RFID-Pose: Vision-Aided Three-Dimensional Human Pose Estimation With Radio-Frequency Identification

Chao Yang, *Student Member, IEEE*, Xuyu Wang, *Member, IEEE*, and Shiwen Mao, *Fellow, IEEE*

***Abstract*—In recent years, human pose tracking has become an important topic in computer vision (CV). To improve the privacy of human pose tracking, there is considerable interest in techniques without using a video camera. To this end, radio-frequency identification (RFID) tags, as a low-cost wearable sensor, provide an effective solution for 3-D human pose tracking. In this article, we propose RFID-Pose, a vision-aided realtime 3-D human pose estimation system, which is based on deep learning assisted by CV. The RFID phase data are calibrated to effectively mitigate the severe phase distortion, and high accuracy low rank tensor completion is employed to impute the missing RFID data. The system then estimates the spatial rotation angle of each human limb, and utilizes the rotation angles to reconstruct human pose in realtime with the forward kinematic technique. A prototype is developed with commodity RFID devices. High pose estimation accuracy and realtime operation of RFID-Pose are demonstrated in our experiments using Kinect 2.0 as a benchmark.**

***Index Terms*—Computer vision (CV), deep learning, high accuracy low rank tensor completion (HaLRTC), human pose estimation, Radio-frequency Identification (RFID).**

I. Introduction

IN RECENT years, human pose tracking has become an important topic in computer vision (CV), evolving from 2-D [1] to 3-D poses [2]. The accuracy of human pose tracking technique is continuously improved by more advanced hardware and machine learning (i.e., deep learning) techniques. Camera-based techniques have been shown effective for human pose tracking. However, such vision-based techniques also raise security and privacy concerns. It is usually annoying if one is being watched by a video camera all day. It is reported that millions of wireless security cameras deployed around the world are at risk of being hacked [3]. The video data used for pose tracking could be intercepted and illegally used by hackers. The privacy issue draws increasing concerns in the age of Internet of Things (IoT), where eHealth based on IoT is an important part. Many techniques have been proposed to improve the privacy and reliability of the IoT [4]–[6].

With rapid development of machine learning, deep learning has been highly promising for improving the safety and reliability of personal software and the IoT, which usually relies on sufficient and high-quality data [7]–[9]. If the human pose data are obtained without using a camera, people will no longer worry about their privacy being threatened. To address this issue, several radio frequency (RF) sensing-based schemes have been proposed for human pose estimation, such as WiFi [10], [11], frequency-modulated continuous wave (FMCW) radar [12], and mmWave radar [13]. Unlike camera-based techniques, such RF sensing-based schemes estimate the human joints from a confidence map constructed by RF signals, so the user's privacy will be preserved. For example, channel state information (CSI) is utilized in WiFi-based systems [11], and the human pose can be estimated with a deep neural network such as a convolutional neural network (CNN). However, due to the multipath effect, WiFi signals are highly sensitive to interference (e.g., movements) in the surrounding environment. Although FMCW radar is more robust to the environment interference than WiFi-based systems, the cost of the system is higher than commodity WiFi, which hinders its wide deployment.

To this end, radio frequency identification (RFID) provides a promising solution for human pose estimation. Compared with the above contact-free RF sensing systems, RFID tags can be used as wearable sensors because of their small size. The interference caused by the multipath effect is much smaller in the RFID system. Furthermore, the cost of RFID systems is lower than the advanced radar-based systems such as the FMCW radar. However, because of the low data rate in RFID systems, generating a joint confidence map for all joints, as in other RF-based systems, is highly challenging. Consequently, the existing RFID-based pose tracking systems are focused on monitoring the movements of one particular limb using the phase data sampled from multiple tags [14], [15]. When multiple joints are moving simultaneously, the performance could be affected by the disturbance of other RFID tags (e.g., the mutual coupling effect) or the intertag collisions. Thus, tracking the entire body with RFID tags is still a challenging and open problem.

Manuscript received June 20, 2020; revised September 19, 2020; accepted October 11, 2020. Date of publication October 28, 2020; date of current version August 31, 2021. This work was supported in part by the NSF under Grant ECCS-1923163 and Grant CNS-1822055, and in part by the Wireless Engineering Research and Education Center (WEREC) at Auburn University. Associate Editor: Y. Lin *(Corresponding author: Shiwen Mao.)*

Chao Yang and Shiwen Mao are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA (e-mail: czy0017@tigermail.auburn.edu; smao@ieee.org).

