be sufficient. We believe that in our experimental setting where all objects are located on the table at a similar depth, taking into account depth (as optionally proposed in our pipeline) is not as relevant in analyzing HOI as it would be in more unconstrained environments.



(a)



(b)

Figure 3.3: Evolution of hand-based (a) and gaze-based (b) NAO prediction over time. Red and blue vertical lines indicate the frames where gaze and hand distances from bottle are respectively the smallest.

Another type of visualization we present to assess the quality of the HOI understanding of our pipeline is shown in Figure 3.3. The two plots show the evolution over time of both hand-based (Figure 3.3a) and gaze-based (Figure 3.3b) NAO predictions.
In particular, the visualization refers to a reaching action (touch bottle) where the target object is the bottle. The red and blue vertical lines respectively represent the two instants where the gaze (at frame 10) and hand (at frame 40) distances from the center of the bottle (the groundtruth interaction object) are the smallest. In Figure 3.3a, the bowl is not listed

because it is never predicted as the hand-based NAO.

Although we use this type of visualization as a qualitative insight, some quantitative considerations can be drawn.
For example, it is possible to estimate the fraction of frames for which the groundtruth interaction object is correctly predicted as the NAO. In addition, for each video the video-wise NAO prediction, namely the object predicted as frame-wise NAO for the largest fraction of frames, could be compared to the groundtruth.
Such comparative classification analysis is not presented since we find more meaningful (for both HOI understanding and anticipation estimation) to explore the interaction at a local (frame-wise, qualitative) instead of global (video-wise, quantitative) level. Moreover, despite the fact that in each video only one action is performed, the interaction with the object does not necessarily occupy the whole length of the video, especially in cases of simple reaching actions like touching the bottle.

Another example of a quantitative consideration that can be drawn from such visualization is presented in Section 3.3.2, and it concerns interaction anticipation: instead of considering "gazing time" and "touching time" to calculate anticipation, we can consider the first instants in time when the interaction object is correctly recognized as gaze-based and hand-based NAO among all other possibilities. For example, in Figure 3.3 we can observe that the "gazing time" (frame 10) and the "touching time" (frame 40) could be in some sense anticipated (to frame 0 and 21 respectively).
We present this possibility more in detail in Section 3.3.2.

One last qualitative consideration we would like to present concerns one of the problems we face during the analysis: bad-quality outputs from both pose estimator and object detector. Incorrect pose estimation or missing object detections impact to some extent the results of our method. Indeed, more accurate object and pose detections correspond to more reliable and useful considerations on gaze-to-hand anticipation and NAO predictions.
Therefore, when possible, we attempt to partially solve or work around them, but we consider this out of the scope of our work.

3.3.2 Quantitative Analysis

In addition to the qualitative exploration of NAO predictions over time, we also observe the quantitative anticipation of head direction with respect to hand movement, both in reaching and transporting actions.

In Figure 3.4, we can observe gaze and hand distances from each object (bottle, bowl, and cup) for a reaching video ("touch bottle"). We identify a meaningful anticipation pattern between gaze and hand only for the bottle, which is the groundtruth interaction object.

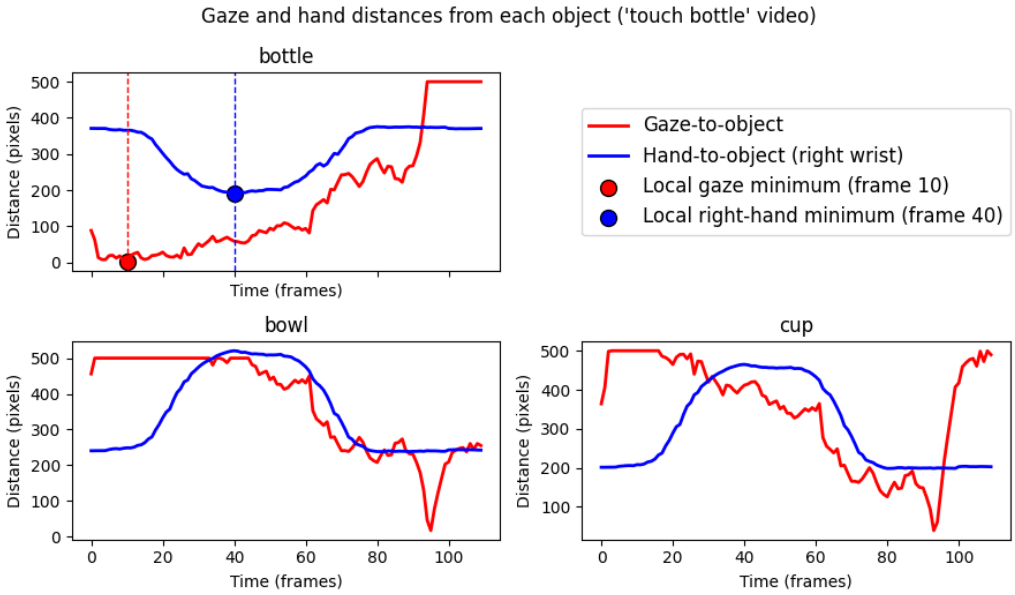


Figure 3.4: Gaze and hand distances from each object in "touch bottle" (reaching) video. The bottle (interaction object) is the only object for which we identify an anticipation pattern between gaze and hand. We set a cap of 500 pixels for gaze-to-object distances.

Differently from [40], where distances are computed with respect to a fixed and known (a priori) target position both for reaching and transporting actions, we compute distances with respect to each object center. In reaching actions where objects are not moved, there are no differences between the two approaches. In transporting actions, it is possible to observe that the hand distance from the manipulated object is low during the interaction. On the contrary, it is possible to observe high gaze distance from the interaction object when the head direction points toward the target position. Such distance gets lower and lower as soon as the object is transported to the final position.
We believe that exploring the relationships between gaze, hand, and object center can be more interesting than considering only a fixed target position known a priori.
For what concerns gaze-to-hand anticipation, it can be derived from Figure 3.4 as the temporal difference between local gaze and hand minimum.

Similarly to Figure 3.4, in Figure 3.5 we can observe gaze and hand distances. In this case, the video considered consists in a reaching and transport back action ("open and close bottle"). Only distances with respect to the bottle are shown (Figure 3.5a).
The plot is enriched with additional pieces of information: first and last local gaze and hand minimum; gaze minimum preceding reach; hand minimum at reach time; gaze minimum preceding transport back; hand minimum at transport back time.

