POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

A Multi-View Skeleton-Based Pipeline for Body Behaviour Recognition in Inclusive Care Settings

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Simona Mastroberardino

Advisor: Prof. Maristella Matera

Co-advisor: Prof. Micol Spitale

Academic year: 2025-2026

1. Introduction and Objectives

Artificial intelligence for human behaviour recognition has achieved relevant progress in activity analysis, skeleton-based action recognition, and the automatic understanding of social interactions. However, many existing datasets and models are still developed in controlled settings and mainly represent neurotypical populations. This limits their applicability in real-world educational and care contexts, where behaviours are spontaneous, social situations are less structured, and individual variability is particularly high. These limitations are especially relevant for people with intellectual and developmental disabilities, who remain underrepresented in current AI benchmarks and whose body behaviours may not correspond to standard action categories [3, 5].

This thesis investigates the feasibility of a privacy-preserving pose-based pipeline for body behaviour recognition in a daycare centre attended by adults with intellectual and developmental disabilities, in collaboration with Fraternità e Amicizia. The work does not aim to implement a deployed assistive prototype or an automatic surveillance system. Instead, it focuses on the methodological and experimental foundations required to transform real-world multi-view video recordings into anonymous skeletal representations, annotated temporal windows, and learning-based benchmarks.

The proposed approach follows a human-centred perspective. The pipeline explicitly avoids facial recognition, voice identification, emotion recognition, biometric profiling, and personal identification. Only observable and non-invasive signals are considered, including body pose, movement, posture, tracking continuity, and coarse spatial relations. Manual annotations were defined with the support of educators and were designed to describe visible behaviours rather than infer emotions, intentions, or clinical states.

In this context, the main objective of the thesis is to assess whether skeleton-based representations can provide a learnable signal for recognising observable body behaviours in a naturalistic and inclusive setting. To this end, the work combines multi-view acquisition, camera calibration, 2D pose extraction, anonymous tracking, 3D reconstruction, ELAN-based annotation, dataset construction, and learning-based evaluation. Several modelling strategies are explored, including SVM baselines on interpretable 2D and 3D features, PoseC3D

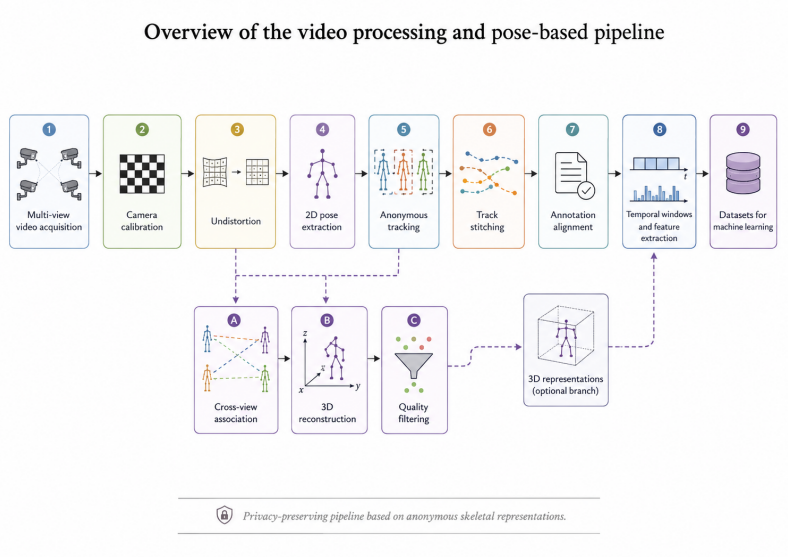

Figure 1: Overview of the privacy-preserving video processing and pose-based pipeline.

experiments on skeleton-based representations, and exploratory relational and annotation-aware analyses [2, 4, 7].

Overall, the thesis shows that pose-derived features contain a partially learnable behavioural signal, while also revealing important limitations related to class imbalance, occlusions, tracking fragmentation, incomplete relational visibility, and weak generalization across sessions. The contribution should therefore be understood as a feasibility and benchmarking study for privacy-preserving body behaviour recognition, rather than as a ready-to-deploy monitoring tool.

2. Pose-Based Data Processing Pipeline

As summarized in Figure 1, the processing workflow was designed to transform raw multi-view recordings into privacy-preserving skeletal representations suitable for annotation alignment and learning-based analysis[1]. Data were collected in collaboration with *Fraternità e Amicizia* in a real daycare environment using two fixed cameras placed in complementary viewpoints of the room. This configuration made it possible to observe everyday activities from different perspectives while preserving the ecological validity of the setting and avoiding intrusive sensing devices.