Xuyu Wang is with the Department of Computer Science, California State University, Sacramento, CA 95819-6021 USA (e-mail: xuyu.wang@csus.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TR.2020.3030952

In this article, we address the challenges in human pose estimation using RFID tags with a novel vision-aided, deep learning solution. We propose the RFID-Pose system for tracking the movements of multiple human limbs in realtime. In the proposed system, RFID tags are attached to the target human joints. The movement of the tags are captured by the phase variations in the responses from each tag. We propose a vision-aided solution to help the proposed deep learning model to learn the features of tag phase variations, rather than localizing these tags with traditional tag localization techniques [16]. The collected RFID phase data are first preprocessed to improve the quality of the raw sampled data, in particular, to mitigate the phase distortion and estimate the large amount of missing samples. Then, we leverage a deep kinematic neural network to learn the features of RFID phase data, where a Kinect 2.0 is used to obtain the ground truth (i.e., labeled data for training). With the assistance of vision data, the deep learning model transforms the phase variation into the spatial rotation angle of each human joint. Since the spatial rotation angle estimation does not require generating a confidence map, the low data rate limitation of RFID systems is no longer an issue. In realtime estimation, human pose is reconstructed by estimated rotation angles from RFID data and the initial human skeleton. The vision data will not be needed anymore in this stage, and so the user's privacy can be well protected.

The main contributions of this article are summarized as follows.

1) To the best of our knowledge, this is the first work for 3-D human pose estimation using commodity RFID reader and tags, which can effectively monitor multiple human joints simultaneously in realtime.
2) We propose a novel data preprocessing approach to mitigate the severe RFID phase distortion and compensate the large amount of missing data in sampled raw RFID data. The tensor completion technique is utilized for data imputation, so that phase data for all RFID tags can be estimated. The greatly improved data quality leads to more effective learning for human pose estimation.
3) We propose a vision-aided solution for training the proposed deep kinematic neural network, to transform sensed RFID phase variations to the spatial rotation of each limb. The proposed approach effectively addresses the challenges of the low data rate in RFID systems, because rotation angle estimation requires much less data than generating a joint confidence map.
4) We develop a prototype system with commodity RFID devices and Kinect 2.0, to evaluate the system performance. Our experimental study validates that the proposed RFID-Pose system can effectively track the human pose with different types of motions in realtime.

The rest of this article is organized as follows. Section II reviews the related work. Section III presents the RFID Pose system overview. In Section IV, the challenges and solutions to RFID data preprocessing are presented. In Section V, the challenges and solutions to RFID-based pose estimation are analyzed and introduced. In Section VI, we present our prototype system evaluation. Finally, Section VII concludes this article. The notation used in this article is summarized in Table I.

TABLE I
NOTATION

Symbol	*Description*
T	Time slot for synchronized RFID data
t	Time slot of raw RFID data
$\vec{P}_n^T$	Position of joint n in time slot T
$\vec{P}_{parent(n)}^T$	Position of joint n's parent joint in time slot T
$\mathbf{R}$	Rotation matrix for 3D coordinates rotation
$\mathbf{R}_n^T$	Rotation matrix in forward kinematic layer for joint n in time slot T
$\ell + x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}$	Unit quaternion format
$\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$	Quaternion units
Q_T	Unit quaternions for FK layer input at time T
$a\boldsymbol{i} + b\boldsymbol{j} + c\boldsymbol{k}$	3D position vector
Φ	Raw RFID phase
Φ_{tag}, Φ_a	Phase offset caused by the RFID circuits and the reader antenna
Φ_α	Phase offset in channel α
ϕ	RFID phase variation
ϕ'	Calibrated phase variation
S	Tag to antenna distance
α	RFID channel index
f_α	Frequency of channel α
N_p	Total number of antennas
p	Antenna index
N_q	Total number of tags
q	Tag index
$\boldsymbol{\psi}$	Sparse phase variation tensor before synchronization
ϕ_{qt}^p	Calibrated phase variation from tag q sampled by antenna p in time slot t
ξ	Compression ratio from ϕ_{qt}^p to $\boldsymbol{\Psi}$
$\boldsymbol{\Psi}$	Sparse phase variation tensor after compression
N_T	Number of synchronized RFID data time slots
$\bar{\phi}_{qT}^p$	Mean phase variation from tag q sampled by antenna p in synchronized time slot T
$\hat{\boldsymbol{\Psi}}$	Estimation of the ideal tensor
$\boldsymbol{\Psi}_{ideal}$	Ideal tensor data
$\boldsymbol{\Omega}$	Mapping tensor composed of 0 and 1 elements
$\epsilon(T)$	mean error of all joints in time slot T
$\hat{P}_n^T$	Estimated position for joint n at time T
$\dot{P}_n^T$	Ground truth position for joint n at time T