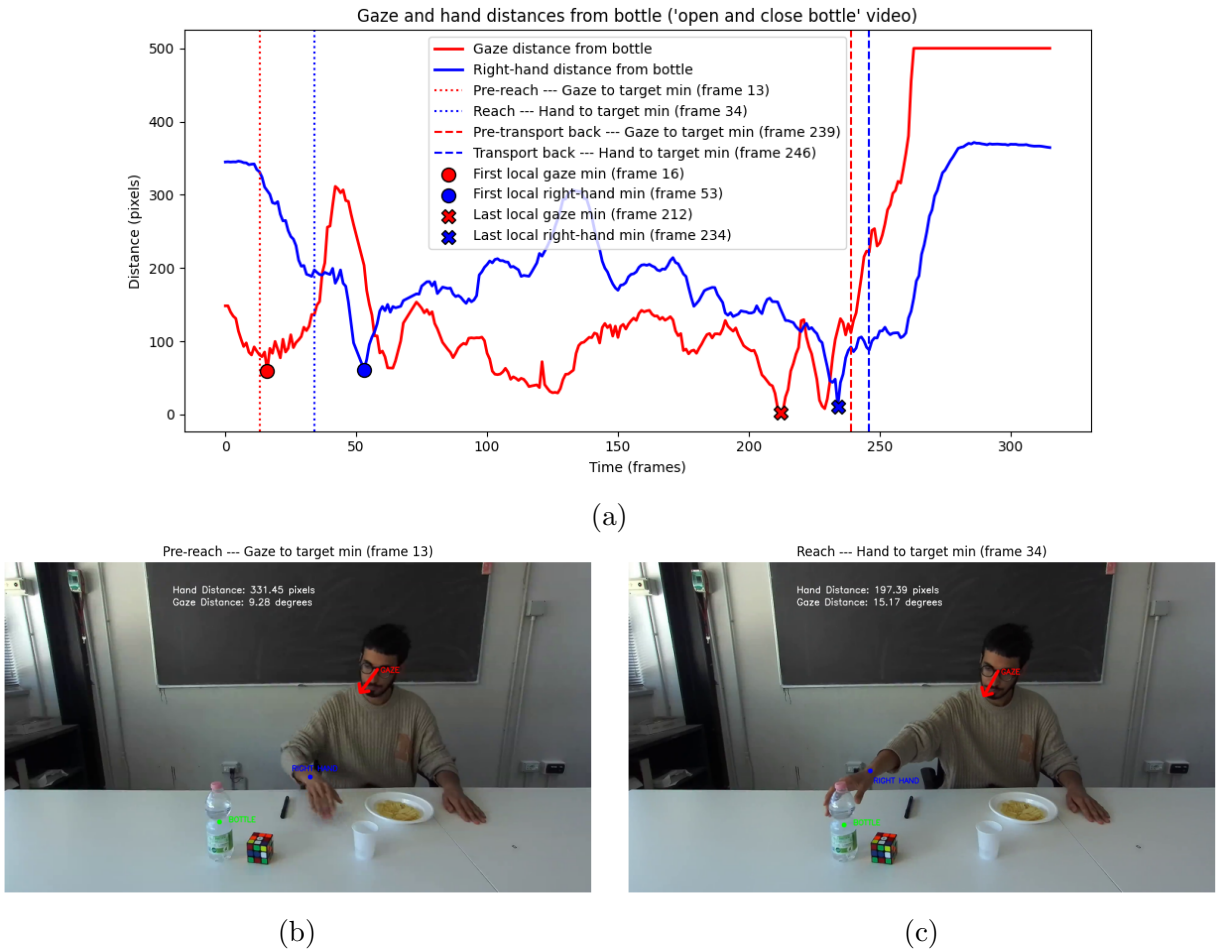


Figure 3.5: Gaze and hand distances (a) from bottle in "open and close bottle" (reaching and transport back) video, with visualization of pre-reach gaze minimum (b) and reach hand minimum (c) frames.

Since distances are now computed with respect to the object center (and not relative to a fixed known position), in case of transporting videos, finding the first or last local minimum is not sufficient anymore to correctly identify the instants in which interaction happens. For this reason, as already explained in detail in Section 3.2, we find the instants of gaze and hand interaction as those frames where a minimum occurs before the object moves (for reaching) or stops moving (for transporting).

In Figure 3.5b we show the frame where gaze distance is minimum with respect to the bottle, before reaching happens, while in Figure 3.5c we show the frame where hand distance is minimum with respect to the bottle, as soon as reaching happens. Similar visualizations

Table 3.1: Anticipation between gaze and hand for reaching actions.

Trial	**Gaze (fr.)**	**Hand (fr.)**	**Anticipation (fr. = s)**
DrinkCup (trial 1)	4	25	$-21 = -0.70s$
DrinkCup (trial 2)	1	24	$-23 = -0.77s$
EatBowl (trial 1)	14	36	$-22 = -0.73s$
EatBowl (trial 2)	0	17	$-17 = -0.57s$
OpenCloseBottle (trial 1)	16	33	$-17 = -0.57s$
OpenCloseBottle (trial 2)	13	34	$-21 = -0.70s$
TouchBottle (trial 1)	10	40	$-30 = -1.00s$
TouchBottle (trial 2)	6	21	$-15 = -0.50s$
TransportBottle (trial 1)	38	51	$-13 = -0.43s$
TransportBottle (trial 2)	14	32	$-18 = -0.60s$

Table 3.2: Anticipation between gaze and hand for transporting actions.

Trial	**Gaze (fr.)**	**Hand (fr.)**	**Anticipation (fr. = s)**
DrinkCup (trial 1)	112	124	$-12 = -0.40s$
DrinkCup (trial 2)	117	135	$-18 = -0.60s$
OpenCloseBottle (trial 1)	248	254	$-6 = -0.20s$
OpenCloseBottle (trial 2)	239	246	$-7 = -0.23s$
TransportBottle (trial 1)	86	96	$-10 = -0.33s$
TransportBottle (trial 2)	95	92	$+3 = +0.10s$

may be obtained for the two frames where respectively gaze distance is minimum prior to the transport back and hand distance is minimum as soon as transporting back happens.

Table 3.1 shows all the computed anticipations for reaching actions, while Table 3.2 lists those for transporting actions.
Moreover, as a different approach to estimate how early the interaction object is correctly identified first by gaze and then by hand, we present in Table 3.3 the temporal anticipations derived from gaze-based and hand-based frame-wise NAO predictions.

Finally, Table 3.4 resumes the results with summary statistics such as mean and standard deviation. Although we consider a very small set of videos (10 reaching actions, 6 of them also involving a transporting action), we believe that such statistics can still be indicative of the following trend: in simple reaching and transporting actions, gaze consistently anticipates hand movement.

From the tables, we can observe that in almost every case anticipation values are negative. This indicates that gaze consistently orients toward the target position (both for reaching and transporting) before hand does.

Table 3.3: Anticipation between gaze-based and hand-based NAO for reaching actions.

Trial	**Gaze (fr.)**	**Hand (fr.)**	**Anticipation (fr. = s)**
DrinkCup (trial 1)	28	0	$+28 = +0.93s$
DrinkCup (trial 2)	22	0	$+22 = +0.73s$
EatBowl (trial 1)	0	12	$-12 = -0.40s$
EatBowl (trial 2)	0	6	$-6 = -0.20s$
OpenCloseBottle (trial 1)	0	18	$-17 = -0.60s$
OpenCloseBottle (trial 2)	0	15	$-21 = -0.50s$
TouchBottle (trial 1)	0	21	$-21 = -0.70s$
TouchBottle (trial 2)	0	4	$-4 = -0.13s$
TransportBottle (trial 1)	0	32	$-32 = -1.07s$
TransportBottle (trial 2)	0	16	$-16 = -0.53s$

Table 3.4: Summary statistics (mean $\pm$ standard deviation) of anticipation across different interaction settings. (*) Outliers, corresponding to cases where hand seems to precede gaze due to intrinsic limitations of the method, are excluded from the computation.

Setting	**Anticipation (fr.)**	**Anticipation (s.)**
Reaching (gaze-hand) - 10 values	19.7 ± 4.6	$-0.66s \pm 0.15s$
Transport (gaze-hand) - 5 values (*)	10.6 ± 4.3	$-0.35s \pm 0.14s$
Reaching (NAO-based) - 8 values (*)	16.1 ± 8.4	$-0.54s \pm 0.28s$

The positive value in 3.2 is due to object detection problems: for many frames, the bottle is "lost" while transported, so the distance from the object center cannot be computed.
On the other hand, the positive values in 3.3 are due to the intrinsic limitations of using 2D head direction projection as a gaze proxy: in such cases, the cup is almost at the same horizontal level as the bowl (vertically aligned with it). While the hand is closer to the cup, the gaze distance with respect to the bowl is slightly lower. Therefore, the cup is predicted as the hand-based NAO before than being predicted as the gaze-based NAO. Here, considering depth-aware gaze distances (which penalizes objects far from the subject) would not solve the misclassification either, since the bowl is closer to the subject.

