Virtual IMU Data Augmentation by Spring-Joint Model for Motion Exercises Recognition without Using Real Data

Chengshuo Xia
Keio University
Japan
csxia@keio.jp

Yuta Sugiura
Keio University
Japan
sugiura@keio.jp

ABSTRACT

A conventional motion exercises recognition system only tracks designated motion types, and it enables users cannot use a customized system according to personal needs. The virtual IMU data provides a new opportunity to reduce the cost of training datasets and flexibly design the activity recognition system using online resources. To better design a user-customized motion exercises recognition system using virtual IMU data, this paper proposes a virtual IMU sensor module with a spring-joint model to augment the virtual acceleration signal from the limited online 2D video. The original virtual acceleration signal is extended with data from different acceleration distributions generated by the spring-joint model and used to train a motion exercises recognition system. The proposed method can design a classifier for three motions with limited video resources, showing an average accuracy of 85.5% on the real motion data of seven individuals.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous computing**.

KEYWORDS

Data Augmentation; Motion Exercises Recognition; Virtual IMU

ACM Reference Format:
Chengshuo Xia and Yuta Sugiura. 2022. Virtual IMU Data Augmentation by Spring-Joint Model for Motion Exercises Recognition without Using Real Data. In *The 2022 International Symposium on Wearable Computers (ISWC '22), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3544794.3558460

1 INTRODUCTION

Exercise and health have become growing concerns for many people. To stay fit, people have become accustomed to incorporating fitness training into their daily lives [5, 22]. By enabling computing systems to understand a user's fitness behavior, inertial measurement unit (IMU)-based motion exercises recognition enables daily mobile systems such as the Apple Watch-based fitness tracker [4]. However, because of the diversity of end users, the desired features of a motion exercises recognition system differ according to varying physical conditions and demands. In existing systems, users cannot use motion recognition with personalized features, such as setting the tracking motions according to their needs. The reason is that most current IMU-based motion exercises recognition systems have been developed based on machine learning, in which objects recognized by the system are determined by a pre-defined dataset. The costly collection process makes it difficult to obtain a dataset multiple times by following the needs of individuals. Therefore, it is important to use a low-cost dataset acquisition method to address end users' personalized needs in motion exercises recognition.

In recent years, innovative machine learning has emerged as a method that relies on 2D RGB video to acquire virtual IMU data. Using the deep vision model, the 3D skeleton joint coordinates of the subject's motion in the 2D video can be extracted and applied to a humanoid avatar, thus achieving conversion from a 2D motion video to 3D virtual motion [9, 14]. Accessing the positional changes of an avatar's motion can calculate the virtual acceleration signals. By mixing the virtual acceleration extracted from public videos with the real acceleration, the robustness of the trained classifier can be enhanced and has been proved to identify locomotion and finger motion [9, 12].

The approach to extracting the virtual acceleration could help expand IMU-based motion recognition training datasets. Since the acquisition of sensor signals is more flexible and cheaper via this approach, it can be used for flexible selection of motions in traditional motion exercise recognition. The system input can be 2D videos so that users can use a wide range of online motion instruction videos to select their desired motion, obtain the acceleration signal of the virtual IMU, and train a customized recognition system. However, because most online motion instructional videos have short motion time lengths (20-40 seconds) for a single exercise, fewer virtual IMU signals are obtained frame-by-frame. The performance of limited training data worsens in the face of inter-difference and intra-difference in users' motions. Thus, augmenting virtual IMU signals corresponding to the motion of limited exercise videos can help build a classifier with good performance for application by individuals.

In this study, a data augmentation method was developed for virtual acceleration data from online video to recognize motion exercises. A spring-joint-based sensor module was designed in the virtual environment, and nonlinear kinematic characteristics of the spring-joint model were used to generate acceleration signals with different spatial distributions. Other temporal distributions were achieved by changing the playback speed of the 3D motion. Thus, the limited virtual acceleration signal dataset was extended to train the classifier and adapt the virtual data to recognize the real

ISWC '22, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9424-6/22/09.
https://doi.org/10.1145/3544794.3558460

motions of the user. Using this approach, it is possible to implement online motion exercises video in a user-customized recognition system design without the need for real-world data.

