Enhancing GAN-Based Motion Data Augmentation through Dynamic Time Warping Distance Filtering

Junwon Yoon[1], Hyun-Joon Chung[1]*, Jeon-Seong Kang[1], Jung-Jun Kim[1], Kwang-Woo Jeon[1], SeungWoo Kim[1], Myounghoon Shim[2] and Jae-Kwan Ryu[2]

[1]AI Robotics R&D Division, Korea Institute of Robotics & Technology Convergence, Seoul 06372, Republic of Korea

[2]Unmanned/Intelligent Robotic Systems R&D, LIG Nex1, Seongnam 13488, Republic of Korea

Email: {jyoon, hjchung, kjs2605, jjkim, jeonkw, ksw6035}@kiro.re.kr, {myounghoon.shim, jaekwan.ryu}@lignex1.com

***Abstract*—Motion capture data is crucial but creating a large dataset can be challenging due to complexities in acquisition. Generative Adversarial Network (GAN)-based motion data augmentation offers a potential solution to this issue. However, GANs often struggle with learning from limited data, resulting in poor quality output. In this study, we propose a Dynamic Time Warping (DTW) filtering method that filters out generated data significantly deviating from real-world examples. Through this approach, we have achieved an improvement in the fidelity of the generated data, even with dataset size constraints, as evidenced by an increase in action recognition accuracy.**

***Keywords*—data augmentation, generative adversarial network, dynamic time warping, motion capture**

I. Introduction

As artificial intelligence (AI) progresses, the importance of data for training AI models has been increasingly emphasized. This is particularly significant in fields where data acquisition presents considerable challenges, leading to an interest in AI generative models for data augmentation. One such challenging area is motion capture, which plays a pivotal role in wearable robotics design [1], gaming, animation [2], sports science [3], ergonomics [4], and training AI models related to human motion [5], [6]. However, collecting large datasets of motion capture is challenging due to cost constraints and complexities involved in the acquisition process [7], [8].

Generative Adversarial Networks (GANs) are renowned for their successful application in image generation research [9]–[11], a technique that can be effectively applied to motion generation tasks. Although GANs have demonstrated promising results with extensive open datasets for motion generation [12], acquiring large-scale motion datasets specific to certain tasks or users is prohibitively expensive in real-world settings. The nature of GANs is such that utilizing small amounts of data often leads to lower-quality results [13]. Consequently, there is an evident demand for methods capable of enhancing the quality of motion capture data generated by GANs, even when the available data is limited.

Dynamic Time Warping (DTW) is an algorithm that has been widely used in comparing time-series data like speech recognition signals [14] or wearable sensor data [15]. Its strength lies in the ability to compare sequences of different lengths while taking shifts during similarity evaluation into account. Although DTW has been applied successfully onto motion capture data primarily within classification tasks [16], [17], its potential towards improving generation tasks remains largely unexplored.

This study introduces a novel DTW filtering method designed to enhance the fidelity of augmented motion data effectively. Although the DTW filtering method was applied to GAN in this study, it is versatile and can be utilized with any other generative models without requiring significant architectural changes or being constrained by data size. In addition, we explored the impact of the filter threshold on the fidelity and diversity of the generated motion data by comparing action recognition accuracy and Fréchet Inception Distance (FID) across various filter threshold values.

II. Methods

A. Dataset

In our experiments, we utilized the H3.6M dataset [18], a popular open resource in motion generation and prediction research. This dataset includes 15 human activities such as walking, sitting, and discussing, enacted by 7 actors with each activity performed twice. Adapting the data processing methods from previous studies, we converted the skeleton joint angle data into the exponential map representation [19], [20] and removed the global rotation and translation of the root joint, as well as any joints with constant angles [21], [22]. For the purposes of training the GAN and evaluating the generated motion data, we selected 5 activities (directions, discussion, greeting, sitting, walking) commonly observed in daily life. Each activity contained 14 sequences, a relatively small number for conventional GAN training.

B. Conversion of Skeleton Sequence to Pseudo-images

We adopted a transformation approach from joint angle motion data to a 2D image format that is suitable for GANs to learn and generate. This method is referred to as pseudo-image representation in previous studies [12], [23]. Each component of the exponential map vector, which represents 3D joint angles, was converted into the R, G, and B channels of the 2D image. In this pseudo-image format, time frame and joint types were represented as rows and columns, respectively. This arrangement allowed the convolutional filters used in GANs

*The correspondence should be addressed to hjchung@kiro.re.kr

to capture both temporal and spatial features of the skeleton sequence data. The details of this transformation process are provided in Fig. 1.