Despite all the limitations (few videos, 2D head direction as a proxy for gaze) and the problems (missing detections) of our experiment, we believe that our proposed approach can be useful to investigate interaction and anticipation in HOI settings.
Although in trivial cases 2D head direction alone could be sufficient to detect and anticipate the interaction object (among others) in a scene, exploring a combination of tools (as we propose in our pipeline), such as depth estimation or hand proximity, may be useful to have a more complete scene understanding.

Chapter 4

Use Case 2: Human-Human Interaction

In Chapter 3, we analyzed the role of HPE in a controlled scenario where a single subject interacts with different objects. We framed the experiment as a Human–Object Interaction (HOI) use case.
In this chapter, we present a different application scenario where two subjects interact with each other in a controlled environment. We present the experiment as an example of Human-Human Interaction (HHI) use case.

Specifically, we consider a mother–child interaction scenario which has been presented to us during a research collaboration with the Italian Institute of Technology (IIT). The method and the results of the collaboration are presented in [20], a work-in-progress paper in which we are co-authors. Such study introduces a multimodal vector-based framework to exploit head orientation during mother-child engagement for the analysis of dyadic interactions.

During social interactions, traditional measures to quantify attentional coordination between two subjects have primarily focused on gaze behaviors. In this scenario, the head orientations of both subjects are exploited as indexes of attentional coordination. The main goal is to infer and analyze patterns of interaction between mother and child. In particular, the approach we propose relies exclusively on 2D computer vision (CV) techniques applied to monocular videos. Our results are compared with both clinical video coding and 3D motion capture (MoCap) measurements, potentially representing more expensive and less accessible alternatives in real-world contexts. It has been hypothesized that there would be consistency across the three tracking techniques: clinical video coding, 3D motion capture (Vicon MoCap system) and our 2D computer vision approach.

4.1 Setting

The experimental setting consists in the same structured mother-child interaction scenario described in [19]. Ten mother-child dyads engages in a naturalistic play inside a controlled environment, while head orientation is concurrently estimated using clinical video coding, motion capture and computer vision.

More in detail, each couple is recorded over 5 minutes of ecological interaction (subdivided into 5 trials), in which the mother (caregiver) remains seated in a corner, while the child is free to move and interact with the toys placed in the scene. The resulting dataset consists in a total of 50 videos of approximately one minute of length each.

Children are aged between 2 and 4 years, and are evenly divided into two groups: 5 children with visual impairments, and 5 typically developing children. From a clinical perspective, one of the aims is analyzing potential differences in interaction patterns between these two groups.

Interaction takes place in a close ecological playground (3×3 meters), surrounded by a baby-friendly gate (1 meter high). The mother sits in a frontal position with respect to a single RGB camera placed at 1.50 meters of distance from the gate enclosing the environment. The recordings from such camera consist in the only inputs for our CV-based analysis.

In parallel, a Vicon MoCap system is employed by IIT researchers. Both mother and child wear a headband equipped with three reflective markers positioned on its left, center, and right frontal parts. Marker trajectories are captured by eight near-infrared cameras mounted on the ceiling: four located at the corners of the room and four placed at the center of each side.
In addition to both CV and MoCap predictions, a clinician coded the videos annotating interaction behaviors every single second.

Therefore, for each trial, three parallel pieces of information are obtained:

- Clinical interaction annotations, manually provided by an expert;
- 3D Vicon MoCap-based interaction estimates;
- 2D CV-based interaction predictions, obtained using our HPE pipeline.

4.2 Method

In this section, we describe the CV-based approach we adopt to infer head orientation from 2D video frames, calculate the interaction angle between the two subjects and classify such interaction. In addition, we briefly present the clinical video coding and the 3D MoCap approaches adopted by IIT researchers during the collaboration. Finally, we define the metrics we use to quantify the level of agreement between our approach and each one of the other classification models.

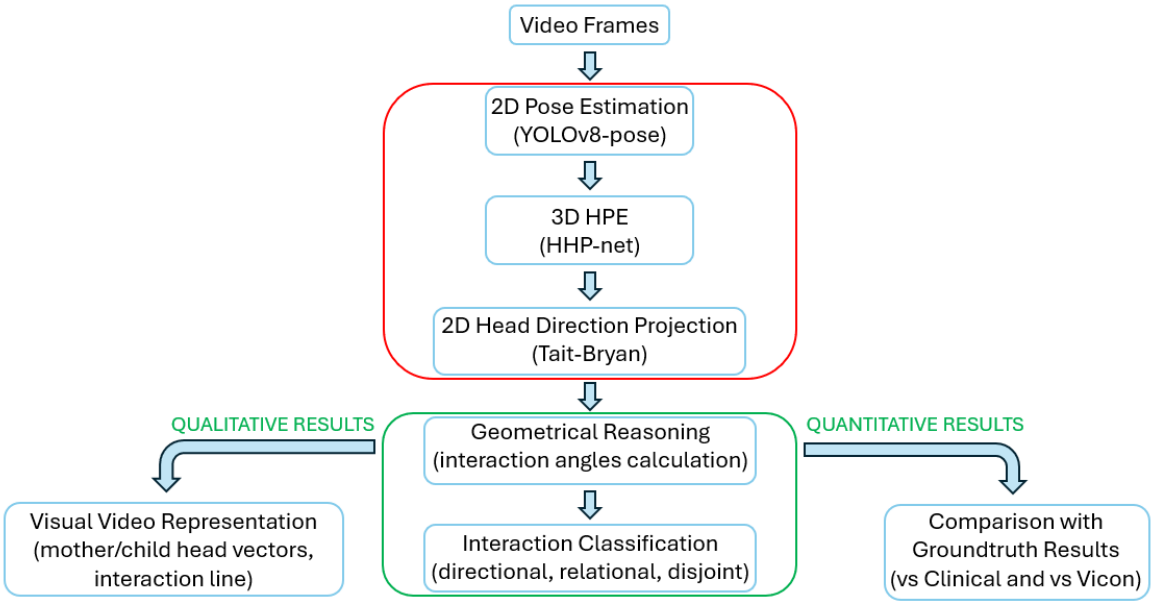


Figure 4.1: End-to-end pipeline for our Use Case 2. The steps highlighted in red are based on the literature ([7], [39], [40]), and are in common for both Use Case 1 and Use Case 2. The steps highlighted in green are the main novelty of our Use Case 2 method. Results we obtain are both qualitative and quantitative.

The full pipeline of the proposed approach is presented in Figure 4.1.
As a first step, 2D head direction is obtained as presented in Section 2.2.

Geometrical Reasoning: Interaction Angles Calculation

Once the head centers and the 2D head direction unit vectors are obtained for both subjects, geometrical reasoning is applied to infer the occurring type of social interaction between mother and child (steps highlighted in green in Figure 4.1).

Specifically, both head centers $\mathbf{p}_M = (x_M, y_M)$ and $\mathbf{p}_C = (x_C, y_C)$ as well as the 2D head direction unit vectors $\mathbf{h}_M$, $\mathbf{h}_C \in \mathbb{R}^2$, associated respectively with the mother and the child, are represented in 2D image coordinates.
We calculate the unsigned angular difference between the two head directions as:

$$\theta = atan2(\|\mathbf{h}_M \times \mathbf{h}_C\|, \mathbf{h}_M \cdot \mathbf{h}_C),$$

with $\theta \in [0, \pi]$. For two generic vectors $\mathbf{a} = (a_x, a_y)$ and $\mathbf{b} = (b_x, b_y)$, $\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y$ and $\|\mathbf{a} \times \mathbf{b}\| = |a_x b_y - a_y b_x|$. Geometrically, $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos\theta$ and $\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\|\|\mathbf{b}\| |\sin\theta|$. Since we are operating with unit vectors, $\|\mathbf{a}\|\|\mathbf{b}\| = 1$. Thus, $atan2(|\sin\theta|, \cos\theta)$ recovers the unsigned angle $\theta \in [0, \pi]$ between the two vectors, discarding directional information (if looking more to the left or to the right).
In addition, we compute the interaction angle of the mother toward the child as:

$$\theta_M = atan2(\|\mathbf{h}_M \times \mathbf{v}_{MC}\|, \mathbf{h}_M \cdot \mathbf{v}_{MC}),$$

where $\mathbf{v}_{MC} = \mathbf{p}_C - \mathbf{p}_M$ is the vector connecting the mother's head center to the child's head center. This angle quantifies the unsigned angular deviation between the mother's head orientation and the (mother-to-child) connecting vector, independently of whether the child lies to the left or right.
Analogously, we compute the interaction angle of the child toward the mother as:

$$\theta_C = atan2(\|\mathbf{h}_C \times \mathbf{v}_{CM}\|, \mathbf{h}_C \cdot \mathbf{v}_{CM}),$$

where $\mathbf{v}_{CM} = \mathbf{p}_M - \mathbf{p}_C = -\mathbf{v}_{MC}$ is the vector connecting the child's head center to the mother's head center. This angle quantifies the unsigned angular deviation between the child's head orientation and the (child-to-mother) connecting vector.

