Microwave/millimeter wave sensors

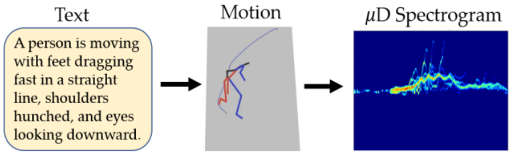
Text2Doppler: Generating Radar Micro-Doppler Signatures for Human Activity Recognition via Textual Descriptions

Yi Zhou[1,2,3], Miguel López-Benítez[3,4], Limin Yu[5] and Yutao Yue[6,1,2]

[1] Institute of Deep Perception Technology, JITRI, Wuxi 214000, China
[2] XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, Chin
[3] Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3GJ, UK
[4] ARIES Research Centre, Antonio de Nebrija University, 28040 Madrid, Spain
[5] School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[6] Thrust of Artificial Intelligence and Thrust of Intelligent Transportation, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China

Manuscript received June 7, 2017; revised June 21, 2017; accepted July 6, 2017. Date of publication July 12, 2017; date of current version July 12, 2017.

Abstract—Radar-based Human Activity Recognition (HAR) is popular because of its privacy and contactless sensing capabilities. However, a major challenge in this area is the lack of large and diverse datasets. In response, we present a novel framework that uses generative models to transform textual descriptions into motion data, thereby simulating radar signals. This approach significantly enriches the realism and diversity of the dataset, especially for infrequent but critical activities such as falls and abnormal walking. Textual descriptions capture the semantic complexity of human actions, thereby improving intra-class diversity. Our framework scales the data generation process by using a lightweight physics-based simulator and improves diversity by controlling gait variation, multi-viewpoint adaptation and background noise modelling. The experiments show that data diversity is a critical factor for fair model comparisons, and that the simulated data can effectively improve performance through sim-to-real transfer learning.


Text
A person is moving with feet dragging fast in a straight line, shoulders hunched, and eyes looking downward.
Motion
μD Spectrogram

Index Terms—radar simulation, text-driven motion synthesis, human activity recognition

I. INTRODUCTION

Radar sensing, known for its privacy and contactless sensing capabilities, has become a growing area of interest in Human Activity Recognition (HAR). The sensing pipeline can be divided into two paradigms: one based on high resolution point clouds [1], the other using Doppler velocity patterns [2]. Radar point clouds are typically sparse due to the low angular resolution of radar and the weak reflection from the human body. Improving spatial resolution typically requires a larger aperture, which increases sensor size, power consumption and cost. A promising alternative approach is to use motion signatures for classification. Since frequency modulated continuous wave (FMCW) radar can measure Doppler velocity at high resolution, the micro-Doppler distribution of non-rigid body motion over time can serve as a distinctive motion signature for human activity.

A number of recent studies [2], [3] have harnessed the power of deep learning to classify radar micro-Doppler spectrograms. The prerequisite for deep learning is a high quality dataset. However, radar datasets are often orders of magnitude smaller than vision datasets and with lack diversity. Radar data collection typically involves volunteers performing a predefined set of activities. While this may be sufficient for simple activities such as walking, it is insufficient for less common activities such as falling or abnormal gait. These instructional activities often appear unnatural and lack contextual richness and diversity, leading to issues with corruption robustness [4].

To address the challenges of data collection, several studies have focused on high-fidelity simulation for radar-based HAR tasks. This process involves two main steps. The first is the acquisition and extraction of skeletal information of the human pose, which can be accurately captured using marker-based motion capture systems [5] or estimated from videos [6]–[9]. The second step involves radar simulation and signal processing to produce the micro-Doppler spectrogram. In addition to the simulation, generative models, like GAN [10], are also used to synthetic radar samples, but with the problem of mode collapse and kinematic inconsistency. Although these methods expand the potential sources of human activity data, they still fall short in capturing critical activities such as falling, which are infrequent in daily life and therefore underrepresented in data collected by other modalities.

To overcome these difficulties, our work combines text-to-motion generation and physics-based simulation to generate a diverse synthetic dataset. Specifically, we exploit knowledge from large language models (LLMs) and the power of generative models to transform textual descriptions into human motion. We then use physical simulation to generate radar micro-Doppler spectrograms from the skelekton data.