II. RELATED WORK

This article is closely related to prior works on RFID–based sensing [17] and human pose estimation [18]. We mainly focus on these two classes of systems in the following.

Recently, passive RFID tags have attracted great interest because of their easy deployment and low-cost features [19]. The low level reader protocol used by the reader can provide useful low-level information such as received signal strength indicator (RSSI), phase, Doppler frequency shift, timestamp, etc. [20]. As a result, many RFID-based sensing techniques have been developed for many applications, such as indoor localization [16], [21]–[24], vital sign monitoring [25]–[31], user authentication [32], material identification [33], object orientation estimation [34], vibration sensing [35], anomaly detection [36], temperature sensing [37], and drone localization and navigation [38]–[40]. Particularly, the RF-wear system [15] and RF-Kinect system [14] utilize RFID tags attached to the human joints to estimate the movement of a particular limb, such as front arms, front legs, and thighs [14], [15]. We adopt the same approach in RFID-Pose. However, these systems may not be suitable for realtime human pose estimation, especially

when multiple moving joints need to be tracked simultaneously. These RFID-based sensing systems inspire us to develop an RFID-based pose estimation system.

Prior works on human pose estimation are mainly based on CV techniques [18], [41]. For human pose estimation using video data, deep learning-based method has been shown effective for 2-D human pose with conventional RGB cameras [1], [42], and 3-D human pose with RGB-Depth cameras [43] and VICON systems [44]. These camera-based techniques can achieve high accuracy, but all require sufficient lighting condition and may raise privacy concerns.

These limitations motivate the development of RF-based pose estimation techniques, because detecting RF signals do not require any lighting [45]. Moreover, since no video is used in the RF systems, the privacy issues are effectively addressed. However, collecting labeled pose data from RF signals is very challenging. Therefore, several RF-based techniques leverage vision data as labeled pose data to train the deep learning network. This approach is also taken in the proposed RFID-Pose system. For example, RFPose is the first work to use RF signals with an FMCW radar for 2-D human pose estimation, where a teacher–student deep learning model is utilized [12]. RFPose3D is the later version for 3-D human pose estimation with FMCW radar [45]. Moreover, mmwave radar is also utilized for human pose estimation with deep learning [13]. Recently, WiFi CSI has been exploited to create 2-D skeletions [10] and 3-D human poses [11] using cross-modal deep learning techniques. However, Radar and WiFi-based human pose estimation are easily influenced by the environment noise and interference, and the FMCW radar technique is limited by the relatively higher cost [e.g., implemented with universal software radio peripherals (USRP)].

The proposed RFID-Pose system, to the best of our knowledge, is the first to apply RFID-based sensing for 3-D human pose estimation. The proposed system consists of a novel and effective solutions for cross-modal 3-D human pose estimation using RFID and CV, which is much more robust compared with WiFi and Radar-based methods.

III. RFID-Pose System Overview

In this article, we propose an RFID-based sensing system, termed RFID-Pose, to estimate and track 3-D human pose in realtime. The RFID-Pose system can sense the 3-D positions of all the RFID tags attached to the human body by exploiting the phase data collected at the reader antennas. The training process of the system is supervised by the labeled vision data collected by a Kinect2.0 device, but only RFID data will be required for online human skeleton estimation. Human pose can be effectively constructed by mapping the positions of the attached RFID tags into 3-D coordinates. The overview of the RFID-Pose system architecture is presented in Fig. 1, which is mainly composed of four components, including 1) RFID phase data collection, 2) Kinect skeleton data collection, 3) RFID data preprocessing, and 4) Skeleton reconstruction using a deep kinematic neural network.