Geometrical Reasoning: Interaction Classification

The 2D geometrical reasoning we use to compute the interaction angles between mothers and children is conceptually equivalent to the 3D geometrical reasoning applied by our IIT collaborators in the Vicon MoCap approach (explained in detail in [20]).
Specifically, 3D motion data is tracked with the Vicon System after performing spatial calibration. Then, a 3D model is created using the head as a single kinematic segment where the three markers (left, central, and right foreheads) are identified. The x, y and z coordinates of the markers are extracted and filtered using a low-pass filter to remove noise. For what concerns the head orientation estimation, such approach relies on the computation of 3D local reference systems (associated with yaw, pitch, and roll measures) for both mothers and children' heads. After this computation, the only axis considered is the one representing the yaw rotational angle of the head, which refers to the rotation on the horizontal anatomical plane, and it is most directly associated with facial orientation.

For both approaches, the dyadic interactions are classified into three attentional categories: Relational, Directional and Disjoint. To operate the classification, a cutoff angle selection is assigned to each attentional category. This logic is applied to the interaction angles between mothers and children computed both with 2D CV and 3D Vicon MoCap approaches. Specifically, the three interaction classes (schematized in Figure 4.2) are the following:

- Relational attention describes a situation where mother and child are looking at each other. This interaction (coded as -2) corresponds to angles between 150° and 180°. From a computational point of view, in our 2D CV approach it occurs when both angles θ_M and θ_C between each subject's head direction vector and the connecting vector are below the fixed threshold of 30°;
- Directional attention describes a situation where mother and child are looking approximately in the same direction. This interaction (coded as 0) corresponds to angles between 0° and 30°. From a computational point of view, in our 2D CV approach it occurs when the angle θ between the two head direction vectors is below the fixed threshold of 30°;
- Disjoint attention describes a situation where mother and child are looking in different directions, but not toward each other. This interaction (coded as 1) corresponds to angles between 30° and 150°. From a computational point of view, in our 2D CV approach it occurs in every other case, when neither of the above conditions is satisfied.

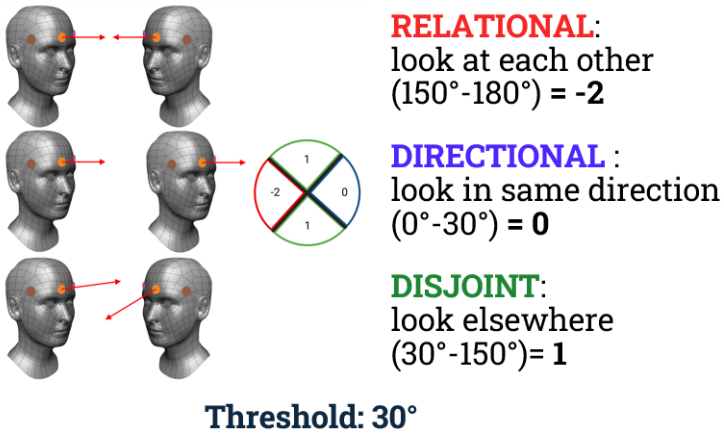


Figure 4.2: Graphical representation of the three considered attentional categories: Relational (in red), Directional (in blue), and Disjoint (in green). The angular threshold (which identifies the categories cutoffs) used in our computations is set to 30°.

In agreement with our IIT collaborators, these fixed (predefined) cutoff ranges have been selected after a preliminary check ensuring that such thresholds corresponded to head orientations consistent with the intended attentional categories. Although this categorical approach provides clear interpretability, we are aware that it relies on rigid thresholds that may not fully capture the continuous and graded nature of attentional dynamics.

WLS and GMM Complementary Models

To address these limitations, two complementary models have been developed by our IIT collaborators: a Weighted Least Squares (WLS) model and a Gaussian Mixture Model (GMM), explained in detail in [20]. Here, we just briefly introduce them.

WLS model combines angular relationships with regression-based weighting. Interaction angles are modeled as graded expressions of three prototypical configurations (0° for Directional, 90° for Disjoint, 180° for Relational) using Gaussian weighting profiles. Samples closer to these canonical angles are given greater influence in a weighted linear regression that outputs smooth continuous predictions. Final attentional states are obtained by discretizing these predictions into the previously discussed three categories.
This strategy allows smooth transitions between states by combining the interpretability of categorical coding with the flexibility of continuous modeling.

GMM model uses probabilistic clustering to take into account the variability in head orientation strategies. The distribution of interaction angles is modeled in a data-driven way using a three-component GMM. After parameter estimation, the resulting components are ordered by ascending mean angle and respectively mapped to Directional, Disjoint, and Relational attentional states. Each observation is assigned to the component with the highest posterior probability.
This strategy allows attentional categories to emerge directly from the empirical distribution without arbitrary angular cutoffs.

For what concerns clinical video coding, which does not involve vectors construction and angles calculation, the cutoff logic cannot be applied. In this case, a clinician is asked to code each video, annotating the observed behaviors every single second based on observational criteria. The instructions are to observe each frame and classify the head orientation between mother and child as one of the same three categories presented above: Relational, if heads are approximately facing each other; Directional, if heads are approximately pointing in the same direction; Disjoint, if heads are pointing in different directions.

WLS and GMM computational models are applied to both 2D CV and 3D Vicon MoCap estimated angles to obtain a different classification, independent from the selection of a rigid cutoff. Since clinical video coding relies on discrete categorical annotations rather than continuous head orientation angles, the two models cannot be applied in this case.

Standardization and Comparison

After the initial frame-wise classification, temporal resolution is standardized across methods operating at different frame rates to ensure comparability. Specifically, clinical annotations are provided at 1 Hz (one frame per second), while CV and MoCap predictions are provided at 30 Hz and 100 Hz respectively. Therefore, CV and MoCap interaction labels are respectively aggregated into bins of 30 and 100 frames (over 30-frame and 100-frame temporal windows), matching the one-second resolution of clinical video coding. For each bin, it is computed the most frequent attentional category (mode).

The classification agreement between methods can then be evaluated. We perform two different comparisons: between CV and clinical video coding, and between the two automated approaches (CV and Vicon MoCap).
In Section 4.3, in addition to a qualitative assessment of our CV head direction estimates (Section 4.3.1), we present the results of the comparisons.

To quantitatively assess the agreement between methods (in Section 4.3.2), we first present the confusion matrices deriving from the second-by-second comparisons between the CV predicted interaction label (using cutoffs) and both Clinical and Vicon MoCap values, used once at a time as groundtruth. Similarly, we present the confusion matrices of the second-by-second comparison between CV and Vicon MoCap labels obtained after the application of WLS and GMM models.

Finally, we also compute the agreement at event-level using an event-based temporal matching index, with a behavior-specific tolerance window. Such window accounts for natural variability in behavioral timing and allows a more flexible matching between annotations that are not perfectly synchronized in time but conceptually aligned ([5]).