[1]Data acquisition and processing were authorized by Politecnico di Milano's Ethical Committee - authorization no. 71/2025

The raw videos were first converted into a standard format and pre-processed through camera calibration and lens undistortion. Intrinsic calibration was used to estimate the camera matrix and distortion coefficients of each view, while the relative configuration between cameras provided the geometric basis for the multi-view branch. This step was necessary to obtain more consistent observations across views and to support subsequent 3D reconstruction.

For the 2D branch, human poses were extracted independently from each camera view. Each detected person was represented through body landmarks, from which anonymous tracks were constructed over time. Since the recordings were collected in an unconstrained real-world environment, pose extraction was affected by occlusions, seated postures, partial visibility, and close interactions between people. For this reason, the tracking pipeline was refined to favour temporally consistent trajectories, remove short or low-quality detections, and preserve informative partial upper-body poses when the full body was not visible.

A further stitching step was applied to reduce track fragmentation. Compatible track segments were merged only when temporal, spatial, and geometric consistency criteria were satisfied. This strategy was intentionally conservative: uncertain associations were discarded rather than forcing potentially incorrect identity links. As a result, the pipeline produced anonymous person-level trajectories suitable for

alignment with manual annotations, without relying on face recognition, appearance matching, voice identification, or biometric information.
In parallel, a multi-view 3D branch was developed to reconstruct skeletal information from the two camera views. The approach relied on explicit geometric triangulation when corresponding 2D keypoints were available, followed by reprojection-based filtering to remove unreliable points. A fusion step was then used to improve continuity by retaining triangulated joints whenever possible and completing missing information only when necessary. This design is consistent with recent work emphasizing the role of multi-view geometry for robust 3D pose estimation under occlusions and viewpoint changes [1, 4].
The final outputs of the pipeline consist of 2D pose tracks for each camera, stitched anonymous trajectories, and multi-view 3D or fused skeletal representations. These outputs do not contain facial, vocal, emotional, or identity-related information. Instead, they provide geometric and temporal descriptors of observable body behaviour, which are later transformed into window-level features for supervised learning.

3. Annotation and Dataset Construction

The annotation process was designed to convert real-world observations into temporal labels suitable for supervised learning, while preserving the human-centered nature of the project. Since the recorded activities were spontaneous and socially situated, the codebook was not defined as a list of standard actions, but as a set of observable body and relational behaviours relevant to the educational context. The annotation protocol was developed iteratively, combining exploratory video review with feedback from educators at Fraternit'a e Amicizia.
Annotations were created in ELAN and organized into four conceptual groups: individual behaviours, relational behaviours, support-context annotations, and tag notes. Individual labels describe behaviours associated with a single person, such as transitions between sitting and standing, closed posture, disengagement from the group, or repetitive movements. Relational labels describe dyadic configurations, such as proximity, mutual orientation, or affiliative contact. Support-context and tag annotations were used to mark educator interventions, ambiguity, occlusions, or visibility issues, but were not treated as main target classes.
For the supervised person-level task, the original codebook was reduced to four target classes: *Stand*, *Sit*, *ClosedPosture*, and *DisengagedRepetitive*. The last class aggregates disengagement from the group and repetitive movements, in order to obtain a more robust label for learning. Relational labels, such as *SocialOrientation* and *AffiliativeContact*, were kept separate and used only in exploratory pair-level analyses, because they refer to pairs of people rather than to single-person samples.
A key step was the alignment between manual annotations and automatic pose tracks. ELAN annotations refer to anonymous session-level IDs, while the processing pipeline produces technical stitched-track IDs. These two levels were connected through a conservative manual mapping procedure based on temporal overlap, spatial coherence, and visual inspection. Only high-confidence mappings referring to participants and marked as usable for machine learning were retained. Educators, ambiguous mappings, and unreliable tracks were excluded from the supervised datasets.
After mapping, each annotated participant was represented through a discrete timeline sampled at 2 Hz. The annotated intervals were projected onto this timeline, and fixed-length sliding windows were generated with durations of 2, 4, and 6 seconds and a stride of 0.5 seconds. Each window was assigned a label through majority voting and was retained only when one positive class covered at least 50% of its timesteps. Windows without a sufficiently represented target behaviour, or windows affected by ambiguous label assignments, were discarded.
This process produced separate dataset versions for the two 2D camera views and for the multi-view 3D branch. The main 2-second person-level datasets contain 17,372 samples for CAM1, 11,580 samples for CAM2, and 607 trainable samples for the 3D representation. The resulting datasets are strongly imbalanced, especially toward *ClosedPosture*, reflecting the natural distribution of behaviours in the observed setting rather than an artificially balanced experimental

protocol.

