Pointing Gesture Recognition from 3D Human Skeleton Data

Miquelly Nicolini Lucas*, Gabriel Donna Altoé*, Luiz Carlos Cosmi Filho*, and Raquel Frizera Vassallo*
*Electrical Engineering Department
Federal University of Espírito Santo, Vitória, Brazil
{miquelly.lucas, gabriel.altoe, luiz.cosmi}@edu.ufes.br, raquel.vassallo@ufes.br

***Abstract*—Gesture recognition in Human-Machine Interaction (HMI) refers to the automatic detection of human gestures, assigning semantic meaning to physical movements and enabling interaction with computers, robots, or the analysis of human behavior. This work proposes a method for static gesture recognition, focusing specifically on the "pointing" gesture, based on 3D human skeletons reconstructed through a multi-camera system. The objective is to automatically detect the presence or absence of the gesture using spatial pose data derived from the three-dimensional reconstruction of the human body from multiple viewpoints. Following manual annotation and segmentation of the gesture sequences, structural features were extracted from the skeletons. A normalization process was applied, which performs translation to the origin and rotation to align the skeleton with the X-axis. Classical supervised machine learning models were then employed to classify body poses: Logistic Regression, Random Forest, and Decision Tree. Experiments were carried out using Leave-One-Subject-Out (LOSO) cross-validation. The results demonstrate the viability of the proposed approach for applications in intelligent environments, where recognizing the pointing gesture could be used to indicate goals, highlight objects of interest, or define target positions for mobile robots.**

I. Introduction

In Human-Machine Interaction (HMI), gestures are regarded as one of the least intrusive and most natural forms of communication. They are often preferred over speech, as they typically do not require a grammatical structure for their formulation [1]. These gestures can be dynamic or static: (i) static gestures do not vary in time, e.g. pointing or grasping; (ii) dynamic gestures can be defined as a set of small movements that makes possible the communication between individuals, e.g. waving or drawing letters in the air [2].

In recent years, various approaches have been proposed for static gesture recognition. Osipov *et al.* [3] proposed a real-time method for static gesture recognition based on hand skeletons, specifically targeting sign-digit gestures, using a Support Vector Machine (SVM) classifier. Similarly, Cosmi-Filho *et al.* [4] introduced a decision tree-based approach for recognizing static gestures using 3D skeletons obtained from a multi-camera system, aiming to identify when users raised their left hand, right hand, or both.

Among the various types of static gestures, pointing gestures have gained significant attention due to their intuitive nature and versatility in real-world applications. Lorentz *et al.* [5] present an application of human-robot interaction using pointing gestures and a humanoid robot. Čorňák *et al.* [6] developed a collaborative method for interaction between a human operator and robotic manipulator also using pointing gestures. Medeiros *et al.* [7] used a pointing gesture interface for human-drone interaction in a firefighting scenario.

To recognize the pointing gesture, many existing methods rely on single monocular RGB cameras or stereo camera setups. In contrast, this work presents an approach for recognizing the static pointing gesture using a system composed of multiple calibrated cameras. This setup enables the 3D reconstruction of skeletal data, which is then employed to identify the pointing gesture. To this end, we propose the use of classical machine learning models to classify the gesture. Once the gesture is detected, a 3D line is estimated based on the shoulder-to-wrist joints. This methodology thus allows seamless integration with other systems, enabling functionalities such as indicating goals, highlighting objects of interest, or defining target positions for mobile robots (see Fig. 1).

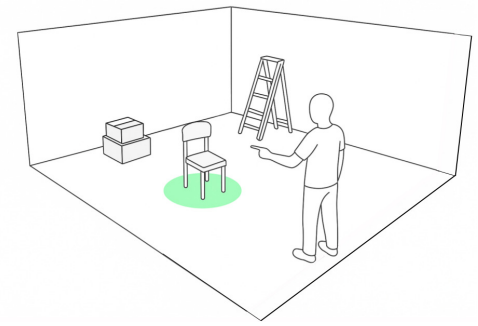


Fig. 1. Example of application using pointing gesture recognition: highlighting object of interest.

II. Proposal

For better understanding, our proposal is structured into five main stages: the method employed for 3D skeleton pose estimation using a calibrated multi-camera system (Section II-A), the dataset used in this study (Section II-B), the preprocessing techniques applied to the data (Section II-C), the machine learning models used for gesture classification (Section II-D), and the method for estimating the 3D line (Section II-E).

A. 3D pose estimation

The skeleton estimation method employed in this work is based on the adaptation proposed by Cosmi-Filho *et al.* [4],

which builds upon the work of Chu *et al.* [8].

