Human Motion Prediction Metrics: from Time to Frequency

Michael Vanuzzo, Marco Casarin, Mattia Guidolin,
Stefano Michieletto, and Monica Reggiani

University of Padova,
Department of Management and Engineering (DTG),
Stradella S. Nicola, 3, 36100 Vicenza, Italy
{michael.vanuzzo, marco.casarin.4}@phd.unipd.it,
{mattia.guidolin, stefano.michieletto, monica.reggiani}@unipd.it

Abstract. Collaborative robotics has the potential to revolutionize industrial applications by integrating human and robot capabilities. However, for efficient and seamless collaboration, predicting human motion is essential. This allows robots to dynamically adjust their behavior and avoid potential collisions. Despite significant progress in this field in recent years, there is still uncertainty surrounding the metrics needed for a complete and accurate evaluation of algorithm performance. Currently, the evaluation of Human Motion Prediction (HMP) techniques is based on metrics focusing exclusively on geometric aspects. This work proposes a HMP metric to evaluate the realism and naturalness of predicted human motion sequences based on their frequency spectra. Using the Human 3.6M dataset, several experiments were conducted to demonstrate the effectiveness of the proposed metric. The results showed the ability of this metric to capture insights related to the realism of the predicted motion sequences, making it a valuable complementary tool alongside existing metrics for evaluating HMP algorithms.

Keywords: Human Motion Prediction · Metrics · Collaborative Robotics

1 Introduction

Collaborative robotics focuses on combining the precision and repeatability of robots with the adaptability and problem-solving skills of humans. This is particularly effective in industrial settings, where robots can provide valuable support to human operators in dangerous and physically demanding tasks. To achieve seamless Human-Robot Collaboration, these systems need to accurately anticipate human movements and dynamically self-adapt based on the operator's behavior. Several Human Motion Prediction (HMP) algorithms have been proposed in the literature based on modeling the human body through a skeletal representation [4–6, 8].

The current metrics used for evaluating human movements only analyze geometric aspects of the predicted movements. However, these metrics fail to consider the realism of the movements being predicted as they focus only on the rotation angles or 3D position of the body skeleton. It has been observed [2] that the

frequency spectra of motion sequences are strongly correlated with the realism of human movements. In this work, we propose a novel metric for HMP based on the analysis of the motion frequency spectra of the predicted human movements. The effectiveness of the proposed metric has been addressed through several experiments conducted using the Human 3.6M (H36M) dataset [3], comparing three state-of-the-art HMP models [4,5,8], as well as the Zero-Velocity (ZeroVel) baseline [6].

2 Methodology

This section describes the metrics commonly used for HMP and proposes a novel metric based on the frequency spectrum of the motion sequences. In the context of HMP, a sequence is a time series of human poses, each defined by a set of features that fully describe the skeleton's configuration.

2.1 Geometric Accuracy Metrics

These metrics aim to measure the difference between each predicted frame and the ground truth within the K sequences of the test set, each spanning T frames.

Mean Angle Error (MAE) The MAE, also known as Euler Error, represents the standard metric to evaluate HMP algorithms [1, 4–6], and its definition is:

$$\text{MAE} = \frac{1}{K \cdot T} \sum_{k=1}^{K} \sum_{t=1}^{T} \|\hat{x}_{k,t} - x_{k,t}\|_2 \tag{1}$$

Here, $\hat{x}_{k,t}$ and $x_{k,t}$ denote the predicted pose and the ground truth,respectively, for frame t in sequence k. Each pose $x_{k,t}$ is represented by a vector containing $3 \cdot J$ elements, corresponding to the 3 Euler angles that describe the relative rotation of each of the J joints with respect to its parent joint.