Agreement Metrics

To evaluate agreement between methods, we quantify and compare their performances in two different but complementary ways: we compute both second-level (second-by-second) and event-level agreement.

At a second-by-second level, each time point is treated as an independent categorical observation belonging to one of the three interaction behaviors (Relational, Directional or Disjoint). For each comparison, we compute the corresponding 3×3 confusion matrix.
To quantify the overall agreement (while correcting for chance) we compute the Cohen's Kappa (κ) coefficient:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement (sum of diagonal entries divided by the total number

of observations N), and p_e is the expected agreement by chance:

$$p_e = \sum_{i=1}^{3} \left(\frac{\text{row}_i \cdot \text{col}_i}{N^2} \right),$$

with row_i and col_i representing the marginal totals for class i.
Cohen's Kappa ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating chance-level agreement. In our study, κ provides a global summary of second-level agreement across the three classes.
In addition to κ, we compute for each class c per-behavior Precision, Recall, and F1-score derived from the second-level confusion matrices:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \quad \text{F1}_c = \frac{2TP_c}{2TP_c + FP_c + FN_c},$$

where TP_c, FP_c and FN_c are respectively the number of per-behavior true positives, false positives and false negatives.
We compute as well their overall values, averaged across behaviors assigning equal weight to each class, independently of its frequency.

While second-by-second level metrics evaluate instantaneous agreement, they penalize temporal misalignments. To assess agreement at a more behaviorally meaningful scale, we compute it at an event-level.
Second-wise labels are first converted into contiguous event intervals by grouping consecutive time points with the same labels. At each event is therefore assigned a duration.
To account for natural variability in behavioral timing, we introduce an adaptive and behavior-specific tolerance window. For each class c, we compute the standard deviation σ_c of event durations in those we consider as the groundtruth annotations. Such value defines a tolerance window of $\pm\sigma_c$ around the starting time of each groundtruth event.
A predicted event is considered a valid match with the groundtruth if it belongs to the same behavior class, the difference between the starting times is within $\pm\sigma_c$, and the two events temporally overlap (intersection > 0).
Matching is performed using temporal Intersection-over-Union (IoU), ensuring one-to-one correspondences between events.
For each class c, we compute per-behavior Precision, Recall, and F1-score, as well as the Event-level Temporal Matching Index (ETMI), defined as:

$$\text{ETMI}_c = \frac{TP_c}{TP_c + FP_c + FN_c}.$$

Differently from the F1-score, the ETMI represents a symmetric measure of temporal alignment. It directly quantifies the proportion of correctly matched events out of the total number of both groundtruth and predicted events.
Once again, overall values are obtained as the average of each per-behavior metric.

Second-level agreement captures instantaneous consistency between predictions of different methods. In contrast, event-level agreement evaluates whether the methods identify the same behavioral episodes, allowing small timing deviations that could naturally arise in human and automated classification processes.
Together, these complementary metrics should provide a comprehensive assessment of both lower-level classification agreement and higher-level behavioral episode alignment.

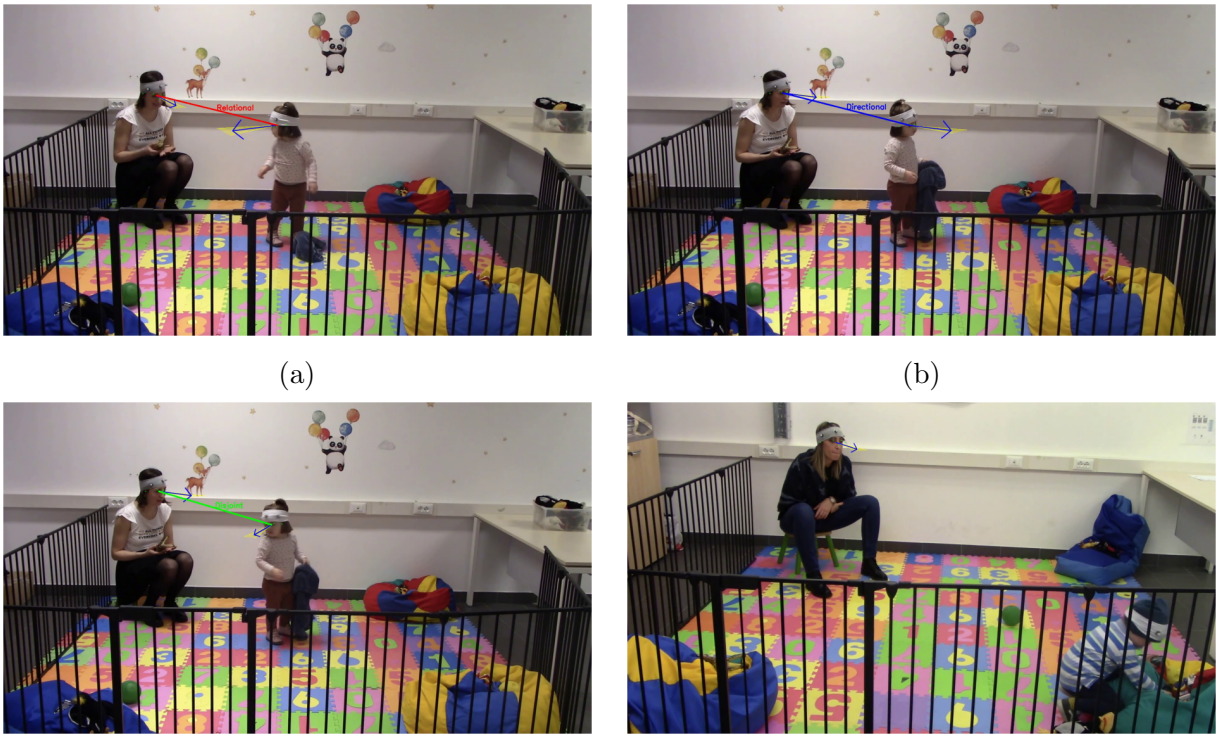
4.3 Results

The experimental results of our CV approach are analyzed from both qualitative and quantitative perspectives.
From a qualitative point of view, we overlay onto the video frames the five facial keypoints, head vectors and the (mother-child) connecting line. The goal is to visually assess the quality of the frame-by-frame head direction estimates, and the predicted interaction classification (based on the rigid cutoffs).
From a quantitative point of view, we compare our interaction predictions with Clinical and Vicon MoCap classifications. We present the comparison as a formal assessment of the agreement between methods both at second-level and at event-level.

4.3.1 Qualitative Analysis

From a qualitative point of view, predicted head direction vectors (in blue) are visually overlaid onto the video frames. Additionally, a visual cone (in yellow) is symmetrically built around the head direction exploiting the HPE uncertainty (obtained as output of the HHP-net): the higher the uncertainty, the wider the cone. Moreover, the line connecting mother and child head centers is also drawn and colored depending on the frame-wise occurring type of interaction: red for Relational, blue for Directional and green for Disjoint.
This visualization should allow an intuitive evaluation of the HPE, highlighting scenarios in which predictions are either more or less reliable.

In Figure 4.3, we can observe four frames, derived from two different dyads, where different types of interactions are occurring.
In Figure 4.3a, mother and child are looking at each other, while in Figures 4.3b and 4.3c they are respectively looking in the same direction and in completely different directions.
In Figure 4.3d, we present a quite frequent case of missing detection of one of the subjects (here the child). This is due to the fact that, in this particular frame, the pose detector is not able to recognize the child as a person, because he is crouching down and "hiding" himself behind the gate. Therefore, his head pose cannot be retrieved and no interaction can be computed.