C. GAN-based Motion Generation

We used Wasserstein GAN with gradient penalty (WGAN-GP) [11] for generating motion data. Our model was trained over 50,000 epochs with a batch size of 14. The noise dimension was 128 and the shape of the generated images was defined as (256, 256, 3). We employed the Adam optimizer with a learning rate of 0.0001 for both generator and discriminator components.

D. Dynamic Time Warping (DTW) Distance Filtering

To address the issue of low-fidelity data generation when training GANs with a limited amount of data, we proposed a novel approach: incorporating DTW distance filtering into the data generation process.

DTW excels at comparing two time-series data of varying lengths or time shifts. Rather than merely calculating the difference between two values at the same time frame, DTW considers values surrounding the time step to minimize the distance and find an optimal path for comparing the two time-series data.

Essentially, goal of the DTW algorithm is to find optimal warping path which minimize cost function calculated from local cost matrix associated with the optimal warping path [24]. For two time-series data $X = (x_1, x_2, ..., x_N)$ and $Y = (y_1, y_2, ..., y_M)$, local cost matrix C_l is:

$$C_l \in \mathbb{R}^{N \times M} : c_{i,j} = \|x_i - y_j\|, i \in [1:N], j \in [1:M] \quad (1)$$

For a given warping path $p = (p_1, p_2, ..., p_K)$ with $p_l = (n_i, m_j) \in [1:N] \times [1:M]$ for $l \in [1:K]$, warping path must satisfy following criteria:

- Boundary condition

$$p_1 = (1,1),\ p_K = (N, M) \quad (2)$$

- Monotonicity condition

$$n_1 \leq n_2 \leq ... \leq n_K,\ m_1 \leq m_2 \leq ... \leq m_K \quad (3)$$

- Step size condition

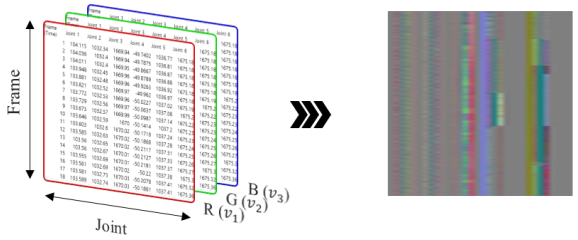
$$p_{l+1} - p_l \in \{(1,1), (1,0), (0,1)\} \quad (4)$$

cost function of the warping path with respect to the local cost matrix $c_p$is:

$$c_p(X, Y) = \sum_{l=1}^{K} c(x_{n_l}, y_{m_l}) \quad (5)$$

If we denote the optimal warping path as P^* which has minimal cost among the warping paths, the DTW distance function will be

$$DTW(X, Y) = c_{p^*}(X, Y) = \min\{c_p(X, Y),\ p \in P^{N \times M}\} \quad (6)$$



(a) Skeleton sequence motion data (b) Pseudo-image representation

Fig. 1. Transformation from (a) skeleton sequence motion data to (b) pseudo-image representation.

Due to the inefficiency of comparing every possible cost function for each warping path, DTW utilize dynamics programming-based algorithm. An accumulated cost matrix or global cost matrix D is defined as follows:

- First row

$$D(1, j) = \sum_{k=1}^{j} c(x_1, y_k),\ j \in [1, M] \quad (7)$$

- First column

$$D(i, 1) = \sum_{k=1}^{i} c(x_k, y_1),\ i \in [1, N] \quad (8)$$

- All other elements

$$D(i, j) = \min\{D(i-1, j-1),\ D(i-1, j),\ D(i, j-1)\} + c(x_i, y_j),\ i \in [2, N],\ j \in [2, M] \quad (9)$$

Once accumulated cost matrix is calculated, optimal warping path could be found by backtracking minimal values among adjacent from the end point.

Our method is designed to enhance the fidelity of generated data by calculating the dynamic time warping distance between each generated and real data instance. This measure provides an indication of their similarity in terms of temporal dynamics. If the DTW distance exceeds a certain threshold - indicating substantial dissimilarity - we discard that particular instance and prompt the generator to create new data. This filtering loop can be integrated with any generative models, as illustrated in Fig. 2, allowing for improved control over the fidelity and diversity of generated motion capture data. In this experiment, we examined motion data generated both with and without the DTW filtering method, using filter thresholds of 1.2, 1.6, 2.0, and 2.4 radian. We conducted both qualitative and quantitative analyses.