The remainder of this letter is organised as follows. Section II presents the proposed Text2Doppler pipeline. Section III first discusses the necessity of using the dataset with diversity to compare models

Corresponding authors: Limin Yu and Yutao Yue (e-mail: limin.yu@xjtlu.edu.cn, ytyue@ustc.edu).
Associate Editor:
Digital Object Identifier 10.1109/LSENS.2023.0000000

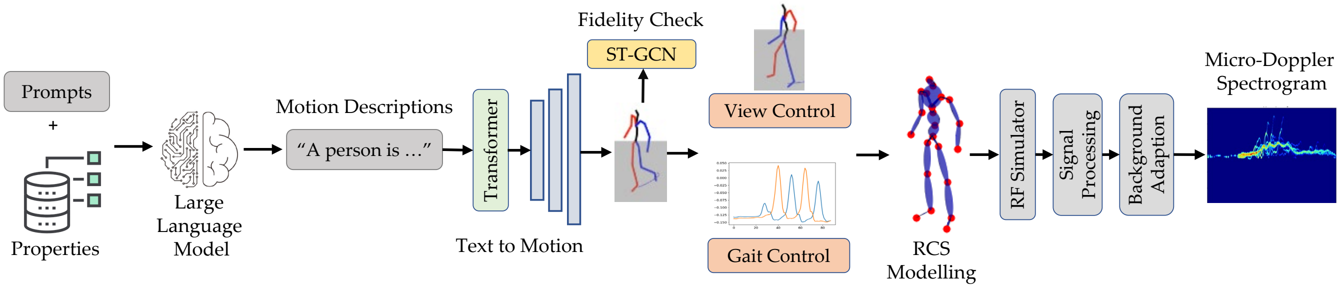


Fig. 1. Text2Doppler simulation pipeline

fairly, then introduces sim-to-real transfer learning to effectively use the simulated data even without high fidelity. Finally, Section IV concludes the paper and summarizes the whole work.

II. TEXT2DOPPLER SIMULATION PIPELINE

A. Text to Motion Generation

Our simulation pipeline, as shown in Fig. 1, starts with the automatic generation of motion descriptions for daily indoor human activities using LLMs. To avoid oversimplified motion descriptions, we provide a database of motion properties in the prompts as contextual information. These properties include elements such as action, direction, body part involved, objects interacted with, and a descriptive narrative. The inclusion of these fine-grained properties significantly enriches the motion descriptions, resulting in more detailed and diverse motion simulations.

After obtaining the motion descriptions, we directly use the pre-trained model MoMask [11] for text-to-motion generation. MoMask uses a generative masked modelling framework to convert textual descriptions into 3D human motions. When inspecting the generated results, we observe that the quality of the generated motion sequences varies across categories, possibly due to the distributional shift between the training dataset and our cases. For example, the generation of falling motions often fails, while the quality of walking motions is generally good. Therefore, it is necessary to add a fidelity check module to discriminate the unrealistic examples. To achieve this, we use the Spatial Temporal Graph Convolutional Networks (ST-GCN) [12] for the classification of the skeleton sequences. The results show that the model achieves 90.02% accuracy and 86.17% recall, which is acceptable for fidelity checking and anomaly filtering.

TABLE 1. Database of Properties

Properties	Items
Action	walk, fall, sit, bend, lie down, stand up, climb, stretch, squat, dance, run, jump, turn
Direction	left, right, clockwise, counterclockwise, anticlockwise, forward, back, up, down, straight
Body Part	arm, foot, feet, hand, leg, waist, knee
Object	stair, chair, floor, ball, handrail
Description	slowly, carefully, fast, careful, slow, quick, happily, angry

B. View Control and Gait Speed Control

The different motion descriptions for a given activity class address the semantic diversity of real-world data. In addition, we address some physics-based diversity, including the different viewing angles and different walking speeds for a given motion sequence. In the traditional pipeline, such diversity is approximated at the received data stage by data augmentation [4]. In comparison, introducing diversity at the skeleton stage offers significant advantages by maintaining physical fidelity and interpretability.

Regarding the change in viewpoint, it is important to note that, as radar primarily measures radial velocity, different viewpoints can lead to variations in the motion pattern. To account for viewpoint changes, we first determine the main direction of the motion trajectory using Principal Component Analysis (PCA). The first two principal components represent the direction of motion. We then calculate the new viewpoint location based on the desired angular difference, ensuring that the distance from the start point of the motion remains unchanged. For gait speed control, we start by retrieving the foot trajectories of both feet. By identifying the peaks, we can divide the whole trajectory into distinct segments. These segments are then interpolated to either speed up or slow down the motion.