4. Learning-Based Behaviour Recognition

The learning-based part of the thesis investigates whether the pose-derived representations constructed from the annotated recordings contain a signal that can be used to distinguish observable body behaviours. The main task was formulated as a four-class person-level classification problem, using the final individual labels *Stand*, *Sit*, *ClosedPosture*, and *DisengagedRepetitive*. Each sample corresponds to an annotated person observed within a temporal window and described through features extracted from the corresponding pose tracks.

The first experimental branch focused on 2D tabular features extracted separately from the two camera views. These features summarize tracking continuity, pose availability, detection quality, bounding-box position and shape, movement of the body centre, and coarse body-geometry measures. This representation was intentionally simple and interpretable, since the goal was to establish a robust baseline before introducing more complex models. Support Vector Machines were used as the main classifier, with feature standardization, class balancing, and hyperparameter tuning. Performance was evaluated primarily through macro-F1 and balanced accuracy, due to the strong imbalance among behavioural classes.

A second branch explored the contribution of the multi-view 3D representation. In this case, the input features do not describe only image-plane motion, but also the availability and quality of the reconstructed 3D signal, including usable 3D timesteps, joint availability, reconstruction quality, and multi-view or fused-signal ratios. This branch was more selective than the 2D one, because a window could be retained only when the multi-view reconstruction was sufficiently reliable. Although smaller, the 3D dataset made it possible to test whether geometric information across views could provide a more balanced representation of body behaviour.

In addition to tabular baselines, the thesis explored PoseC3D as a skeleton-based deep-learning model. PoseC3D represents skeleton sequences through pose-derived heatmap volumes and has been proposed as a robust alternative to graph-based skeleton action recognition, especially in the presence of pose estimation noise and multi-person settings [2]. In this thesis, PoseC3D was evaluated both on 2D pose-derived samples and on a dedicated 3D-derived export. These experiments were treated as exploratory, because the available dataset is small and strongly imbalanced compared with standard action-recognition benchmarks [7].

The thesis also included two exploratory extensions beyond the main person-level task. First, a relational branch was designed to model dyadic behaviours such as *SocialOrientation* and *AffiliativeContact*. This branch was motivated by research on bodily behaviour and social interaction analysis [5, 6], but it was limited by the low number of windows in which both subjects of a dyad were simultaneously visible and reliably mapped. For this reason, relational modelling was treated as a feasibility analysis rather than as a consolidated benchmark.

Second, an annotation-aware extension was introduced to test whether contextual annotations could enrich the feature space without leaking the target label. These features summarized the presence of overlapping relational annotations, support-context annotations, tag notes, and other-person events within the same temporal window. The goal was not to replace pose-based recognition with annotation-derived labels, but to assess whether contextual information could improve or explain model behaviour.

Overall, the experimental design compares increasingly rich representations: monocular 2D tabular features, multi-view 3D quality-aware features, skeleton-based deep representations, relational pair-level features, and annotation-aware contextual features. This structure makes it possible to evaluate not only predictive performance, but also the methodological limits of behaviour recognition in a real, inclusive, and privacy-sensitive setting.

5. Results and Discussion

The experimental results show that pose-derived representations contain a learnable signal, but also that behaviour recognition in this setting remains challenging. As reported in Table 1, the monocular 2D baselines achieved the highest overall accuracy, especially in the CAM2 setting. However, the gap between accuracy and

Table 1: Summary of the main experimental results.

Setting	Model	Results
CAM1 2D, 2s only	SVM	Acc. 0.69; BAcc. 0.30; MF1 0.31; WF1 0.70
CAM2 2D, 2s only	SVM	Acc. 0.73; BAcc. 0.35; MF1 0.32; WF1 0.79
Multi-view 3D, 2s trainable	SVM	Acc. 0.52; BAcc. 0.38; MF1 0.34; WF1 0.44
3D-derived export	PoseC3D	Top-1 0.68; Mean class acc. 0.59