E. Action recognition accuracy for fidelity evaluation

To quantitatively evaluate the fidelity of the generated motion data, we trained an action recognition model and verified its classification accuracy with respect to the generated data. Given the ease of training and superior performance of 1D convolution-based models, even with small amounts of data [25], we trained a 1D convolution-based action recognition

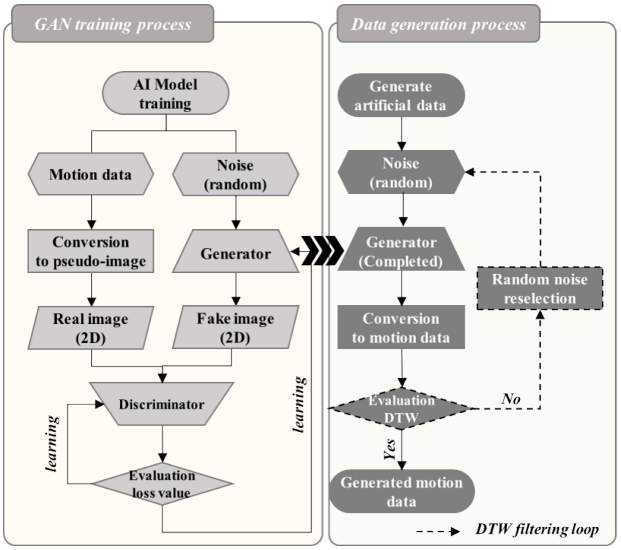


Fig. 2. The DTW distance filtering process.

model on five types of activity data. As the sequence length varied in each motion data, we conducted preprocessing to standardize all sequences to 600 timesteps through interpolation, facilitating learning and classification. To ascertain the accuracy of the action recognition model on real data, we performed 5-fold cross-validation. As a result, the model demonstrated an average classification accuracy of 88.57% and a standard deviation of 3.50% across the five actions. We then had the trained action recognition model classify data generated for each activity, 100 instances each, based on whether the DTW filter method was applied. This allowed us to quantify the fidelity based on how similar the features of the generated data were to those of the actual data used for training.

F. Fréchet Inception Distance (FID)

FID is one of the most widely used metrics for evaluating generative models, encompassing both fidelity and diversity [26]. We calculated the FID values for the generated pseudo-images. By comparing the FID values before and after the application of the DTW filtering method, as well as across different filter threshold values, we were able to analyze the impact of the proposed method on the fidelity and diversity of the generated motion data.

III. Results

A. Qualitative evaluation of generated motions

We conducted a qualitative evaluation by visualizing the skeleton motion data for the walking activity. The visualization results can be seen in Fig. 3. When compared with real data, the motion data generated without DTW filtering exhibits lower fidelity, as evidenced by skewed bodily postures. The generated motion is somewhat awkward to perceive as a human's walking motion.

In contrast, the motion data generated with the application of DTW filtering demonstrates characteristics of human walking motion that are relatively similar to the actual data. As we apply stricter standards to fidelity, i.e., lower filter threshold values, the generated actions visually resemble the real data more closely. However, when comparing the visual results of data generated with higher filter threshold values, we observe a reduction in diversity and the generation of more similar actions.

B. The Impact of DTW Filtering on the fidelity and diversity of Generated Data

To evaluate the effectiveness of the proposed DTW filtering method, we comprehensively assessed fidelity through action recognition model accuracy and both fidelity and diversity through FID values, each under various filtering threshold values (Table I, Table II).

The action data generated without applying the DTW method had an accuracy of 81.20%, lower than the 5-fold cross-validation result (88.57%) of the action recognition model. This suggests that the data generated without applying the DTW method lacked sufficient features for the action recognition model to identify each action, indicating insufficient fidelity. This aligns with the results of the quantitative analysis, where the visualized actions were hard to identify as walking motions by human observers. The action data generated with the application of the DTW filtering method showed over 90% accuracy with the trained action recognition model. Moreover, lower filtering threshold values resulted in more accurate action classification, implying a closer resemblance of features between the generated and real data.

