Towards Explainable Human Motion Prediction in Collaborative Robotics

Michael Vanuzzo[1†], Francesco Borsatti[2†], Marco Casarin[1], Mattia Guidolin[1], Monica Reggiani[1], and Stefano Michieletto[1]

[1] University of Padova,
Department of Management and Engineering (DTG),
Stradella S. Nicola, 3, 36100 Vicenza, Italy
[2] University of Padova,
Department of Information Engineering (DEI),
Via Gradenigo, 6/B, 35121 Padova, Italy
{michael.vanuzzo, francesco.borsatti.1, marco.casarin.4}@phd.unipd.it,
{mattia.guidolin, monica.reggiani, stefano.michieletto}@unipd.it

Abstract. Predicting human motion is challenging due to its complex and non-deterministic nature. This is particularly true in the context of Collaborative Robotics, where the presence of the robot significantly influences human movements. Current Deep Learning models excel at modeling this complexity but are often regarded as black boxes. Explainable Artificial Intelligence (XAI) offers a way to interpret these models. In this work, we introduce an XAI approach to identify key features in a Human Motion Prediction (HMP) system. Additionally, we semantically associate action labels to the joint rotations representing human motion to further improve the interpretability and precision of the model. We evaluated our system using the AMASS dataset and BABEL labels. Experimental results demonstrated the importance of specific action-related features, enhancing prediction accuracy compared to the Zero-Velocity baseline model.

Keywords: Human Motion Prediction, Human Action Semantic, Explainable AI, Collaborative Robotics

1 Introduction

The anticipation of human motion is a multidisciplinary task that involves the capability of interpreting and understanding human body dynamics. The complexity and the non-deterministic nature of human behavior makes predicting human body poses a very difficult challenge. This is particularly true in a Collaborative Robotics context, where the presence of the robot must be considered due to its strong relation with human behavior. Currently, state-of-the-art approaches are based on Deep Learning (DL) techniques, including Recurrent

[†] These authors have contributed equally to the work.

Neural Network (RNN), Graph Convolutional Network (GCN), Generative Adversarial Network (GAN), and the*attention mechanism* to model human motion [3]. These approaches are very powerful in modeling complex data, however they are very difficult to interpret and generally used as black-boxes in an end-to-end procedure [7].

Explainable Artificial Intelligence (XAI) [5] refers to the ability to understand and interpret the decisions made by Artificial Intelligence (AI) systems. It consists of providing insights about why a given model makes a specific decision, helping to understand potential biases and to identify errors inside the model. Moreover, XAI plays a crucial role in the identification of the most relevant features for making predictions, with the aim of simplifying the dataset through feature selection. Bento et al. [1] exploited an XAI approach in a classification task based on a DL model to identify and remove the undesired information from input images. Javed et al. [2] proposed a multi-sensor Human Activity Recognition (HAR) system based on the most important features selected through XAI approaches. The experiments showed performance improvements in multiple Machine Learning (ML) and DL approaches.

In this work, we introduce action semantics as additional information for the task of Human Motion Prediction (HMP) and we propose an XAI approach to identify the most relevant semantic features. Action semantics can significantly help to improve predictions in a Human-Robot Collaboration (HRC) context given that the tasks performed by the user are often known in advance. Moreover, highlighting the importance of semantic features is more intuitive and consistent for human interpretability compared to joint angle rotations.

The remainder of the paper is organized as follows. Sec. 2 describes the methodology used in this work to perform Human Motion Prediction. Sec. 3 reports on the experiments carried out to evaluate the proposed system, as well as the obtained results. Finally, Sec. 4 concludes the article.

2 Methods

Our approach consists of predicting future poses of an individual while gaining insight into the relevance of the semantic features used to describe human motion. Through this approach, we aim to create a system that is more comprehensible and thus easier to expand. The input of our system consists of both data related to the person's past poses and an embedding representing the semantic information of the action being performed. The person's pose is represented as a sequence of j rotations applied to its 3D skeletal structure, uniquely defining the spatial arrangement of the body elements. To describe a sequence of n frames, these poses can be concatenated into a one-dimensional vector of size $3 \cdot j \cdot n$. The semantic information is a sentence describing the action performed. It is encoded using a one-hot encoding with p elements. Therefore, the overall input consists of a vector of size $3 \cdot j \cdot n + p$. The predictive architecture employed for this task is a random forest [2], that enables the analysis of the importance of the different input features to the system.