(a) (b)

(c) (d)

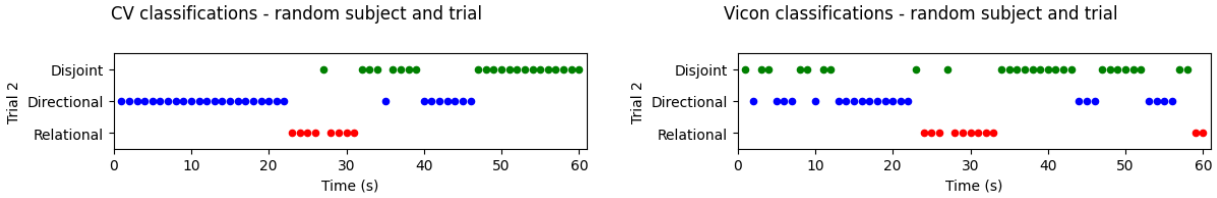
Figure 4.3: Frames where different types of dyadic interaction are observed: Relational (a), Directional (b) and Disjoint (c). In (d) it is shown a case of missing detection: the child is not detected by the pose detector; therefore, no interaction can be computed.

The visual exploration of video frames highlights some of the limitations we face during the experiment:

- Environmental occlusion, caused by elements in the scenario, such as the gate, which partially occlude the visibility of one of the subjects from the camera perspective;
- Human occlusion, caused by the mutual occlusion between subjects, when one of them (the child) gets in front of the other (the mother) from the camera perspective;
- 2D projection limitations, implicit in monocular vision approaches, which cannot fully capture depth cues and therefore differ from 3D-based systems such as MoCap.

Both environmental and human occlusions can lead to missing detections or negatively affect facial keypoints detection and head pose estimation. In other cases, the 2D nature of head direction vectors can lead to a misleading computation of angles and therefore an

incorrect interaction classification.
On the contrary, when both subjects (and in particular their faces) are entirely visible and not occluded, the visualizations highlight the effectiveness of the method. In such cases, angles calculated from head vectors and interaction classifications are more reliable.



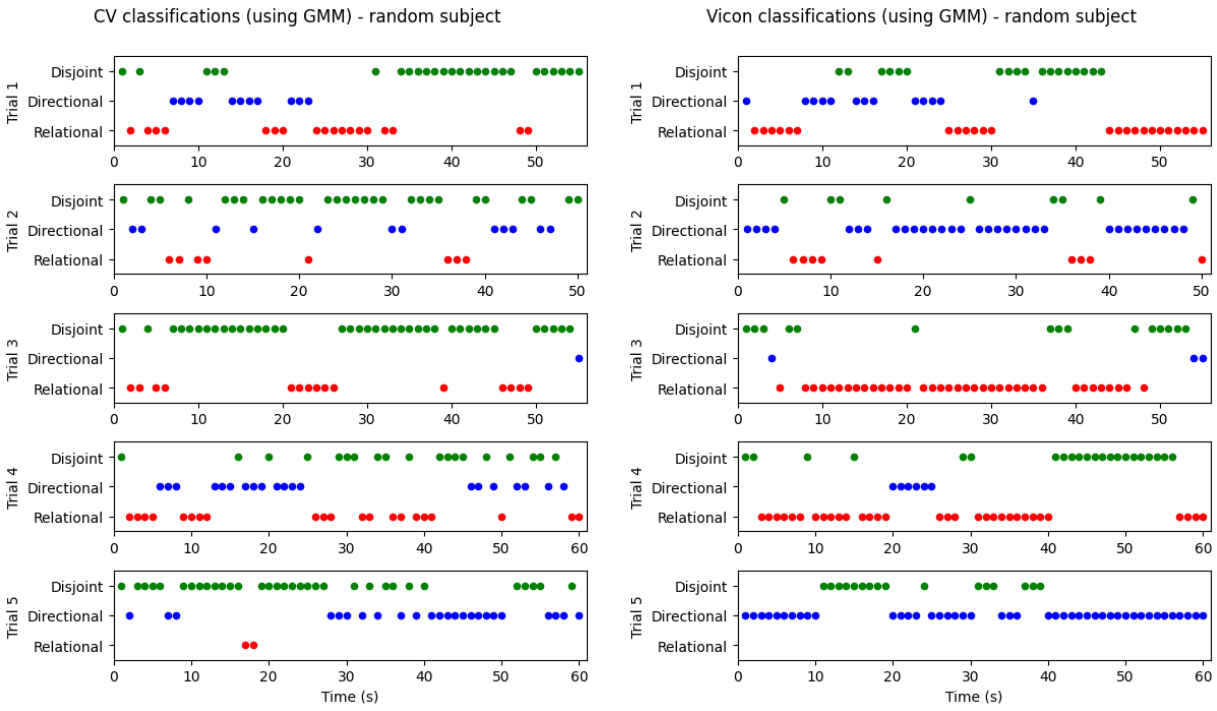
(a) (b)

Figure 4.4: Visual representation of interaction classifications (using rigid cutoffs) throughout a full video (one trial), with both CV (a) and Vicon MoCap (b) approach.

In Figure 4.4, we present a visualization that could qualitatively indicate the performance of our CV method (for a single video) in comparison with the Vicon MoCap approach.
The plot shows the temporal evolution of the interaction prediction over a single trial for both the CV (Figure 4.4a) and the Vicon MoCap (Figure 4.4b) approach. The second-by-second predictions are obtained classifying the computed interaction angles according to rigid cutoffs.
A similar pattern can be recognized for a portion of the video, while for other instants different predictions are assigned by the two methods.

As presented in Section 4.2 rigid predefined cutoffs may not fully capture the graded nature of attentional dynamics. To address this limitation, the WLS and GMM models have been complementary developed.
In Figure 4.5, it can be observed for a random subject (mother-child dyad) the visual comparison between CV (Figure 4.5a) and Vicon MoCap (Figure 4.5b) classifications deriving from the application of the GMM model to the calculated interaction angles.
Although for some portions of the videos the classification pattern is similar, in general we can still observe that the predictions of the two methods are not consistently the same at a second-by-second level.

In Section 4.3.2, we quantitatively assess the agreement of our method with respect to both Clinical coding and Vicon MoCap approaches. To do so, we compute the values of the metrics (both at second-level and at event-level) that we presented in Section 4.2.



(a) (b)

Figure 4.5: Visual representation of interaction classifications (applying GMM-based processing) throughout all the trials for a random subject, with both CV (a) and Vicon MoCap (b) approaches.

4.3.2 Quantitative Analysis

Following the qualitative inspection of interaction patterns, we present a quantitative analysis to formally evaluate agreement between methods at both second-level and event-level.

First of all, we need to point out that, since the dataset consists in 10 subjects (dyads) recorded over 5 trials of 60 seconds, we should have a theoretical total amount of 3000 second-level observations. However, quantitative analysis is conducted on 2777 of them. Specifically, 132 time points (approximately 5% of the total) are excluded due to missing classifications produced by our CV pipeline. These missing (null) outputs derive from failures in the initial pose detection stage, occurring under conditions of partial occlusion. Since our CV classification depends on successful pose estimation, frames where pose detection fails cannot be assigned an interaction label.
Additionally, another small number of observations are excluded to ensure perfect temporal alignment across methods: only time points for which a valid classification is simultaneously available for all compared methods are retained.

Table 4.1: Comparison between our CV (rows) and Clinical (columns) classifications.

	Disjoint (Cl.)	**Relational** (Cl.)	**Directional** (Cl.)	**Total** (Cl.)
Disjoint (CV)	1202	159	359	1720
Relational (CV)	102	69	33	204
Directional (CV)	606	66	181	853
Total (CV)	1910	294	573	2777

Table 4.2: Comparison between our CV (rows) and Vicon (columns) classifications.

	Disjoint (V.)	**Relational** (V.)	**Directional** (V.)	**Total** (V.)
Disjoint (CV)	1170	449	101	1720
Relational (CV)	120	78	6	204
Directional (CV)	602	118	133	853
Total (CV)	1892	645	240	2777

This procedure prevents artificial discrepancies that would arise from unequal observation counts between modalities. Consequently, all reported second-level and event-level metrics are computed on the same set of 2777 temporally aligned observations.