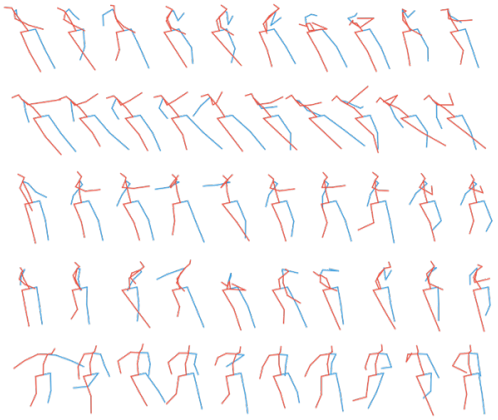
Contrary to the improvement in action recognition accuracy with lower DTW filter thresholds, the FID values increased. This could be inferred as the filtering method limiting the diversity of the generated data. Therefore, indiscriminately applying lower filter threshold values to improve the overall quality of the generated data may reduce diversity, suggesting the need for careful adjustment.

These results demonstrate that integrating DTW distance filtering into GAN models can effectively enhance their ability to generate high-fidelity motion sequences even under conditions where training data are scarce while highlighting need for balance between similarity enforcement and maintaining sample diversity.

TABLE I
Action recognition accuracy of generated data with and without DTW filtering method

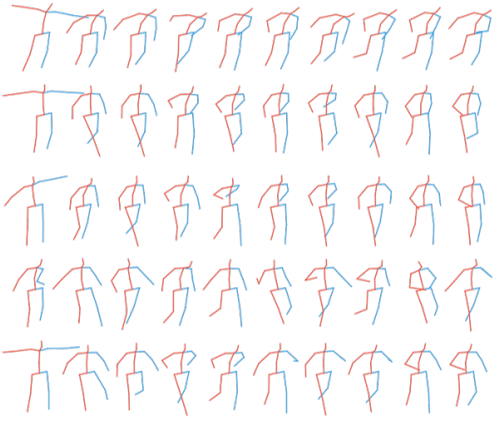
	DTW filtering method				
	Not applied	Threshold 2.4	Threshold 2.0	Threshold 1.6	Threshold 1.2
Action recognition accuracy	81.20%	93.80%	96.60%	96.60%	97.40%

(a) Real data

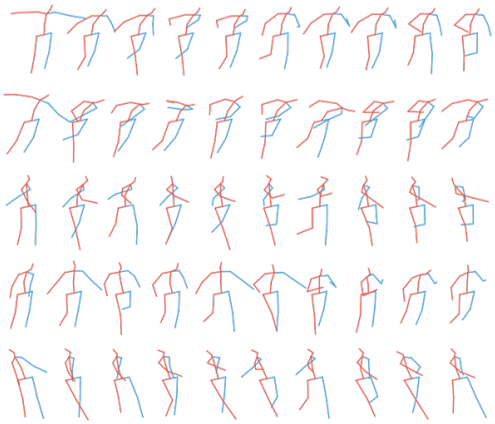


(b) Genetrated data (DTW X)

(c) Generated data (DTW filtered, threshold = 1.2)



(d) Generated data (DTW filtered, threshold = 1.6)



(e) Generated data (DTW filtered, threshold = 2.0)

(f) Generated data (DTW filtered, threshold = 2.4)

Fig. 3. Visualization of real and generated skeleton data (walking motion), with and without the DTW filtering method, across different filter threshold values: (a) Real motion data, (b) Generated data without the DTW filter, and generated data with DTW filter thresholds of (c) 1.2, (d) 1.6, (e) 2.0, and (f) 2.4 radians.

TABLE II
FRÉCHET INCEPTION DISTANCE WITH & WITHOUT DTW FILTER

	without DTW filtering	with DTW filtering (Threshold value)			
		(2.4)	(2.0)	(1.6)	(1.2)
FID	35.873	37.867	37.947	40.244	42.949

IV. CONCLUSION

Our exploration of DTW distance filtering as a technique to improve fidelity of the generated data with limited data has yielded promising results. We demonstrated that incorporating this method into the process of generating motion datasets can significantly enhance the fidelity of motion data even with small amount of data.

Considering the importance of motion capture data and the difficulties in acquiring it, which often limits the availability of large-scale data, the proposed method appears promising. It can contribute to research on motion capture data augmentation by improving the fidelity of GAN models that may struggle to secure sufficient fidelity when trained with limited data.