The predicted sequences can be quantitatively evaluated using the Mean Angle Error (MAE) metric, that can be computed as the Euclidean distance between the predicted and the ground truth pose vectors:

$$\mathrm{MAE_t} = \frac{1}{K} \sum_{k=1}^{K} \|\hat{x}_{k,t} - x_{k,t}\|_2$$

In the provided context, $\hat{x}_{k,t}$ and $x_{k,t}$ represent vectors associated with predicted and ground truth sequences, respectively. Both vectors are defined at frame t and encompass the corresponding j rotations expressed as Euler angles following the "XYZ" sequence. Here, K represents the number of motion sequences under consideration. Once the values of $\mathrm{MAE_t}$ are determined for each frame, it is possible to compute the cumulative MAE by summing the obtained values for each frame in the predicted sequence. To assess the effectiveness of the proposed system, we compared it with the Zero-Velocity baseline. In this model, the predicted sequence consists of a repetition of the last ground truth frame. Although this approach is straightforward, it proves effective in establishing a baseline for comparisons that is commonly employed in the field of HMP [4]. Finally, the significance of a feature is assessed through a metric known as Gini importance or Mean Decrease in Impurity (MDI) [6]. It quantifies the total decrease in node impurity weighted by the probability of reaching that node and averaged over all trees of the ensemble. The probability of reaching a specific node is approximated by the proportion of samples reaching that node.

3 Experimental results

To evaluate the applied approach, experiments were performed using a subset of the Archive of Motion Capture of Surface Shapes (AMASS) dataset. AMASS is a comprehensive repository of human motion data, unifying various optical marker-based motion capture datasets within a common framework and parameterization. The framework employed is Skinned Multi-Person Linear Model (SMPL), which includes a skeletal representation comprising 24 keypoints. Each keypoint describes the relative rotation of a specific joint, except for the first, which represents the pelvis absolute rotation with respect to a global reference system. Each rotation is denoted in the axis–angle representation, which is characterized by a triplet of values. Additionally, the coordinates of the root joint concerning a global reference system are provided, describing the subject's trajectory within the environment during the action. The framerate of these recordings is 120Hz, which was reduced to 10Hz to extend the temporal horizon without significantly increasing the dimensions of input and output information. The input sequence consists of 8 frames, corresponding to the preceding $0.8s$ relative to the current moment. Then, the representation of poses during the action comprises $8 \cdot 25 \cdot 3$ values, that are the concatenation of the poses (24 joints plus translation) over the 8 frames used as input.

To obtain a description of the action in each sequence, the Bodies, Action and Behavior with English Labels (BABEL) dataset was used. BABEL represents a

comprehensive resource with language labels describing the actions performed in numerous motion-capture sequences from AMASS. To construct a one-hot encoded vector, we cataloged all words from the action labels and marked those present in the respective sequence. This approach led to the creation of a sparse vector with a length of 612 elements.

For training the random forest, we used 25 estimators with a maximum depth of 10, selecting the parameters in an empirically fashion. The results obtained on the evaluation set were then compared with the Zero-Velocity model.

The outcomes on the test set exhibit that the random forest architecture achieves superior results compared to the baseline Zero-Velocity model. Specifically, the random forest architecture attains a MAE of 16.22, whereas the Zero-Velocity baseline yields a result of 20.55. Additionally, it is noteworthy that even over extended temporal horizons, the prediction error after $1.2s$ is 1.70 for the random-forest-based model and 1.98 for the Zero-Velocity baseline.