C. Radar Data Simulation

The generated motion data is stored as BVH (BioVision Hierarchy format) files. These files contain the motion data, which is a sequence of frames, each of which contains the position and orientation of each body segment in the skeleton. In line with other gesture simulators [5], [13], we approximate the body segments as ellipsoids. Assume that each body part segment can serve as an ellipsoid parameterized by two semi-axes of equal length a and a principal semi-axis of length c. The radar cross section (RCS) σ of the i-th body segment can then be approximated by the following equation [13]:

$$\sigma_i = \frac{\pi a_i^4 c_i^2}{\left(a_i^2 \sin^2(\psi_i) + c_i^2 \cos^2(\psi_i)\right)^2} \tag{1}$$

where ψ_i describes the aspect angle of the principal axis.

The intermediate frequency (IF) signal for each body segment can then be calculated, and the total radar response is obtained by superimposing all the responses. Suppose we use a FMCW radar with carrier frequency f_c, bandwidth B, chirp repetition time T_{chirp}, chirp duration T_c and speed of light c. The IF signal for the l-th chirp can be modelled as

$$\begin{aligned} s_{\mathrm{IF},l}(t) = A \exp\Bigg(2\pi \mathrm{j}\Bigg[\frac{2f_c R}{c} - \left(\frac{2f_c v_{rad}}{c}\right) \cdot l \cdot T_{chirp} \\ + \left(\frac{2BR}{cT_c}\right) \cdot t\Bigg]\Bigg) \end{aligned} \tag{2}$$

For each body segment, the range R and the radial velocity v_{rad} can be calculated from the locations of the scattering points and the radar. The amplitude A can be calculated from the radar equation as

$$A = \sqrt{P_r G_{IF}} = \sqrt{P_t G_t G_r \frac{\lambda^2 \sigma A_e}{4\pi R^4} G_{IF}} \tag{3}$$

where P_r is the received power, P_t is the transmitted power, G_t is the gain of the transmitting antenna, G_r is the gain of the receiving antenna, λ is the wavelength of the radar signal, σ is the RCS of the target, A_e is the effective aperture of the receiving antenna and G_{IF} is the gain of the IF amplifier.

Once the IF signal has been obtained, the next step is to perform analog-to-digital (ADC) sampling for each received channel. The data acquired by the ADC is then organized into a three-dimensional tensor. A range Fast Fourier Transform (FFT) is then performed along the fast time dimension to obtain the range profile, followed by a Moving Target Indication (MTI) filter to remove static clutter. Finally, a Short-Time Fourier Transform (STFT) is applied to extract the micro-Doppler spectrogram.

D. Background Noise Generation

The above simulations are noise-free and only take into account the reflections from the human. Therefore, we need to add background noise to the generated data to improve the generalisation. The background noise for the micro Doppler spectrogram can be introduced by system noise, sensor noise and environmental factors. Instead of explicitly modelling these complex noises, we learn the background noise in a supervised manner. Specifically, we use real data collected in empty rooms to train a VQ-VAE [14] to learn the background reflections and add randomly generated background noise to the Doppler spectrogram during the data simulation process.

III. EXPERIMENTS AND RESULTS

A. Dataset Specification

We design 8 classes of daily activities for the dataset, including normal walking, abnormal walking, running, falling, sitting, bending, jumping and dancing. For each, we generate 200 motion descriptions per class and 20 ten-second data samples per description. For each data, the viewing angle is randomly selected within 15 to 15 degrees. We then use the fine-tuned ST-GCN model to remove the ambiguous data, resulting in a filtered high quality dataset with a size of 19,165 samples. We also generate a simple version of the dataset with the same size and data distribution, but less motion diversity, by generating the data samples from the same motion descriptions.

Due to the randomness in the generation, the motion data generated for each description are diverse with respect to length, moving direction and detailed motion content. In particular, for the safety critical activities, we design diverse motion descriptions to account for the intra-class variance of these activities. For example, for falling, we consider three types of falls, including slipping, tripping and collapsing, as shown in Fig. 2. For abnormal walking, we describe different walking styles, such as Parkinson's gait, waddling gait, fixation gait, ataxic gait, scissor gait and stepping gait.