However, we also noted an important caveat: over-restricting diversity within generated samples can lead to higher FID scores. This suggests a need for balance between enforcing similarity to real-world examples and maintaining sufficient diversity within generated samples.

In conclusion, while our results confirm that DTW distance filtering can improve both the fidelity of generated data when training on limited datasets, they also underscore the importance of careful threshold selection to avoid compromising sample diversity. Future work may explore more sophisticated approaches for dynamic threshold adjustment that could potentially yield even better results.

ACKNOWLEDGMENT

This research was supported by the Korea Research Institute for defense Technology planning and advancement (KRIT) funded by the Defense Acquisition Program Administration (DAPA) of the Korea government since 2021. (No. KRIT-CT-21-039-00, Development of Techniques of Designing and Operating a Powered Full-Body Exoskeleton with Bulletproof Armor and Military Equipments(Contribution rate : 50%)). This research was also financially supported by the Institute of Civil Military Technology Cooperation funded by the Defense Acquisition Program Administration and Ministry of Trade, Industry and Energy of Korean government under grant No. 19-CM-GU-01

REFERENCES

[1] K.-W. Jeon, H.-J. Chung, E.-J. Jung, J.-S. Kang, S.-E. Son, and H. Yi, "Development of shoulder muscle-assistive wearable device for work in unstructured postures," *Machines*, vol. 11, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2075-1702/11/2/258

[2] S. Sharma, S. Verma, M. Kumar, and L. Sharma, "Use of motion capture in 3d animation: Motion capture systems, challenges, and recent trends," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 289–294.

[3] S. Noiumkar and S. Tirakoat, "Use of optical motion capture in sports science: A case study of golf swing," in *2013 International Conference on Informatics and Creative Multimedia*, 2013, pp. 310–313.

[4] T. Petrosyan, A. Dunoyan, and H. Mkrtchyan, "Application of motion capture systems in ergonomic analysis," *Armenian Journal of Special Education*, vol. 4, no. 2, pp. 107–117, 2020.

[5] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," 2017.

[6] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3d human motion prediction: A survey," *Neurocomputing*, vol. 489, pp. 345–365, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222002077

[7] M. O'Reilly, B. Caulfield, T. Ward, W. Johnston, and C. Doherty, "Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review," *Sports Medicine*, vol. 48, pp. 1221–1246, 2018.

[8] T. Maeda and N. Ukita, "Motionaug: Augmentation with physical correction for human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6427–6436.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017.

[12] W. Xi, G. Devineau, F. Moutarde, and J. Yang, "Generative model for skeletal human movements based on conditional dc-gan applied to pseudo-images," *Algorithms*, vol. 13, no. 12, 2020. [Online]. Available: https://www.mdpi.com/1999-4893/13/12/319

[13] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring gans: generating images from limited data," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[14] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[15] L. Guo, J. Chen, Q. Zheng, J. Wang, J. Bian, X. Wang, and R. Ouyang, "Wearable motion recognition system based on dynamic time warping," *IOP Conference Series: Earth and Environmental Science*, vol. 234, no. 1, p. 012087, feb 2019. [Online]. Available: https://dx.doi.org/10.1088/1755-1315/234/1/012087

[16] K. Adistambha, C. H. Ritz, and I. S. Burnett, "Motion classification using dynamic time warping," in *2008 IEEE 10th Workshop on Multimedia Signal Processing*. IEEE, 2008, pp. 622–627.

[17] A. Switonski, H. Josinski, and K. Wojciechowski, "Dynamic time warping in classification and selection of motion capture data," *Multidimensional Systems and Signal Processing*, vol. 30, pp. 1437–1468, 2019.

[18] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[19] G. W. Taylor, G. E. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19. MIT Press, 2006. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2006/file/1091660f3dff84fd648efe31391c5524-Paper.pdf

[20] F. S. Grassia, "Practical parameterization of rotations using the exponential map," *Journal of Graphics Tools*, vol. 3, no. 3, pp. 29–48, 1998. [Online]. Available: https://doi.org/10.1080/10867651.1998.10487493

[21] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[22] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," 2020.

[23] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4570–4579.

[24] P. Senin, "Dynamic time warping algorithm review," *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, vol. 855, no. 1-23, p. 40, 2008.

[25] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888327020307846

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018.