2 RELATED WORK

IMU is one of the most popular solutions for detecting the kinematic characteristics of human motion [6, 13, 18]. Combined with commercial off-the-shelf devices, the IMU-based detection system has been used to successfully track users' daily exercise in devices such as Apple Watch Fitness and Apple Health [1]. To facilitate the IMU-based motion recognition system building, close attention has been paid to the extraction of virtual IMU data to reduce the cost of collecting a training dataset. IMUSim [23] introduced the first design of a virtual IMU sensor system based on 3D motion sequences from motion capture (MoCap) equipment, including acceleration, angular velocity, and magnetic calculation, as well as a data processing unit and noise simulation. To further apply virtual IMU data in a machine learning-based system, Kang et al. [7] employed the Unity-embedded animation and extracted the virtual IMU data to train the classifier, which then recognized the real activities of standing, walking, and jogging. Previous works generally utilized the advanced MoCap equipment to reconstruct the 3D human motion and calculate the virtual IMU signal [19]. Recently, IMUTude [9] extracted a virtual IMU signal based on 2D video and converted it to 3D motion in Blender [2]. The received virtual IMU data were used to mix the real dataset as an augmentation tool to improve the built classifier's performance. Liu et al. [12] further utilized video sources to contribute a fine-tuned IMU-based finger motion system that extracted a virtual IMU signal from an online sign language video tutorial and trained the classifier to recognize real hand motions.

In addition to working on a 3D avatar to calculate the virtual IMU data, the deep end-to-end network can be designed to output the virtual IMU data directly from the imported 2D video [8, 10, 24]. Rey et al. [16] presented a network that employed real IMU and laboratory videos to build a regression model that could be used to extract IMU data from opportunistic videos.

3 VIRTUAL IMU DATA AUGMENTATION FOR MOTION EXERCISES

During the motion exercises, the signal distribution of the generated on-body IMU signal is not complex and normally has periodic characteristics. Thus, typical time- and frequency-based features can discriminate different motions. However, because of the differing physical conditions of individual users (e.g., body size, physical fitness, etc.), applying a trained classifier model requires considering inter-user variations (i.e., different users making the same motion) and intra-user variations (i.e., signal differences when a single user makes the same motion).

Therefore, in performing data augmentation with finite signals, it is necessary to add 'random' factors to the original signal to simulate the possible intra- and inter-difference but not affect the distribution of signal features between different movements. Therefore, in this work, we used the reconstructed 3D human motion and designed the related acceleration signal augmentation in time and space to generate more acceleration signals. Figure 1 presents an overview of the developed method.

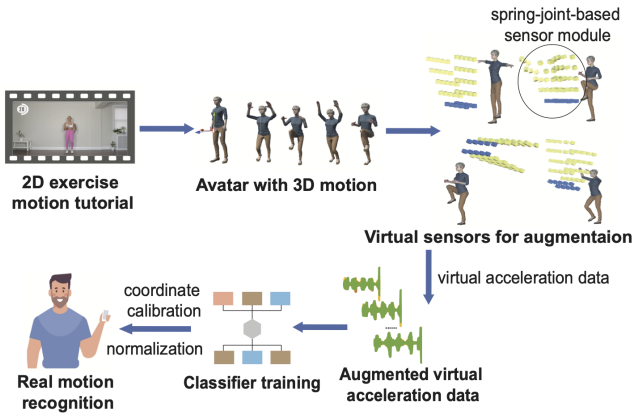

Figure 1: Overview of the designed virtual IMU data augmentation method and its corresponding usage.

3.1 Spatial Augmentation: Spring-Joint-based Sensor Module

In an avatar's motion, an individual limb's movement can be recognized as a translation and rotation process. Thus, linear acceleration and angular velocity distributions basically remain uniform across the surface of the limbs. To change the signal's spatial distribution, we adopted a soft-body structure to obtain the different acceleration distributions in a single motion.