Using images captured simultaneously from multiple calibrated cameras, the system processes the set of different viewpoints by applying a person detector, aiming to isolate each individual in the scene and extract their respective bounding boxes. These bounding boxes are then passed through a 2D pose estimator, which determines the bidimensional locations of body parts and joints, resulting in 2D joint coordinates. For this purpose, YOLOv11 [9] was used as the person detector, and RTMPose [10] for pose estimation.

Then, 2D poses are associated across multiple views, and a 3D pose reconstruction is performed to estimate the full-body skeletons of all individuals in the scene. To achieve this, the method leverages temporal consistency, associating the 2D poses detected in each camera view with previously reconstructed skeletons. To handle imprecise or noisy samples, the method incorporates a part-aware measurement that evaluates the affinity between specific body parts, along with an outlier filtering mechanism applied to the 2D estimations during the reconstruction process [4].

B. Dataset

The dataset used in this study was built from recordings conducted in a laboratory environment equipped with four cameras (see Fig. 2). Approximately 5 minutes of video were collected per participant, during which individuals were instructed to perform the pointing gesture in various positions and directions, covering the following scenarios: (i) right arm, (ii) left arm, and (iii) arms at rest.

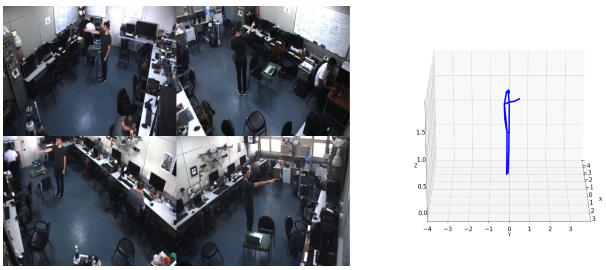


Fig. 2. Reconstruction example using a four-view camera system.

Subsequently, the frames from each recording were segmented and labeled as '1' (pointing gesture) or '0' (no gesture). Then, the method described in the previous section was used to estimate the 3D joint coordinates of the subjects. For training purposes, only the 3D coordinates of the shoulder, elbow, and wrist from both arms were used.

The final dataset comprises a total of 4,606 frames, collected from four individuals (the authors of this work) with variation in height and gender. The class distribution in the dataset consists of 60.6% "gesture" frames and 39.4% "no gesture" frames.

C. Data pre-processing

To ensure invariance to the global position and orientation of the skeletons, a two-step normalization process was applied to each frame in the sequence:

- Translation to the origin: All joints in the skeleton are translated so that a reference joint (e.g., the nose) is positioned at the origin of the coordinate system.
- Alignment with the X-axis: The skeleton is then rotated so that the vector defined between the shoulder joints becomes aligned with the X-axis.

This normalization is applied to every skeleton at each time step, producing a standardized spatial representation of the motion.

D. Machine Learning Models

The following machine learning models were employed in our experiments to classify to gesture of interest:

- Random Forest (RF) [11]: An ensemble learning method, which combines multiple individual models (in this case, decision trees) to create a more powerful and reliable model. It improves predictive performance and reduces overfitting by leveraging the diversity of the ensemble.
- Logistic Regression (LR) [12]: A linear classification model that estimates the probability of a binary outcome based on a linear combination of input features.
- Decision Tree (DT) [13]: A tree-based model that recursively splits the feature space based on criteria like information gain or Gini impurity. It models decisions as a sequence of simple rules, making it both interpretable and effective for capturing non-linear relationships.

E. 3D line estimation

After classifying a given sample as a pointing gesture, the pointing direction is defined using the parametric equation of a line in 3D space. Let P_{shoulder} and P_{wrist} denote the 3D coordinates of the shoulder and wrist, respectively. The direction vector v of the line is obtained by the difference between these points, defined in Equation (1).

$$v = \overrightarrow{P_{\text{shoulder}} P_{\text{wrist}}} \tag{1}$$

The wrist is selected as the reference point P_0 lying on the line, due to its proximity to the hand. Thus, the line r is characterized as a parametric equation, defined in Equation (2).

$$r : P = P_0 + t\vec{v} \quad ; \quad t \in \mathbb{R} \tag{2}$$

Here, P denotes a generic point on the line, P_0 corresponds to the 3D coordinates of the wrist, v is the direction vector derived from the shoulder–wrist displacement, and t is a real-valued scalar parameter.

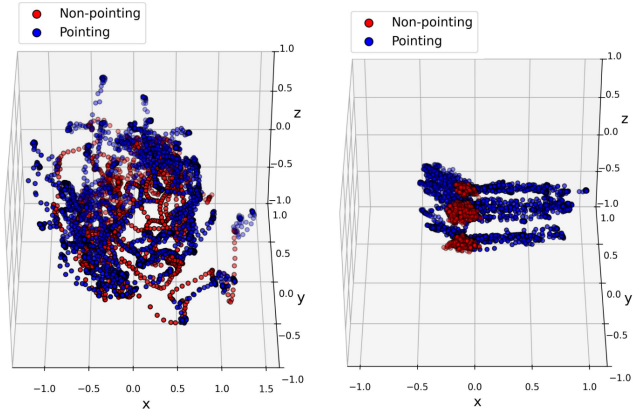
To determine which arm is performing the pointing gesture (right, left, both, or none), a distance-based criterion is applied individually to each arm: if the distance between the shoulder and wrist of a given arm exceeds a predefined threshold (determined empirically through a parameter sweep), the arm is considered to be executing the gesture.