Tables 4.1 and 4.2 present the second-by-second confusion matrices for the comparisons between the CV method and the two reference approaches (clinical video coding and Vicon MoCap, respectively).

Tables 4.3 and 4.4 report the confusion matrices for the CV and Vicon MoCap comparisons after respectively applying WLS and GMM modeling to the computed interaction angles. At this temporal resolution, agreement is evaluated at the level of instantaneous (second-by-second) labels.

When compared with Clinical annotations (Table 4.1), the CV method shows a tendency to overestimate Directional interactions and underestimate Relational ones. Although the marginal distributions are similar (Disjoint being the most frequent class), the overall agreement remains limited, with Cohen's Kappa coefficient $\kappa = 0.05$ (just slightly over chance-level agreement).
The overall averaged F1-score is 0.40, indicating a low agreement at second-level.

A similar pattern emerges in the comparison with Vicon MoCap (Table 4.2), where $\kappa = 0.06$ and the overall averaged F1-score is 0.36. We can observe that Vicon identifies a higher prevalence of Relational interactions than Directional ones, producing systematic differences with respect to CV predictions.
These differences could probably reflect the two distinct measurement modalities (2D for CV, 3D for Vicon MoCap).

Table 4.3: Comparison between CV and Vicon classifications using WLS data.

	Disjoint (V.)	**Relational** (V.)	**Directional** (V.)	**Total** (V.)
Disjoint (CV)	713	517	129	1359
Relational (CV)	154	173	16	343
Directional (CV)	583	229	263	1075
Total (CV)	1450	919	408	2777

Table 4.4: Comparison between CV and Vicon classifications using GMM data.

	Disjoint (V.)	**Relational** (V.)	**Directional** (V.)	**Total** (V.)
Disjoint (CV)	456	410	363	1229
Relational (CV)	327	401	102	830
Directional (CV)	259	153	306	718
Total (CV)	1042	964	771	2777

Applying WLS-based processing (Table 4.3), second-level agreement ($\kappa = 0.09$) slightly increases.
GMM-based processing (Table 4.4) outputs the highest second-level agreement among the tested configurations ($\kappa = 0.12$, overall F1-score = 0.42). Under GMM, class distributions become more balanced and inter-class confusions are reduced.

Overall, second-level agreement remains low across comparisons, indicating that label differences due to small temporal misalignments substantially affect instantaneous metrics. Therefore, to evaluate agreement at a behaviorally meaningful level, we compute event-level metrics using an adaptive tolerance window ($\pm 1\sigma$), as described in Section 4.2. Results reveal a very different picture.

For the comparison between CV and Clinical, we obtain the following overall values: Precision = 0.92, Recall = 0.73, F1-score = 0.81, and ETMI = 0.69. Disjoint and Directional events show high precision and recall, while Relational events still show a low recall (0.50). This suggests that CV identifies fewer Relational episodes, even with temporal tolerance. For the comparison between CV and Vicon MoCap, overall F1-score is 0.69 and ETMI is 0.55. Directional events achieve perfect recall (1.00), but their precision decreases. This indicates that many episodes predicted as Directional events do not correspond to real events (identified by the Vicon MoCap method). On the contrary, Relational events show high precision and low recall. This highlights systematic differences between modalities.

After WLS-based processing, agreement significantly improves: overall F1-score = 0.77 and ETMI = 0.65. It reaches its highest values after GMM-based processing: overall F1-score = 0.87 and ETMI = 0.78. With GMM, all three behaviors show consistently high precision and recall (F1-scores between 0.82 and 0.90), indicating a strong alignment of episodes.

We observe that event-level agreement is always higher than second-level agreement across all comparisons. This confirms that many observed differences at a second-level could arise from small temporal shifts rather than real disagreement.

Expanding the tolerance window from $\pm 1\sigma$ to $\pm 2\sigma$ only produces small improvements in overall metrics. This behavior suggests that many temporally aligned events already fall inside the $\pm 1\sigma$ window. Therefore, another relaxation would not particularly modify agreement patterns. We can conclude that the adaptive $\pm 1\sigma$ window is sufficient to capture natural behavioral timing variability without artificially overestimate performances.

It is important to note that agreement metrics depend on which annotation source is treated as groundtruth. Precision and Recall are asymmetric by definition. Therefore, exchanging prediction and groundtruth swaps false positives and false negatives, modifying the relative contribution of over-segmentation and under-segmentation errors.

For example, when CV is evaluated against Vicon MoCap, high Directional recall indicates that many Directional events identified by Vicon (groundtruth) are detected by CV. However, swapping roles would highlight whether Vicon captures all episodes identified by CV. This strong dependency on the perspective highlights that disagreement may arise from differences in segmentation strategy rather than from complete classification errors.
The symmetric ETMI metric partially mitigates this issue by jointly accounting for unmatched predicted and groundtruth events.

Taken together, second-level analysis reveals a modest instantaneous label agreement, while event-level analysis demonstrates a substantially stronger alignment of behavioral episodes, particularly under GMM-based processing. These results indicate that discrepancies between approaches are primarily due to a temporal variability rather than a systematic misclassification.

Overall, this use case demonstrates that head direction estimation from 2D monocular videos can provide useful insights into HHI patterns, even in real-world scenarios such as the presented mother-child setting.
We are aware that limitations of our 2D CV approach arise from partial occlusions and the lack of depth information. However, our proposed pipeline offers a low-cost alternative to traditional MoCap systems or manual video coding, as it shows promising alignment with both other approaches.

Conclusion and Future Works

This thesis investigated whether HPE, interpreted as a proxy for visual attention, can support interaction modeling in monocular RGB settings. We designed a lightweight and interpretable framework that integrates HPE and geometrical spatial reasoning, instead of presenting a new end-to-end deep architecture.

The central hypothesis, that interaction targets can be inferred from structured geometrical relations between head orientation vectors and scene elements, was evaluated in two complementary contexts.

In the HOI use case, head orientation was combined with object localization and depth estimation to anticipate interaction targets. Results showed that head-based geometrical reasoning provides informative signals for short-term anticipation. Our findings support the assumption that visual attention precedes manipulation. Therefore head direction can be exploited for predictive interaction modeling.

In the HOI use case, dyadic attention events were detected by analyzing the angular relationships between head orientation vectors. Second-level agreement with both clinical video coding and MoCap groundtruths was low. On the contrary, the event-level analysis revealed substantially stronger alignment, particularly when temporal smoothing and probabilistic clustering (WLS and GMM) were applied. This suggests that disagreement at an instantaneous level is mainly due to temporal differences in classifications, rather than to substantial limitations of head-based interaction modeling.

Overall, the results indicate that meaningful interaction patterns can emerge from 2D head orientation cues alone, even without multi-camera 3D reconstruction or eye tracking.
Our framework demonstrates that low-cost geometrical reasoning represents a competitive and interpretable alternative approach in contexts often dominated by deep learning approaches.

Future works may explore the use of a different detector (or newer versions like YOLO26 [36], publicly released in January 2026), which could lead to more promising and accurate results. Moreover, the framework could be extended to multi-person or non-trivial environments to test its scalability and generalization capabilities. Finally, hybrid approaches

incorporating both geometrical reasoning and learning-based architectures could represent a promising alternative to balance interpretability and performance.

Overall, this thesis shows that head orientation is not just a low-level feature, but is also a visual cue that could support human-object and human-human interaction modeling in 2D vision-based social understanding.

Bibliography

[1] A.F. Abate, C. Bisogni, A. Castiglione, M. Nappi (2022). Head pose estimation: An extensive survey on recent techniques and applications. *Pattern Recognition* **127**.

[2] A. Abele (1986). Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior* **10**, 83–101.