Specifically, a spring-joint model was used to build a hierarchical structure sensor module. The sensor module contained several child nodes that were connected to the root node via a spring joint. Thus, to modify the spring's characteristics and the root's movement, each child node produced different acceleration distributions but followed the same motion pattern. Figure 2 shows the designed sensor module and movements applied by force to the root node.

The spring-joint model was developed based on Unity 3D with *PhysKit*, an externally developed physics simulation tool [15]. Two key factors are essential in a spring-joint model: *spring characteristics* and *node number*. The *spring characteristics* usually include two important parameters: the damping coefficient, which determines the oscillations after movement, and the stiffness. As stiffness increases, the structure becomes increasingly rigid and the kinematic distribution becomes the same. The *node number* indicates the number of different linear accelerations that can be produced in a single-sensor module. Theoretically, the higher the number of nodes, the greater the variety of distributions. However, when the number of hierarchical nodes increases, stochastic movements increase in the end nodes. Therefore, three nodes were chosen for connection in this sensor module.

3.2 Temporal Augmentation: Playback Speed

In terms of time, different users in the real world may complete tasks at different speeds. Therefore, acceleration signals generated at different speeds can be simulated in virtual space. Because the 3D

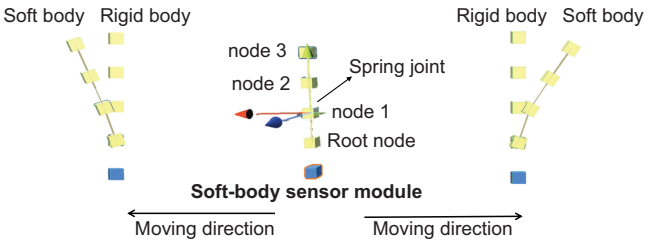

Figure 2: The developed spring-joint-based sensor module.

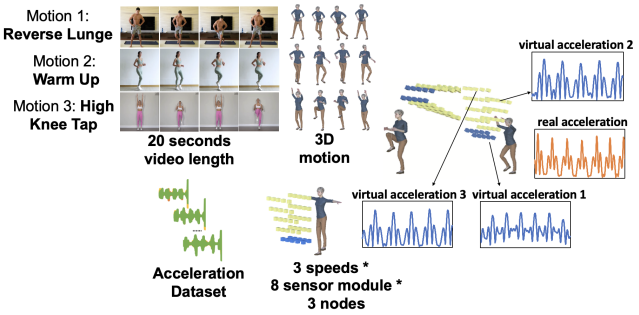

Figure 3: Aerobic motions in the experiment and the virtual signals produced (in a *High Knee Tap* motion).

motion is input frame-by-frame, it is possible to simulate the avatar in completing the same motion at different speeds by changing the playback speed, a simple step in Unity 3D.

3.3 Virtual IMU Data Generation

In this work, the commercial 3D motion conversion software *DeepMotion* was used to convert a 2D video to 3D motion [3]. Following the instructions of the *DeepMotion* tool, the selected video was supposed to maintain a clear resolution, whole-body appearance, and relatively larger movement. An exercise tutorial video was sufficient to satisfy the requirements. The virtual acceleration was calculated from the position data as introduced in IMUSim [23].

3.4 Domain Adaptation from Virtual to Real World

Domain adaptation is one of the most challenging problems that limit the application of classifiers trained on virtual data for real-world recognition. It aims to eliminate the possible differences between the virtual and real-world signals caused by the coordinate system of virtual space, avatar body size, and so on. Traditional domain adaptation is based on transfer learning using a small amount of real data [21]. To further decrease the requirement for real data, we employed sensor coordinate calibration with a normalized signal to realize a virtual-to-real-world domain adaptation, including sensor coordinate system calibration and signal fusion and normalization.

3.4.1 Sensor Coordinate System Calibration. The signal of a real IMU sensor is usually in the sensor's local coordinate system, while virtual data are obtained from the global coordinate system in the virtual environment. Thus, we converted the acceleration data into a sensor-fixed coordinate system to compare the sensor data in the two domains. We specified the coordinate system corresponding to the attitude of the body-attached IMU before the motion starts as a fixed coordinate system and convert the signal to this coordinate system.