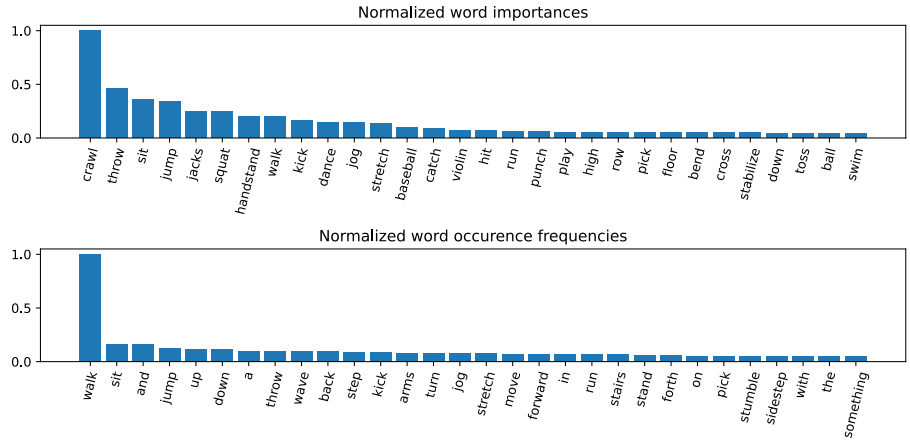

Fig. 1. Comparison between the normalized number of occurrences of words and the importance of them considered as semantic features for Human Motion Prediction

Subsequently, the importance associated with each semantic feature was computed using the Gini importance metric. The most important features were compared with the results obtained from the analysis of the word frequencies in the labels associated with the motion data. As depicted in the Fig. 1, the words that occur most frequently are often prepositions, articles, and conjunctions. However, these words did not have significant relevance for motion prediction and were not prominent in terms of importance. On the contrary, looking at the feature importance, the most crucial terms were associated with specific actions, providing valuable information about the future pose's evolution. It is also important to note that the most frequently occurring word, namely *walk*, appears in multiple descriptions due to the dataset nature. While this term is present on

the feature importance scale, its rank is lower due to its significantly reduced discriminating capability in this context.

4 Conclusions

In this paper, we proposed an architecture for HMP that incorporates semantic information about the performed actions within the sequence, demonstrating prediction performance and interpretability of the solution. This strategy enables an analysis of the importance of the utilized semantic features, paving the way for adopting XAI methods in the filed of HMP. XAI is a crucial concept providing better insights into complex problems and being potentially pivotal both in designing improved modeling architectures and in processing input data. In particular, applying XAI into the most recent DL architectures could prove particularly effective when integrated into HRC systems, which strongly benefit from the comprehension of the underlying logic guiding motion prediction. An innovative design of this nature will provide superior understanding of human intentions, enabling rich and precise interaction between human and robot agents, all while ensuring the system's overall safety.

Acknowledgment

This study was partially funded by Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004) within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership, CUP: C93C22005280001 and by the PRESENCE (anticiPatoRy bEhaviors for Safe and Effective humaN-robot CoopEration) project (BIRD221598).

References

1. Bento, V., Kohler, M., Diaz, P., Mendoza, L., Pacheco, M.A.: Improving deep learning performance by using Explainable Artificial Intelligence (XAI) approaches. Discover Artificial Intelligence 1, 1–11 (2021)
2. Javed, A.R., Khan, H.U., Alomari, M.K.B., Sarwar, M.U., Asim, M., Almadhor, A.S., Khan, M.Z.: Toward Explainable AI-empowered cognitive health assessment. Frontiers in Public Health 11, 1024195 (2023)
3. Lyu, K., Chen, H., Liu, Z., Zhang, B., Wang, R.: 3D human motion prediction: A survey. Neurocomputing 489, 345–365 (2022)
4. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2891–2900 (2017)
5. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.: Explainable AI: interpreting, explaining and visualizing deep learning. Spring er LNCS 11700 (2019)
6. Nembrini, S., König, I.R., Wright, M.N.: The revival of the Gini importance? Bioinformatics 34(21), 3711–3718 (05 2018)
7. Yang, B., Hu, L., Peng, Y., Wang, T., Fang, X., Wang, L., Fang, K.: Human Pose Prediction Using Interpretable Graph Convolutional Network for Smart Home. IEEE Transactions on Consumer Electronics (2023)