Mean Per Joint Position Error (MPJPE) The MPJPE is a widely used metric for both pose estimation and prediction [5, 7]. It inherently considers both the distance between joints, i.e., the length of the links defined in the skeleton, and the accumulation of errors along the kinematic chain. This metric is mathematically defined as follows:

$$\text{MPJPE} = \frac{1}{K \cdot T \cdot J} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{j=1}^{J} \|\hat{p}_{k,t,j} - p_{k,t,j}\|_2 \tag{2}$$

Here, $\hat{p}_{k,t,j}$ and $p_{k,t,j}$ represent the predicted 3D position of joint j and its corresponding ground truth for frame t in sequence k.

2.2 Frequency Spectrum Similarity Metric

When evaluating a HMP algorithm, it is crucial to consider the realism of the generated sequences. However, a prediction yielding high accuracy in geometric metrics may fail to account for unnatural movements that present sharp discontinuities. Therefore, this study focuses on evaluating the realism of the generated motion sequences based on the analysis of their frequency spectrum.

Given a motion sequence $p_k(t)$ described by joint positions over time, its power spectral density $P_{k,norm}(f)$, normalised across all joints, can be computed using the Fourier transform. Then, different power spectral densities can be compared using the Wasserstein Distance (WD), a distance function that can be used between probability distributions. The proposed Power Spectral Densities Similarity (PSDS) metric, defined as the WD of order 1, can be computed as follows:

$$\mathrm{PSDS}\left(P_{k_1,norm}(f), P_{k_2,norm}(f)\right) = \int \left|F_{k_1}(x) - F_{k_2}(x)\right| \mathrm{d}x \tag{3}$$

Here, $F_{k_i}(x)$ is the cumulative distribution function of $P_{k,norm}(f)$, similar to a probability distribution describing the probability that human joint movements will excite specific frequencies. A higher value is achieved when the power distribution shifts towards either low or high frequencies, a behavior in contrast to typical human movement.

3 Experiments and Results

To assess the effectiveness of the proposed PSDS metric, we conducted multiple experiments evaluating different HMP algorithms on the metrics described in section 2. The study was based on the H36M dataset [3] and on three state-of-the-art Deep Learning (DL) models: History Repeats Itself (HRI) [5], Dynamic Multiscale Graph Neural Networks (DMGNN) [4], and Position-Velocity Recurrent Encoder-Decoder (PVRED) [8]. HRI employs an attention mechanism to identify similarities in past and current action sequences. DMGNN implements Graph Convolutional Network to discern relationships among skeleton joints at various abstraction levels. PVRED is based on a Recurrent Neural Network with Gated Recurrent Units to capture temporal relationships by considering both positional and velocity information. Additionally, the results include scores from the ZeroVel model proposed in [6], in which all the prediction frames are identical to the last input frame. Despite its simplicity, this model is commonly used as a valuable baseline. Notably, many algorithms performed worse than this model [6]. This highlights the challenges of accurately predicting future human poses, but also emphasizes the limits of current evaluation metrics.

Fig. 1 presents the results of the three algorithms in terms of MAE and MPJPE metrics. All models perform better than the ZeroVel baseline, with the HRI model showing superior accuracy. However, it is important to note that the improvement relative to the ZeroVel model is not very pronounced.

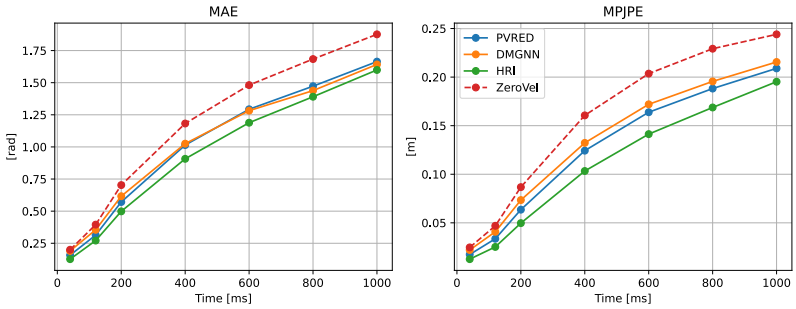

Fig. 1. MAE and MPJPE values during the first second of anticipation for the three models and the baseline, represented by the ZeroVel model.