3.4.2 Signal fusion and normalization. Besides the coordinate system, the signal's magnitude and altitude may also cause the two domains to differ. Therefore, three-axis acceleration data were fused to calculate the magnitude, independent of the sensor's orientation. Normalization was then conducted to eliminate different scale standards regarding the acceleration value [11].

4 EVALUATION

4.1 Motion Exercises and Data Collection

Three exercises tested consisted of aerobic motions selected from online YouTube sources (i.e., *Reverse Lunge*, *Warm Up* and *High Knee Tap*). Each motion was extracted as a segment from the initial online video, and the time length was 20 seconds. According to the examined motion, we selected the *right upper leg* as the sensor placement and generated the related virtual dataset via the proposed method. To evaluate the data, we recruited seven participants to collect a real IMU dataset. A real IMU sensor (Xsens Dot [17]) was placed on the *right upper leg* of each participant. We showed the initial motion tutorial to the participants and requested that they conduct the exercise. Each motion was recorded in 90 seconds. Figure 3 shows the aerobic motions performed by each participant.

4.2 Classification System Design

As the acceleration signal was normalized, the dimension of the obtained signal remained relatively low. Thus, we applied hand-crafted features-based machine learning; that is, the support vector machine (SVM), random forest (RF), and decision tree (DT) models were examined. The sensor data were segmented into four-second lengths with two-second overlaps. The features were as follows: *mean*, *variance*, *standard variance*, *75th percentile*, *25th percentile*, *mean and median value of power spectrum*, *mean and median frequency of power spectrum*, *Entropy*, *Empirical Mode Decomposition* and the coefficients of *Fast Fourier transform* result. After the features were extracted, a principal component analysis (PCA) was conducted to decrease the number of feature dimensions to two.

4.3 Test Different Spring-Joint Module Characteristics

The spring-joint-based sensor module produces different kinematic characteristics related to its node connection structure and spring parameters. We chose to compare two structures, type 1 (where all nodes are connected vertically) and type 2 (where the nodes are connected in 3D space in a distributed manner). Figure 4 shows different acceleration signal produced with various stiffness parameters for a *Reverse Lunge* motion. The damping coefficient was set as 0.1 to reduce the oscillations, and stiffness was varied from 0.2

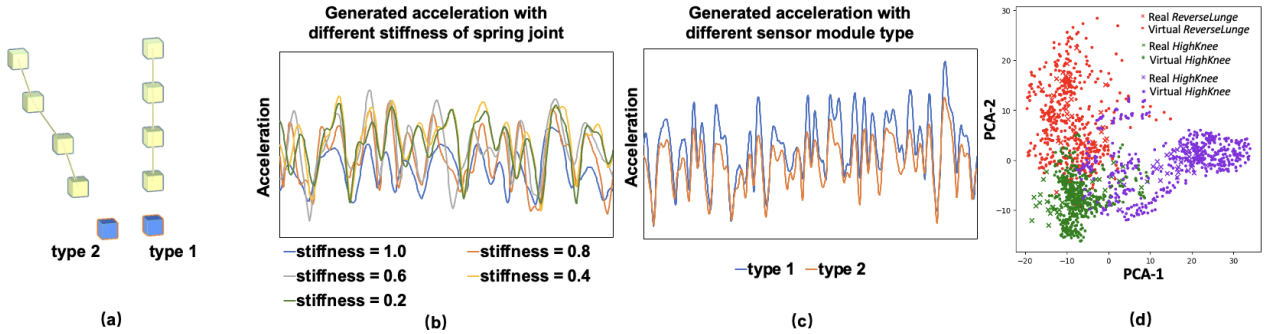

Figure 4: (a) is the used two types of connection structure. (b) and (c) are the different virtual acceleration with various spring parameters (in a *Reverse Lunge* motion). (d) is the PCA-based features from both virtual and real signal with three motions.

to 1.0 (while 1.0 means a rigid body). From the result, the smaller the stiffness, the greater the variation of the generated acceleration signal compared to the signal corresponding to the rigid body (stiffness = 1). In order to reduce the distortion of the original virtual signal, the main design is to vary the stiffness between 0.8 and 0.9 (0.86, 0.88, 0.90, 0.92) in this paper.