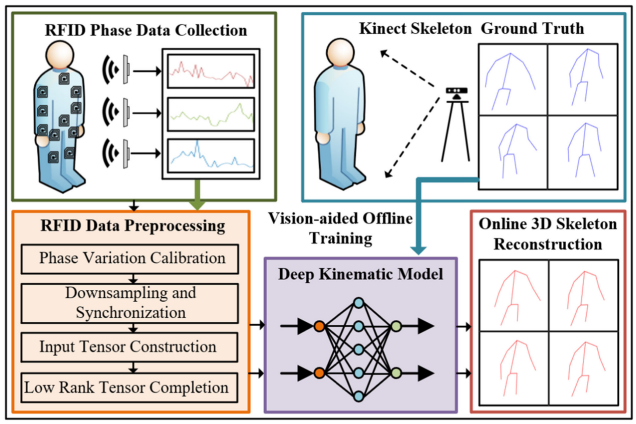


Fig. 1. Overview of the RFID-Pose system architecture.

A. RFID Phase and Kinect Pose Data Collection

In the proposed system, training data are sampled by both the RFID antennas and the Kinect 2.0 device simultaneously. The collected RFID data will be used as the input to the deep kinematic neural network, and the Kinect 3-D pose data will be used as labeled data for the supervised training. To collect RFID data, we attach passive RFID tags on the 12 joints of the human body. Three reader antennas are used to collect the phase and timestamp data from all the attached RFID tags. Kinect 2.0 is a depth camera widely used for capturing 3-D poses in interactive video games. The 3-D position of each human joint is estimated by both the RGB camera and the infrared sensors, and all measured joint positions are stored as 3-D coordinates.

B. RFID Data Preprocessing

Since the sampled RFID raw phase data suffers from considerable distortion caused by channel hopping and phase wrapping, the RFID phase calibration must be applied to cleanse the data before using it to train the deep neural network. We first calibrate the phase variation to mitigate the influence of channel hopping and phase wrapping. Next, we downsample the calibrated RFID data and synchronize it with the 3-D pose time sequence obtained by Kinect. However, because of the slotted ALOHA-like transmission in the RFID system, tags are not evenly interrogated by the antennas. In order to synchronize the RFID data with the collected pose data from Kinect, we should obtain the phase for all tags corresponding to each Kinect data frame. To this end, we propose to employ low rank tensor completion to estimate the missing phase values from the tags. Finally, the calibrated phase data are used as input to train the deep neural network for human skeleton reconstruction.

C. Human Skeleton Reconstruction With a Deep Kinematic Neural Network

In RFID-Pose, we incorporate the deep kinematic neural network to learn the features of the RFID phase data. Unlike monitoring one particular limb movement as in traditional

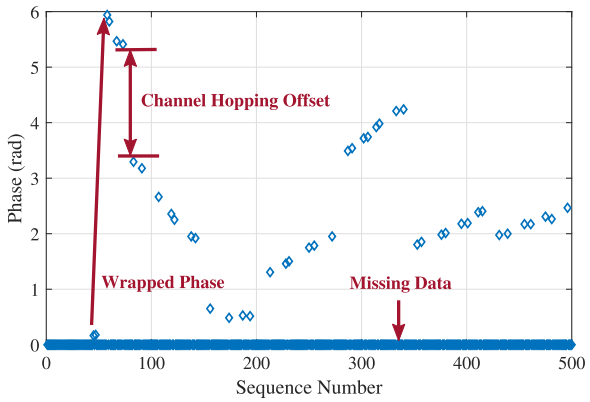


Fig. 2. Raw phase sampled from one of the RFID tags by a single reader antenna.

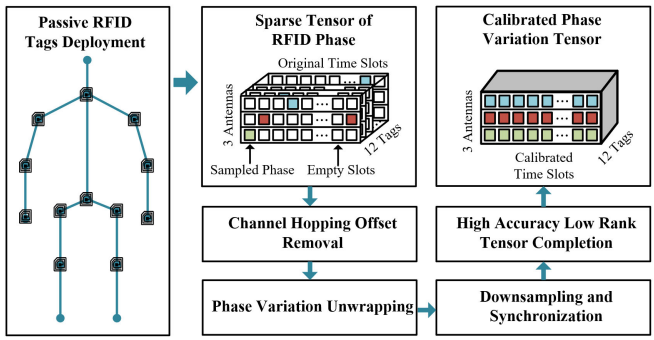


Fig. 3. Flow chart of RFID data preprocessing.