For the simulation we model a 77 Ghz radar with the same configuration as described in [3]. For the signal processing, the FFT points and window size are set to 256. A significant overlap of 128 samples between windows is implemented to increase the resolution in the time-frequency spectrum, which is essential to capture subtle changes in the Doppler signatures.

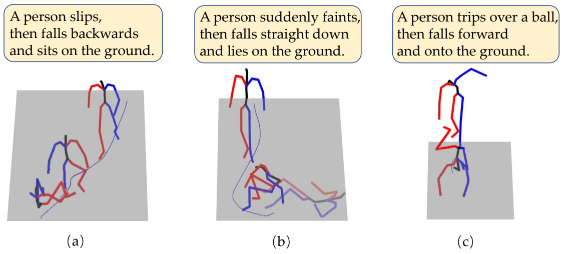

A person slips, then falls backwards and sits on the ground.
A person suddenly faints, then falls straight down and lies on the ground.
A person trips over a ball, then falls forward and onto the ground.
(a)
(b)
(c)

Fig. 2. Generate different types of falling from text descriptions

B. Importance of Dataset Diversity

In this subsection, we demonstrate the importance of dataset diversity in model comparisons. The test models include VGG-7[15], ResNet-18[16], ViT-tiny[17], CRNN[18], and Conv1D-LSTM[2]. According to Table 2, it can be observed that for the simple case, all evaluated models achieve a remarkably high accuracy. This observation suggests that datasets without motion diversity are insufficient to effectively benchmark the capabilities of these models. Conversely, in the context of more challenging datasets, it is evident that the larger convolutional networks significantly outperform both the lightweight RNN-based models and the ViT models.

TABLE 2. Classification Performance

Model	Accuracy(%)		Params(G)	FLOPs(M)
	Text2D Simple	Text2D Hard		
Conv1D-LSTM	98.26	80.80	0.008	0.111
CRNN	99.65	86.95	0.415	0.895
VGG-7	98.44	87.87	2.095	0.298
ResNet-18	98.96	91.26	1.824	11.180
ViT-Tiny	99.48	79.64	1.433	7.261

From the confusion matrix shown in Fig. 3, it is clear that the improvement in performance of the convolutional network is primarily due to its improved ability to distinguish between walking and abnormal walking, and between falling and dancing. This distinction is critical as the detection of abnormal walking and falling is often the core functionality of commercial radar sensor applications. Although a more compact RNN-based model may be sufficient for detecting simple daily activities, the use of a larger convolutional network proves more effective in accurately identifying complex movements, such as dancing, and detailed motions, such as abnormal walking.

C. Sim2Real Transfer Learning

In Fig. 4 we show some examples between the real world measurements and the simulated data. We can see that the simulated data show similar motion patterns to the real data, but the details show differences. Because of this distributional shift, we adopt the sim-to-real transfer learning paradigm to utilize the simulated data. For this analysis, we test an extreme case where the small-size dataset [3] contains 11 types of activity with only 60 samples per class. The direct training of ResNet-18 on this dataset presented difficulties due to the instability of the training process. Consequently, we utilized a

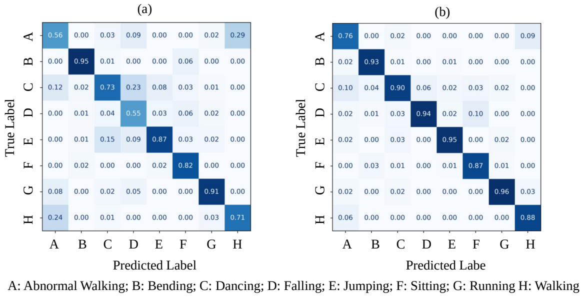


Fig. 3. Confusion matrix: (a) for CRNN and (b) for ResNet-18

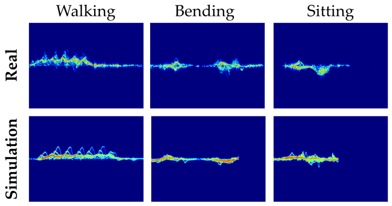


Fig. 4. Comparisons between real-world measurements and simulated data for some typical activities

classical machine learning classifier as the baseline. Specifically, we extract HOG features [19] from the spectrogram and train an SVM classifier. For the sim2real transfer learning, we pre-train ResNet-18 on the simulated dataset, followed by fine-tuning on the real dataset.