Alongside the previous results, the outcomes obtained with the proposed novel metric are presented. Each point in Fig. 2 shows the PSDS metric computed on a 1 s sliding window, thus achieving a frequency resolution of 1 Hz. Values are reported in logarithmic scale over a 30 s span.

The results show how the three models outperform the ZeroVel model, highlighting a greater capability of generating natural movements. Furthermore, it is observable that the HRI model, despite providing the best accuracy with MAE and MPJPE, turns out to be the least effective in generating movements with frequencies that resemble natural human motion. The ZeroVel model, generating a constant pose throughout the prediction timeframe, leads to high PSDS values as the only spectral component generated is at 0 Hz. Therefore, this result highlights that ZeroVel predictions are highly implausible. Furthermore, the effectiveness of the PSDS metric is confirmed by the *train-test* value, which is computed using motion sequences from the test set. Given that the latter consists of real human motion recordings, the *train-test* represents the lowest achievable error.

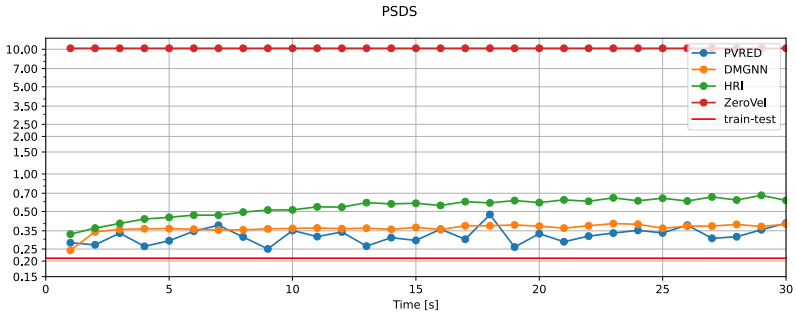

Fig. 2. PSDS values during 30 s of anticipation for the three models, the baseline (ZeroVel model), and the lowest achievable error (train-test).

4 Conclusions

This paper presents PSDS, a novel metric that emphasizes the significance of frequency analysis in determining the quality of generated movement sequences. While commonly used metrics focus on assessing geometric accuracy, PSDS uniquely evaluates the realism of predicted movements. Consequently, it introduces key information complementary to the existing metrics, providing essential insights for developing innovative prediction algorithms. This is particularly crucial in Human-Robot Collaboration field, where predictions must be geometrically accurate and also ensure realism and naturalness. By introducing a novel metric that evaluates these aspects, this paper contributes to a more comprehensive evaluation framework, paving the way for the development of enhanced predictive models.

Acknowledgments. This study was 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. Aksan, E., Kaufmann, M., Cao, P., Hilliges, O.: A spatio-temporal transformer for 3d human motion prediction. In: 2021 International Conference on 3D Vision (3DV). pp. 565–574. IEEE (2021)
2. Gopalakrishnan, A., Mali, A., Kifer, D., Giles, L., Ororbia, A.G.: A neural temporal model for human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12116–12125 (2019)
3. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(7), 1325–1339 (2014)
4. Li, M., Chen, S., Zhao, Y., Zhang, Y., Wang, Y., Tian, Q.: Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 214–223 (2020)
5. Mao, W., Liu, M., Salzmann, M.: History repeats itself: Human motion prediction via motion attention. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 474–489. Springer International Publishing, Cham (2020)
6. 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)
7. Véges, M., Lőrincz, András: Absolute human pose estimation with depth prediction network. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2019)
8. Wang, H., Dong, J., Cheng, B., Feng, J.: Pvred: A position-velocity recurrent encoder-decoder for human motion prediction. IEEE Transactions on Image Processing **30**, 6096–6106 (2021)