RFID-based skeleton tracking systems [14], [15], the deep kinematic neural network is designed to simultaneously estimate the spatial rotation of all human joints relative to their parent joints. Once the initial human skeleton (i.e., the length of the limbs of target) is given, the network could effectively learn the features of calibrated RFID tensor data, and reconstruct the positions of human joints with estimated rotation angles. In RFID-Pose, the Kinect pose data are only used as benchmark for evaluating the accuracy of 3-D pose reconstruction in the online testing process.

IV. Challenges and Solutions: RFID Phase Distortion Mitigation and Data Imputation

The proposed RFID-Pose system reconstructs 3-D human pose from RFID phase data with a deep kinematic neural network. However, the raw RFID phase data cannot be directly used for training and testing. The raw phase dataset from one of the tags sampled by a reader antenna in 500 time slots is plotted as diamond in Fig. 2. The figure shows that the collected RFID phase data are severely interfered during transmission by channel hopping and phase wrapping. Furthermore, there are many samples with a 0 value, which means the tag is not successfully sampled in the time slot. This is due to the slotted ALOHA transmission in RFID systems; only one tag is allowed to respond to the reader's query in each time slot. Such sparse, low quality RFID data makes the RFID-based 3-D pose tracking highly challenging unless an appropriate data preprocessing is conducted.

Therefore, we propose the following RFID data preprocessing for the sampled RFID phase data, as illustrated in Fig. 3. In the preprocessing procedure, we first calibrate the overall phase interference in the raw data and then synchronize the RFID phase data with the collected Kinect data (used as labels for training). Next, the RFID data are used to construct a third-order tensor, where the element at location (x, y, z) is the data collected from antenna x in time slot y from RFID tag z. We leverage high accuracy low rank tensor completion (HaLRTC) to recover the missing samples and form the input data tensor, which is fed into the deep kinematic neural network for training and inference. More details are provided in the following.

A. Combating Collected Phase Interference

1) Frequency Hopping Offset Mitigation: In the proposed system, we leverage an RFID reader to extract the phase data from received RFID tag responses using the low level reader protocol, which is indicative of the tag-to-antenna distance [20]. The phase value is obtained when the RFID reader receives the electronic product code (EPC) from the interrogated tag. The sampled phase value can be written as

$$\Phi = \mathrm{mod}\left(\frac{4\pi S f}{c} + \Phi_{tag} + \Phi_a,\ 2\pi\right) \tag{1}$$

where S denotes the distance between the interrogated tag and the reader antenna; and Φ_{tag} and Φ_a represent the phase offset caused by the circuits in the RFID tag and the reader antenna, respectively; f is the center frequency of the channel; and c is the speed of light. The equation shows that the phase value is indicative of the variation of the tag-to-antenna distance S, but it is also affected by the phase offset caused by the tag Φ_{tag} and the antenna Φ_a.

According to the FCC regulations, the Ultra-High Frequency (UHF) RFID system should hop among 50 channels during operation to avoid collisions among multiple RFID readers. In (1), the sum phase offset $\Phi_\alpha = \Phi_{\text{tag},\alpha} + \Phi_{a,\alpha}$ is determined by both the hardware and the current frequency f_α used for the interrogation. So a considerable phase offset will be generated each time when the system hops to a new channel. As shown in Fig. 2, the severe phase offset is caused by channel hopping, which leads to considerable interference in the collected phase data. To mitigate the interference, we first rewrite the sampled phase in (1) from each channel α as

$$\Phi = \mathrm{mod}\left(\frac{4\pi S f_\alpha}{c} + \Phi_\alpha,\ 2\pi\right),\ \alpha = 1, 2, \ldots, 50 \tag{2}$$

where α is the RFID channel index ranging from 1 to 50. The equation shows that the channel hopping offset is a constant value for each particular channel, which can be canceled by subtracting two phase samples on the same channel. Thus, rather than using the RFID phase data, we calculate the RFID phase variation on the same channel to mitigate the interference caused by the channel hopping offset.