III. Experiments and Results

First, we analyzed the impact of the normalization procedure through a Principal Component Analysis (PCA) (Section III-A). Then, we described the performance evaluation of different classification models using a Leave-One-Out Cross-Validation strategy (Section III-B).

A. Normalization effects

To evaluate the effectiveness of the normalization procedure, Principal Component Analysis (PCA) was applied to the samples both before and after normalization. In Figure 3a, all samples from the dataset were projected into a lower-dimensional space (three dimensions) without any prior normalization. In contrast, Figure 3b shows the same projection applied to the samples after the normalization process described in Section II-C.



(a) w/o normalization (b) w/ normalization

Fig. 3. Principal Component Analysis (PCA) with and without normalization.

The results revealed that, prior to normalization, the pointing gestures exhibited high variability in orientation and position. After applying the normalization process a spatial pattern emerged, with gesture samples becoming more consistently aligned and clustered. This indicates that the normalization step effectively reduces pose variability, facilitating the learning process for the classifiers.

In this context, the PCA algorithm was used solely to illustrate the importance and necessity of data normalization. For the classification task, the normalized joint information is employed to perform binary classification of pointing and non-pointing gestures.

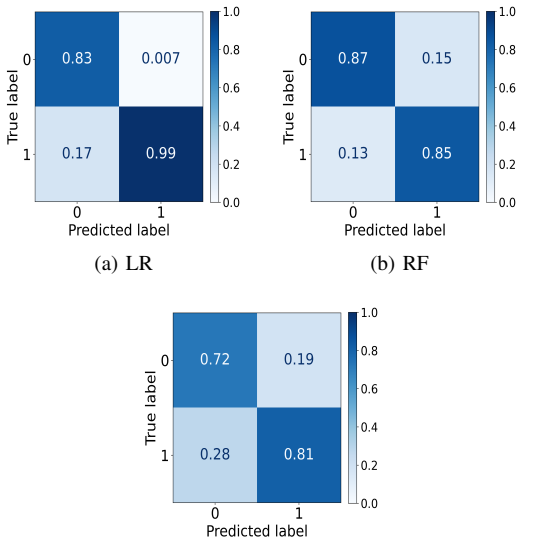
B. Cross validation

To train the models, first we perform a grid search cross-validation parameter optimization. After finding the best parameters for each machine learning model, we used the leave-one-subject-out cross-validation strategy for evaluation [14]. This ensures that samples from the same person are never used simultaneously for training and testing. Additionally, to mitigate class imbalance within each fold, the models were configured to automatically adjust class weights inversely proportional to class frequency.

The detailed results (accuracy, precision, recall and F1-score) for each fold and the corresponding averages for each classifier are presented in Table I. Additionally, Figure 4 shows the mean confusion matrices for each classifier. These matrices represent the average across all validation folds. As shown, the Logistic Regression (LR) classifier achieved the highest average performance among the evaluated models.

Method	Test Subject	Acc.	Prec.	Rec.	F1
LR	1	92.35	99.62	88.13	93.52
	2	92.63	99.85	86.96	92.96
	3	88.70	99.28	81.66	89.61
	4	91.79	97.73	90.02	93.72
	Mean	91.37	99.12	86.69	92.45
RF	1	87.51	86.97	94.18	90.43
	2	85.97	84.91	91.14	87.91
	3	79.88	77.32	93.79	84.76
	4	89.57	91.76	93.04	92.40
	Mean	85.73	85.24	93.04	88.87
DT	1	82.60	81.34	93.73	87.10
	2	59.46	64.42	61.65	63.00
	3	87.56	92.80	85.80	89.16
	4	90.84	94.72	91.65	93.16
	Mean	80.11	83.32	83.21	83.11

TABLE I
Cross validation results with different classifiers.



(a) LR (b) RF

(c) DT

Fig. 4. Mean confusion matrices obtained across all folds for the different classifiers.