4.4 Test on the Real User's Data

We applied the trained classifier to each participant to compare it with the performance trained using an augmented virtual dataset (total 20s * 8 modules * 3 nodes * 3 speeds * 3 motions = 4320s data length). The test dataset was 90s * 3 motions * 7 participants. The baseline comparison used the classifier trained by the initial virtual acceleration signal (i.e., not use the proposed augmentation method). Therefore, the time length of the baseline classifier dataset was the same as that of the original video (total 20s * 3 motions = 60s data length). Both initial virtual acceleration and virtual acceleration generated from spring-joint-based sensor module followed the same processing steps, including the domain adaption, feature extraction and so on. We also evaluated the classifier training by real dataset through the leave-one-subject-out (LOSO) method as the real-world baseline method. As shown in Table 1, the performance of the classification using data augmentation was significantly improved. When it was applied to the seven subjects, it showed an improved recognition rate. The average recognition rate was 85.3%. Compared with the baseline method, the accuracy of the classifier without augmentation was 45.5%. And the real-world LOSO testing approach showed 78.2% accuracy.

Table 1: Results of the proposed augmentation method for different participants. (Avg. is average recognition accuracy.)

Classifier		Subject & Result (%)							Avg.
		P1	P2	P3	P4	P5	P6	P7	
Spring-joint module	RF	65.9	87.9	92.4	94.7	81.1	90.9	84.5	85.3
	SVM	65.9	88.6	94.6	95.5	78.8	90.2	81.9	85.1
	DT	65.9	85.6	92.4	93.9	79.5	86.3	83.3	83.8
Initial virtual signal	RF	55.3	42.2	41.6	43.2	43.9	49.2	43.2	45.5
	SVM	51.2	49.2	43.1	50.1	38.6	45.5	43.2	45.8
	DT	59.8	43.2	44.6	40.2	38.6	46.2	44.6	45.3
Real dataset	RF	81.1	79.5	77.3	78.0	71.9	76.5	83.3	78.2
	SVM	76.5	80.3	75.0	61.4	77.3	76.5	84.1	75.8
	DT	71.9	77.3	75.0	75.0	71.2	73.4	75.8	74.2

5 CONCLUSION AND DISCUSSION

This work proposed a spring-joint-based sensor module to augment the virtual acceleration signal based on exercise video input. Three different aerobic videos were tested with limited length. The motions examined in this experiment were limited to decrease the difference between the two domains, and the employed data dimension was very low. Therefore, the classifier's performance is not high for complex or similar motions.

In future work, more attention can be paid to investigating virtual sensor-dedicated feature mining or classifier structures, such as the domain-invariant features from virtual to real. In addition, traditional IMU-based signal augmentation is a direct secondary modification of collected signals [20], which lacks specific evidence. The method designed in this paper relied on a reconstructed motion to increase the realism of the augmented signal to some extent. Through the spring-joint connection, each child node can generate a different distribution of acceleration signals. In fact, the aerobic motions tested produce large displacement and acceleration signals mainly in the vertical direction of the body (i.e., perpendicular to the transverse plane). Therefore, the child node connection structure designed in this paper is connected vertically, which helps to generate different acceleration signals in the vertical direction by different vertical forces for enhancement. Exploring how the spring-joint connection generate different signal could be the next step.

To further enhance the realism, a secondary modification could be made to the original motion sequence to fine-tune the motion trajectory of some joints to achieve the simulation of different human motions. The resulting virtual sensor signal distribution would closely resemble a real situation.

ACKNOWLEDGMENTS

This work was supported by JST PRESTO Grant Number JPMJPR2134.