The phase variation is calculated by subtracting a sampled phase data from the previous one on the same channel α, as

$$\phi = \mod\left(\frac{4\pi(S_n - S_{n-1})f_\alpha}{c},\ 2\pi\right)$$
$$\alpha = 1, 2, \ldots, 50,\ n = 2, 3, \ldots \quad (3)$$

where S_n represents the tag-to-antenna distance for the nth sampled data on the current channel. It can be seen that the phase variation in (3) is not affected by the phase offset anymore. Since $(S_n - S_{n-1})$ is the change of distance relative to the previous sample, phase variation is also suitable for tracking the movement of RFID tags. Therefore, to mitigate the interference caused by the frequency hopping offset, the input RFID data to the deep kinematic network is composed of the phase variation calculated for each RFID channel.

2) Phase Data Unwrapping: After calculating the phase variation for each channel, the phase distortion caused by channel hopping will be effectively mitigated. However, as shown in Fig. 2, since the sampled phase is wrapped in $[0, 2\pi]$ rad, the wrapped phase data also leads to severe interference in calculated phase variation. For example, if the phase changes from 0.1 rad to -0.1 rad, calculated phase variation will be $2\pi - 0.2$ rad, but the real phase variation is only -0.2 rad. To avoid the influence of phase wrapping, we apply a simple algorithm to unwrap the phase variation.

Considering that the frequency range of the reader antenna is 902–928 MHz with a wavelength about 33 cm, we assume that all the tag position variations between two adjacent samples is smaller than 16.5 cm (half of the weave length), which is reasonable given the 110-Hz sampling rate. Thus, we calibrate the calculated phase variation when its absolute value is larger than π as follows:

$$\phi' = \phi - 2\pi\frac{\phi}{|\phi|},\ \text{if } |\phi| > \pi. \quad (4)$$

In (4), $\phi/|\phi|$ returns the sign of ϕ. Then depending on whether the phase variation is positive or negative, a -2π or a 2π offset is added to ϕ. The calibrated phase variation, for the raw phase data shown in Fig. 2, is presented as diamonds in Fig. 4. We can see that, the channel hopping offset is eliminated in the calibrated data, as well as the phase distortion caused by phase wrapping. Notice that there are still missing data samples, which should be addressed. Otherwise, the input data still contains too many empty units (i.e., it is still highly sparse).

B. RFID Data Imputation

Following FCC regulations, the communications between the RFID reader and tags are based on slotted ALOHA. It means the back propagation data of all the tags are received randomly, and only one tag can respond to the reader in each time slot (i.e., only one phase sample can be collected from one of the tags at a time). In RFID-Pose, we employ a commodity RFID reader with three antennas to scan the 12 tags attached to the human joints. The sampling rate for each tag is thus very low. From the calibrated phase variation data in Fig. 4, we can see that this antenna only collects 38 samples for that tag in 500 time slots, while ideally we expect 500 samples. This means more than 90% of the data are missing for this tag. Learning features from such sparse datasets is highly challenging, and we should estimate the missing samples for more effective learning.

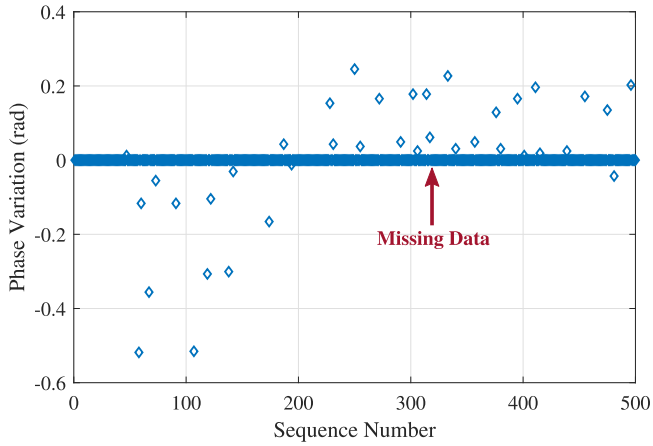


Fig. 4. Calibrated phase variation data from one of the RFID tags (the raw data is plotted in Fig. 2).