We investigate three transfer learning strategies, including linear probing, last layer fine tuning and full model fine tuning [20]. The results of our transfer learning experiments, as detailed in Table 3, show that linear probing has even lower performance than the machine learning method trained on the real dataset and exhibits slow convergence rates due to the distributional shift. Last layer fine tuning slightly improves the performance compared to the machine learning method. The best results, characterised by fast convergence, are achieved by full model tuning, which jointly tunes the feature extractor and the classification head. We further test these methods on a larger dataset [21], and the results are consistent. These results suggest that pre-training on simulated data, even if not high fidelity, is still advantageous to stabilize convergence and transfer prior knowledge to the real dataset.

TABLE 3. Performance in Real World Data

Method	Average Accuracy (%)	Average Number of Epochs
HOG + SVM	76.12	-
Linear Probing	70.31	74
Last Layer Fine Tuning	77.08	52
Full Model Fine Tuning	84.72	18

IV. CONCLUSION

In this study, we present a text-to-radar simulation framework designed to enable large-scale radar simulations that are particularly suited to HAR tasks. A notable advantage of this model is its ability to significantly improve data collection for less commonly observed activities, such as falling and abnormal walking. Experiments suggest that dataset diversity is critical for fair model performance and the effectiveness of sim2real transfer learning. Future research could focus on improving the fidelity of the simulation by utilizing generative models and considering motion interferences.

REFERENCES

[1] Z. Wu, Z. Cao, X. Yu, J. Zhu, C. Song, and Z. Xu, “A novel multi-person activity recognition algorithm based on point clouds measured by millimeter-wave mimo radar,” *IEEE Sensors Journal*, 2023.

[2] Y. Zhou, M. López-Benítez, L. Yu, and Y. Yue, “Improving performance with feature enhancement and ranking constraints for radar-based human activity recognition,” in *IET International Radar Conference (IRC 2023)*, 2023, pp. 1888–1895.

[3] S. Z. Gurbuz, M. M. Rahman, E. Kurtoglu, T. Macks, and F. Fioranelli, “Cross-frequency training with adversarial learning for radar micro-doppler signature classification,” in *Radar Sensor Technology XXIV*, vol. 11408. SPIE, 2020, pp. 58–68.

[4] Y. Zhou, X. Yu, M. López-Benítez, L. Yu, and Y. Yue, “Corruption robustness analysis of radar micro-doppler classification for human activity recognition,” *IEEE Transactions on Radar Systems*, vol. 2, pp. 504–516, 2024.

[5] S. Vishwakarma, W. Li, C. Tang, K. Woodbridge, R. Adve, and K. Chetty, “Simhumalator: An open-source end-to-end radar simulator for human activity recognition,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 37, no. 3, pp. 6–22, 2021.

[6] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, “Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–10.

[7] N. Kern, P. Schoeder, and C. Waldschmidt, “Virtually augmented radar measurements with hardware radar target simulators for machine learning applications,” *IEEE Sensors Letters*, 2024.

[8] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, A. Zhou, and H. Ma, “Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks,” *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–26, 2023.

[9] K. Deng, D. Zhao, Z. Zhang, S. Wang, W. Zheng, and H. Ma, “Midas++: Generating training data of mmwave radars from videos for privacy-preserving human sensing with mobility,” *IEEE Transactions on Mobile Computing*, 2023.

[10] M. Rahman, S. Gurbuz, and M. Amin, “Physics-aware generative adversarial networks for radar-based human activity recognition,” *IEEE Transactions on Aerospace and Electronic Systems*, 2022.

[11] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, “Momask: Generative masked modeling of 3d human motions,” *arXiv preprint arXiv:2312.00063*, 2023.

[12] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[13] N. Kern, J. Aguilar, T. Grebner, B. Meinecke, and C. Waldschmidt, “Learning on multistatic simulation data for radar-based automotive gesture recognition,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 11, pp. 5039–5050, 2022.

[14] A. Van Den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” *Advances in neural information processing systems*, vol. 30, 2017.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[18] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[19] A. Dey, S. Rajan, G. Xiao, and J. Lu, “Radar-based fall event detection using histogram of oriented gradients of binary encoded radar signatures,” *IEEE Sensors Letters*, 2023.

[20] X. Li, S. Liu, J. Zhou, X. Lu, C. Fernandez-Granda, Z. Zhu, and Q. Qu, “Principled and efficient transfer learning of deep models via neural collapse,” *arXiv preprint arXiv:2212.12206*, 2022.

[21] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. L. Kernec, “Radar signatures of human activities,” 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:203152928