REFERENCES

[1] Apple. 2022. Apple Health. https://www.apple.com/ios/health/.
[2] Blender. 2022. Blender. https://www.blender.org/.
[3] DeepMotion. 2022. DeepMotion. https://www.deepmotion.com/.
[4] Paul Dempsey. 2015. The teardown: Apple Watch. *Engineering & Technology* 10, 6 (2015), 88–89.
[5] Ryan S Falck, Jennifer C Davis, John R Best, Rachel A Crockett, and Teresa Liu-Ambrose. 2019. Impact of exercise training on physical and cognitive function among older adults: a systematic review and meta-analysis. *Neurobiology of aging* 79 (2019), 119–130.
[6] Andrea Ferlini, Alessandro Montanari, Cecilia Mascolo, and Robert Harle. 2019. Head motion tracking through in-ear wearables. In *Proceedings of the 1st International Workshop on Earable Computing*. 8–13.
[7] Cholmin Kang, Hyunwoo Jung, and Youngki Lee. 2019. Towards machine learning with zero real-world data. In *The 5th ACM Workshop on Wearable Systems and Applications*. 41–46.
[8] Hyeokhyen Kwon, Gregory D Abowd, and Thomas Plötz. 2021. Complex Deep Neural Networks from Large Scale Virtual IMU Data for Effective Human Activity Recognition Using Wearables. *Sensors* 21, 24 (2021), 8337.
[9] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.
[10] Arttu Lämsä, Jaakko Tervonen, Jussi Liikka, Constantino Álvarez Casado, and Miguel Bordallo López. 2022. Video2IMU: Realistic IMU features and signals from videos. *arXiv preprint arXiv:2202.06547* (2022).
[11] Xi'ang Li, Jinqi Luo, and Rabih Younes. 2020. ActivityGAN: Generative adversarial networks for data augmentation in sensor-based human activity recognition. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 249–254.
[12] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When video meets inertial sensors: zero-shot domain adaptation for finger motion analytics with inertial sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 182–194.
[13] Nobuyuki Oishi, Benedetta Heimler, Lloyd Pellatt, Meir Plotnik, and Daniel Roggen. 2021. Detecting Freezing of Gait with Earables Trained from VR Motion Capture Data. In *2021 International Symposium on Wearable Computers*. 33–37.
[14] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7753–7762.
[15] PhysKit. 2022. PhysKit. https://www.heathen.group/physkit.
[16] Vitor Fortes Rey, Peter Hevesi, Onorina Kovalenko, and Paul Lukowicz. 2019. Let there be IMU data: Generating training data for wearable, motion sensor based activity recognition from monocular rgb videos. In *Adjunct proceedings of the 2019 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2019 ACM international symposium on wearable computers*. 699–708.
[17] Nana Schlage, Andreas Kitzig, Gudrun Stockmanns, and Edwin Naroska. 2021. Development of a mobile, cost-effective and easy to use inertial motion capture system for monitoring in rehabilitation applications. *Current Directions in Biomedical Engineering* 7, 2 (2021), 586–589.
[18] Marcus Schmidt, Carl Christian Rheinländer, Sebastian Wille, Norbert Wehn, and Thomas Jaitner. 2016. IMU-based determination of fatigue during long sprint. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*. 899–903.
[19] Shingo Takeda, Tsuyoshi Okita, Paula Lago, and Sozo Inoue. 2018. A multi-sensor setting activity recognition simulation tool. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 1444–1448.
[20] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 216–220.
[21] Chengshuo Xia, Ayane Saito, and Yuta Sugiura. 2022. Using the virtual data-driven measurement to support the prototyping of hand gesture recognition interface with distance sensor. *Sensors and Actuators A: Physical* 338 (2022), 113463.
[22] Binbin Yong, Zijian Xu, Xin Wang, Libin Cheng, Xue Li, Xiang Wu, and Qingguo Zhou. 2018. IoT-based intelligent fitness system. *J. Parallel and Distrib. Comput.* 118 (2018), 14–21.
[23] Alexander D Young, Martin J Ling, and Damal K Arvind. 2011. IMUSim: A simulation environment for inertial sensing algorithm design and evaluation. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. IEEE, 199–210.
[24] Shibo Zhang and Nabil Alshurafa. 2020. Deep generative cross-modal on-body accelerometer data synthesis from videos. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 223–227.