1) Downsampling and Synchronization: With N_p antennas and N_q tags, we can create a $N_p \times N_q$ phase variation matrix for all the tags and antennas and extend it into an order-3 tensor structure for various time slots. The data tensor for N_p antennas, N_q tags, and N_t time slots is constructed as

$$\boldsymbol{\psi}(:,:,q) = \begin{bmatrix} \phi_{q1}^1 & \phi_{q2}^1 & \cdots & \phi_{qN_t}^1 \\ \phi_{q1}^2 & \phi_{q2}^2 & \cdots & \phi_{qN_t}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{q1}^{N_p} & \phi_{q2}^{N_p} & \cdots & \phi_{qN_t}^{N_p} \end{bmatrix}, q = 1, 2, \ldots, N_q.$$

In the data tensor, ϕ_{qt}^p represents the calibrated phase variation data from tag q sampled by antenna p in time slot t. Note that only one phase variation can be sampled in each $\boldsymbol{\psi}(:,t,:)$. So only up to N_t samples are nonempty in this $N_p \times N_t \times N_q$ tensor, i.e., it is highly sparse. The RFID-Pose system utilizes 12 tags and 3 antennas. Thus the sparsity of the data tensor is as high as 97.22%, which leads to poor learning performance. However, such highly sparse tensors are very hard to be accurately completed with traditional compressed sensing techniques.

Fortunately, since the frame rate of the Kinect data is 30 fps, we can compress the RFID data in multiple adjacent time slots to match the corresponding, single Kinect data frame. Furthermore, since the requirement on the frame rate is not very high for human pose tracking (which mostly involve slow body movements), we can further downsample the Kinect data so that more slices in the sparse tensor can be grouped into one. If we compress tensor $\boldsymbol{\psi}$ into $\boldsymbol{\Psi}$ with ratio ξ, the new tensor after synchronization could be denoted as

$$\boldsymbol{\Psi}(:,:,q) = \begin{bmatrix} \bar{\phi}_{q1}^1 & \bar{\phi}_{q2}^1 & \cdots & \bar{\phi}_{qN_T}^1 \\ \bar{\phi}_{q1}^2 & \bar{\phi}_{q2}^2 & \cdots & \bar{\phi}_{qN_T}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \bar{\phi}_{q1}^{N_p} & \bar{\phi}_{q2}^{N_p} & \cdots & \bar{\phi}_{qN_T}^{N_p} \end{bmatrix}, q = 1, 2, \ldots, N_q$$

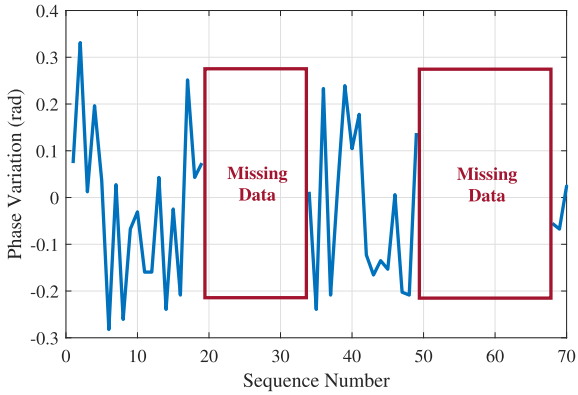


Fig. 5. Downsampled and synchronized RFID phase variation from one RFID tag with $\xi = 50$.

where N_T is the number of synchronized time slots for RFID data, which is the same as the number of downsampled Kinect data units. As the equation shows, for each unit $\mathbf{\Psi}(n_p, n_t, q)$ in the tensor, the first coordinate n_p represents the index of the sampling antenna, the second coordinate n_t indicates the index of the time slot, and the third coordinate q is the index of the attached RFID tag. The tensor structure is also illustrated in the right-hand-side of Fig. 3. In addition, $\bar{\phi}^p_{q,T}$ is the mean phase variation from tag q sampled by antenna p in synchronized time slot T, which is calculated for the ξ adjacent values in ψ as

$$\bar{\phi}^p_{qT} = \frac{1}{\xi} \sum_{t=T}^{T+\xi-1} \phi^p_{qt}. \tag{5}$$

After the downsampling process, the sampling period is also multiplied by ξ. Since phase variation represents the velocity of the overall phase changes, the mean value calculation still keeps the phase variation velocity unchanged. With downsampling and synchronization, the sparsity of the RFID data will be greatly reduced, as illustrated in Fig. 5, which is obtained with $\xi = 50$ from the calibrated phase variation data shown in Fig. 4. As the figure shows, there are now 38 valid data units in 70 time slots. Compared to the original data in Fig. 4, the sparsity is effectively reduced. However, there are still intervals of time with no effective sampled data, which will be addressed next.