[3] R. Algabri, A. Abdu, S. Lee (2024). Deep learning and machine learning techniques for head pose estimation: A survey. *Artificial Intelligence Review* **57**, 288.

[4] M. Alghamdi, N. Alhakbani, A. Al-Nafjan (2023). Assessing the potential of robotics technology for enhancing educational for children with autism spectrum disorder. *Behavioral Science* **13**, 598.

[5] R. Bakeman, V. Quera (2011). Sequential analysis and observational methods for behavioral sciences. *Cambridge University Press.*

[6] R. Birkl, D. Wofk, M. Müller (2023). MiDaS v3.1 – A model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460.*

[7] G. Cantarini, F.F. Tomenotti, N. Noceti, F. Odone (2022). HHP-net: A light heteroscedastic neural network for head pose estimation with uncertainty. In: *WACV*, 3521–3530.

[8] Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 172–186.

[9] F. Chang, J. Zeng, Q. Liu, S. Shan (2023). Gaze pattern recognition in dyadic communication. In: *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, 1–7.

[10] S.L. Colyer, M. Evans, D.P. Cosker, A.I.T. Salo (2018). A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports Medicine* **4**, 24.

[11] D. Das, M.G. Rashed, Y. Kobayashi, Y. Kuno (2015). Supporting human–robot interaction based on the level of visual focus of attention. In: *IEEE Transactions on Human-Machine Systems* **45**, 664–675.

[12] P.A. Dias, D. Malafronte, H. Medeiros, F. Odone (2020). Gaze estimation for assisted living environments. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 279–288.

[13] C. Dong, G. Du (2024). An enhanced real-time human pose estimation method based on modified YOLOv8 framework. *Scientific Reports* **14**, 8012.

[14] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian (2019). CenterNet: Keypoint triplets for object detection. In: *ICCV*.

[15] N.F. Duarte, M. Raković, J. Tasevski, M.I. Coco, A. Billard, J. Santos-Victor (2018). Action anticipation: Reading the intentions of humans and robots. In: *IEEE Robotics and Automation Letters* **3**, 4132–4139.

[16] L. Fan, Y. Chen, P. Wei, W. Wang, S.-C. Zhu (2018). Inferring shared attention in social scene videos. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6460–6468.

[17] M. Farhangi, A. Milan, G. Fallahi, E. Khankeshi-Zadeh (2025). Depth estimation and 3D reconstruction from a single image based on the MiDaS deep learning model. *Kharazmi Journal of Earth Sciences (KJES)* **11**, 152–174.

[18] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, E.-h. Zahzah (2016). Human pose estimation from monocular images: A comprehensive survey. *Sensors* **16**, 1966.

[19] M. Guarischi, E. Montagnani, G. Catalano, E. Saligari, S. Signorini, M. Gori (2025). From motion to interaction: How multisensory information shapes motor behaviors in children with visual impairment. *Research in Developmental Disabilities* **159**.

[20] M. Guarischi, E. Montagnani, M. Memeo, G. Catalano, C. Dagnino, N. Noceti, S. Signorini, M. Gori (2026). The angles between us: Modeling embodied joint attention episodes in mother-child dyads. **Currently work-in-progress paper.*

[21] K. Guo, Y. Huang, S. Sun, X. Song, M. Feng, Z. Liu, H. Song, T. Wang, J. Li, N. Akhtar, A.S. Mian (2025). Beyond human perception: Understanding multi-object world from monocular view. In: *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3751-3760.

[22] P. Her, L. Manderle, P.A. Dias, H. Medeiros, F. Odone (2023). Uncertainty-aware gaze tracking for assisted living environments. In: *IEEE Transactions on Image Processing* **32**, 2335–2347.

[23] Y. Huang, Y. Chen, J. Wang, P. Zhou, J. Lai, Q. Wang (2024). A robust and efficient method for effective facial keypoint detection. *Applied Sciences* **14**, 7153.

[24] C. Jongerius, R.S. Hessels, J.A. Romijn, E.M.A. Smets, M.A. Hillen (2020). The measurement of eye contact in human interactions: A scoping review. *Journal of Nonverbal Behavior* **44**, 363–389.

[25] B. Lai, S. Toyer, T. Nagarajan, R. Girdhar, S. Zha, J.M. Rehg, K. Kitani, K. Grauman, R. Desai, M. Liu (2024). Human action anticipation: A survey. *arXiv preprint arXiv:2410.14045*.

[26] T-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L.. Zitnick (2014). Microsoft COCO: Common objects in context. In: *Computer Vision - ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755.

[27] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C-L. Chang, M.G. Yong, J. Lee, W-T. Chang, W. Hua, M. Georg, M. Grundmann (2019). MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

[28] D. Maji, S. Nagori, M. Mathew, D. Poddar (2022). YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In: *Proceedings of 2022 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2637-2646.

[29] M.J. Marin-Jimenez, A. Zisserman, M. Eichner, V. Ferrari (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision* **106**, 282-296.

[30] S.S. Mukherjee, N.M. Robertson (2015). Deep head pose: Gaze-direction estimation in multimodal video. In: *IEEE Transactions on Multimedia* **17**, 2094-2107.

[31] E. Nicora, V.P. Pastore, N. Noceti (2023). Gck-maps: A scene unbiased representation for efficient human action recognition. In: *International Conference on Image Analysis and Processing*, 62–73.

[32] N. Noceti, A. Sciutti, F. Rea, F. Odone, G. Sandini (2015). Estimating human actions affinities across view. In: *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, 130–137.

[33] N. Noceti, A. Sciutti, G. Sandini (2015). Cognition helps vision: Recognizing biological motion using invariant dynamic cues. In: *V. Murino and E. Puppo (Eds.): ICIAP, Part II, LNCS 9280*, 676–686.

[34] J. Redmon, S. Divvala, R. Girshick, A. Farhadi (2016). You only look once: Unified, real-time object detection. In: *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] A. Rohan, M.J. Hasan, A. Petrovski (2025). A systematic literature review on deep learning-based depth estimation in computer vision. *arXiv preprint arXiv:2501.05147*.

[36] R. Sapkota, R.H. Cheppally, A. Sharda, M. Karkee, (2026). YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint arXiv:2509.25164*.

[37] J.S. Stahl (1999). Amplitude of human head movements associated with horixontal saccades. *Experimental Brain Research*, 41–54.

[38] R. Sun, Z. Lin, S. Leng, A. Wang, L. Zhao (2025). An in-depth analysis of 2D and 3D pose estimation techniques in deep learning: Methodologies and advances. *Electronics* **14**, 1307.

[39] F.F. Tomenotti, N. Noceti, F. Odone (2024). Head pose estimation with uncertainty and an application to dyadic interaction detection. *Computational Vision and Image Understanding* **243**.

[40] F.F. Tomenotti, N. Noceti (2024). Anticipation through Head Pose Estimation: A preliminary study. *arXiv preprint arXiv:2408.05516*.

[41] M. Væver, B. Beebe, O.I. Kirk, N. Snidmann, S. Harder, E. Tronick (2015). An automated approach for measuring infant head orientation in a face-to-face interaction. *Behavior Research Methods* **47**, 328–339.

[42] P. Verma, R. Srivastava, S. Tripathy (2025). An assessment towards 2D and 3D human pose estimation and its applications to activity recognition: A review. *SN Computer Science* **6**, 1–24.

[43] A. Vignolo, N. Noceti, F. Rea, A. Sciutti, F. Odone, G. Sandini (2017). Detecting biological motion for human-robot interaction: A link between perception and action. *Frontiers in Robotics and AI* **4**.

[44] X. Wang, J. Zhang, H. Zhang, S. Zhao, H. Liu (2021). Vision-based gaze estimation: A review. In: *IEEE Transactions on Cognitive and Developmental Systems* **14**, 316–332.

[45] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, M. Shah (2023). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys* **56**, 1–37.