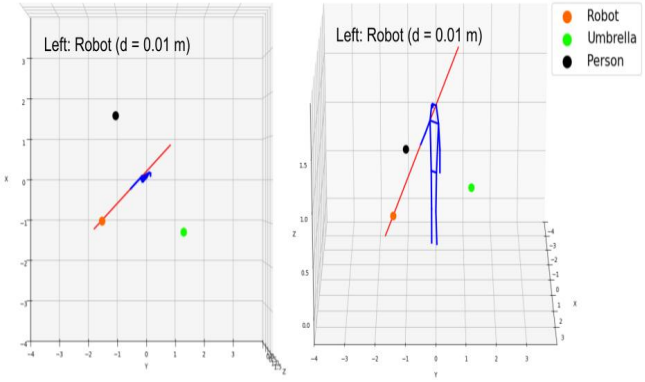
C. Evaluating system

To evaluate our method in a real-world scenario, we conducted an experiment where specific objects were placed at known and fixed positions. These positions were not estimated, but inserted virtually into the system.

To verify whether the system could correctly identify when a person is pointing and to which object, we designed an experiment placing objects in known and fixed positions. In

the frame illustrated in Figure 5a, the subject is pointing to the object labeled "robot". The gesture was correctly classified, the pointing line was estimated, and the distance between the object and the line was $d = 0.01, m$, confirming that the person was indeed pointing to the "robot". A video of this experiment can be seen at this link.

(a)



(b)

Fig. 5. Experiment in a multi-camera system with four views.

IV. Conclusion

This work presented a method for static pointing gesture recognition based on 3D human pose estimation from a calibrated multi-camera system. By integrating person detection, 2D pose estimation, and multi-view 3D reconstruction, the system generates skeletal representations of individuals in a scene. These representations are preprocessed through spatial normalization and used to train machine learning classifiers capable of distinguishing pointing gestures. Once a gesture is detected, a 3D pointing vector is estimated using the shoulder–wrist joint configuration, enabling integration with downstream applications.

The proposed approach may be used in several applications, including human-robot interaction, context-aware systems, and spatial referencing tasks in smart homes, museums, and collaborative workspaces. However, the method depends on reliable multi-view detection and accurate calibration between cameras. Future work will explore the use of temporal models to capture gesture dynamics, the incorporation of deep learning approaches for improved classification, and the evaluation of the method on more diverse and complex datasets.

Acknowledgment

The authors would like to thank FAPES - Fundação de Amparo à Pesquisa e Inovação do Espírito Santo for the financial support through the project "Cooperação Humano-Robô-Ambiente em um Espaço Inteligente A2AI", T.O. 777/2024, and the scholarships granted to two of the authors.

References

[1] R. Arnheim, "Hand and mind: What gestures reveal about thought," *Leonardo*, vol. 27, no. 4, pp. 358–358, 1994. [Online]. Available: http://www.jstor.org/stable/1576015

[2] C. C. Santos, L. C. Cosmi, A. P. d. Carmo, J. Samatelo, J. Santos-Victor, and R. F. Vassallo, "Reconhecimento online de gestos dinâmicos para ambientes interacionais multicâmeras," in *XV Simpósio Brasileiro de Automação Inteligente - SBAI 2021*, 01 2021.

[3] A. Osipov and M. Ostanin, "Real-time static custom gestures recognition based on skeleton hand," in *2021 International Conference "Nonlinearity, Information and Robotics" (NIR)*, 2021, pp. 1–4.

[4] L. C. Cosmi Filho, M. D. d. Oliveira, M. N. Lucas, L. F. Follador, and R. F. Vassallo, "An approach based on a programmable intelligent space for human-robot interaction," in *2024 Latin American Robotics Symposium (LARS)*, 2024, pp. 1–6.

[5] V. Lorentz, M. Weiss, K. Hildebrand, and I. Boblan, "Pointing gestures for human-robot interaction with the humanoid robot digit," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023, pp. 1886–1892.

[6] M. Čorňák, M. Tölgyessy, and P. Hubinský, "Innovative collaborative method for interaction between a human operator and robotic manipulator using pointing gestures," *Applied Sciences*, vol. 12, no. 1, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/1/258

[7] A. C. S. Medeiros, P. Ratsamee, Y. Uranishi, T. Mashita, and H. Takemura, "Human-drone interaction: Using pointing gesture to define a target object," in *Human-Computer Interaction. Multimodal and Natural Interaction*, M. Kurosu, Ed. Cham: Springer International Publishing, 2020, pp. 688–705.

[8] H. Chu, J.-H. Lee, Y.-C. Lee, C.-H. Hsu, J.-D. Li, and C.-S. Chen, "Part-aware measurement for robust multi-view multi-human 3d pose estimation and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1472–1481.

[9] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[10] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-time multi-person pose estimation based on mmpose," *ArXiv*, vol. abs/2303.07399, 2023. [Online]. Available: https://arxiv.org/abs/2303.07399

[11] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[12] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[13] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Chapman and Hall/CRC, 2017.

[14] Q. F. Gronau and E.-J. Wagenmakers, "Limitations of bayesian leave-one-out cross-validation for model selection," *Computational brain & behavior*, vol. 2, no. 1, pp. 1–11, 2019.