macro-F1 reveals the effect of the strong class imbalance, with models tending to favour the dominant class, *ClosedPosture*, over shorter and less frequent behaviours such as *Stand* and *Sit*.
The comparison between CAM1 and CAM2 confirms that the two camera views are not equivalent. Although both views observe the same environment, differences in viewpoint, occlusions, and pose quality lead to different dataset distributions and model behaviour. CAM2 achieved the best monocular SVM performance, with accuracy 0.73 and macro-F1 0.32, while CAM1 obtained accuracy 0.69 and macro-F1 0.31. In both cases, the 2-second window configuration was more stable than mixed window sizes, suggesting that shorter windows better preserve brief transitional behaviours.
The multi-view 3D branch produced a much smaller trainable dataset, because each retained sample required sufficient reconstruction quality, reliable identity alignment, and usable temporal coverage. Despite this reduction, the 3D SVM baseline achieved the best balanced accuracy and macro-F1 among the SVM settings. This suggests that multi-view geometric information may support a more balanced representation of behaviour, even when fewer samples are available. The PoseC3D experiment on the 3D-derived export further supports this observation, reaching top-1 accuracy 0.68 and mean class accuracy 0.59 on the exported subset. However, these results should be interpreted cautiously because the 3D-derived dataset remains small.
The exploratory relational branch highlighted a different limitation. Although relational labels are important for describing social dynamics, the available data did not provide enough reliable windows in which both members of a dyad were simultaneously visible and mapped. As a consequence, graph-based relational modelling was not methodologically robust in the current dataset. Tabular relational classifiers were tested as exploratory alternatives, but their performance remained strongly dependent on the split and on video-specific conditions.
The annotation-aware experiments showed that contextual annotations can introduce an additional signal, especially under random splits. However, grouped and leave-one-session-out evaluations revealed that this signal does not consistently improve generalization to unseen sessions. This suggests that annotation-aware features may capture context-specific patterns, but are not yet sufficient to overcome the variability of the real-world setting.
Overall, the results indicate that privacy-preserving skeletal representations are informative, but not sufficient on their own to obtain robust and generalizable behaviour recognition in this context. The main limitations are not only algorithmic, but also structural: class imbalance, occlusions, partial visibility, fragmented tracks, limited 3D coverage, and the intrinsic variability of spontaneous behaviours all affect model performance. For this reason, the results should be interpreted as a feasibility benchmark rather than as evidence of a ready-to-use recognition system.

6. Conclusions

This thesis investigated the feasibility of privacy-preserving body behaviour recognition in a real daycare setting involving adults with intellectual and developmental disabilities. The work developed an end-to-end methodological pipeline, from multi-view video acquisition and pose extraction to anonymous tracking, educator-informed annotation, dataset construction, and learning-based evaluation. Throughout the process, the analysis was intentionally limited to non-invasive skeletal and contextual signals, avoiding facial recognition,

voice identification, emotion recognition, biometric profiling, and personal identification.
The results show that pose-derived representations can capture a partially learnable signal for observable body behaviours, especially when short temporal windows and multi-view information are considered. At the same time, the experiments highlight the difficulty of this task in naturalistic settings. Class imbalance, occlusions, partial visibility, tracking fragmentation, limited 3D coverage, and session-specific variability strongly affect model performance and generalization. The relational and annotation-aware analyses further show that richer contextual information is promising, but still constrained by data sparsity and incomplete simultaneous visibility.
Overall, the thesis contributes a reproducible feasibility benchmark rather than a ready-to-deploy recognition system. Future work should extend the dataset across more sessions and participants, improve robust multi-person tracking, strengthen multi-view temporal modelling, and further involve educators in the interpretation of model outputs. These directions are essential for developing inclusive AI methods that remain technically reliable, ethically cautious, and meaningful for real educational and care practices.

References

[1] Bo-Han Chen and Chia-chi Tsai. 3DSA: Multi-view 3D Human Pose Estimation With 3D Space Attention Mechanisms. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, volume 15085, pages 323–339. Springer Nature Switzerland, Cham, 2025. Series Title: Lecture Notes in Computer Science.

[2] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting Skeleton-based Action Recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2959–2968, New Orleans, LA, USA, June 2022. IEEE.

[3] Giulia Huang, Maristella Matera, and Micol Spitale. Inclusive ai for group interactions: Predicting gaze-direction behaviors in people with intellectual and developmental disabilities. *arXiv preprint arXiv:2603.14460*, 2026.

[4] Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, and Steven L. Waslander. Multiple View Geometry Transformers for 3D Human Pose Estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 708–717, Seattle, WA, USA, June 2024. IEEE.

[5] Philipp Müller, Michal Balazia, Tobias Baur, Michael Dietz, Alexander Heimerl, Dominik Schiller, Mohammed Guermal, Dominike Thomas, François Brémond, Jan Alexandersson, Elisabeth André, and Andreas Bulling. MultiMediate '23: Engagement Estimation and Bodily Behaviour Recognition in Social Interactions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9640–9645, Ottawa ON Canada, October 2023. ACM.

[6] Chirag Raman, Jose Vargas-Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild.

[7] Bin Ren, Mengyuan Liu, Runwei Ding, and Hong Liu. A Survey on 3D Skeleton-Based Action Recognition Using Learning Method. *Cyborg and Bionic Systems*, 5:0100, January 2024.