2) High Accuracy Low Rank Tensor Completion (HaLRTC): The commodity RFID reader used in RFID-Pose has three antennas. To accurately learn the RFID phase variation features, all tags should be sampled by all antennas in each time slot in the ideal case. However, the phase variations collected from different antennas could be treated as different samples from the same signal source (i.e., tag movement). Since the number of signal sources equals to the number attached RFID tags, the sparse tensor $\mathbf{\Psi}$ can be considered as a low-rank tensor, which can be recovered by low-rank tensor completion. This task is accomplished by solving the following optimization problem [46]:

$$\begin{aligned} \min_{\hat{\mathbf{\Psi}}} \quad & \|\hat{\mathbf{\Psi}}\|_* \\ \text{s.t.:} \quad & \mathbf{\Omega} * \hat{\mathbf{\Psi}} = \mathbf{\Omega} * \mathbf{\Psi} \end{aligned} \tag{6}$$

where $\hat{\mathbf{\Psi}}$ is an estimation of the ideal tensor data $\mathbf{\Psi}_{\text{ideal}}$, which is composed of all the ideal phase variation data; and $\mathbf{\Omega}$ is a tensor of 0 and 1 elements, where $\mathbf{\Omega}_{\text{IJK}} = 1$ when $\mathbf{\Psi}_{\text{IJK}}$ is sampled, and $\mathbf{\Omega}_{\text{IJK}} = 0$ otherwise. In (6), $\|\cdot\|_*$ denotes the trace norm of tensors.

During the optimization procedure, the trace norm of the third-order tensor $\mathbf{\Psi}$ is calculated with the combination of its unfolded matrix in different modes. The optimization problem is represented as [46]

$$\begin{aligned} \min_{\hat{\mathbf{\Psi}}, \boldsymbol{M}_i} \quad & \sum_{i=1}^{3} h_i ||\boldsymbol{M}_i(i)||_* \\ \text{s.t.:} \quad & \mathbf{\Omega} * \hat{\mathbf{\Psi}} = \mathbf{\Omega} * \mathbf{\Psi} \\ & \hat{\mathbf{\Psi}} = \boldsymbol{M}_i, \ i = 1, 2, 3 \end{aligned} \tag{7}$$

where h_i's are constants satisfying $\sum_{i=1}^{3} h_i = 1$, $\boldsymbol{M}_i$ is a tensor with the same size as $\hat{\mathbf{\Psi}}$, and $\boldsymbol{M}_i(i)$ is the matrix unfolded from tensor $\boldsymbol{M}_i$ in mode i. The equation shows that the trace norm of a tensor is a convex combination of norms for all matrices unfolded along each mode. In HaLRTC, the optimization problem (7) is solved with the augmented Lagrange multiplier method (ADMM) [47] with the augmented Lagrangian function defined as

$$\begin{aligned} & \boldsymbol{L}_\rho(\hat{\mathbf{\Psi}}, \boldsymbol{M}_i, \boldsymbol{Y}_i) \\ & = \sum_{i=1}^{3} h_i ||\boldsymbol{M}_i(i)||_* + \left\langle \hat{\mathbf{\Psi}} - \boldsymbol{M}_i, \boldsymbol{Y}_i \right\rangle + \frac{\rho}{2} ||\boldsymbol{M}_i - \hat{\mathbf{\Psi}}||_F^2 \end{aligned} \tag{8}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product of two tensors and $||\cdot||_F$ is the Frobenius norm of the tensor; $\boldsymbol{Y}_i$ is a zero tensor with the same size as $\hat{\mathbf{\Psi}}$, and $\rho > 0$ is the penalty factor in the algorithm. In our system we set $\rho = 1e^{-4}$. Rather than iterate recursively to optimize the target tensor $\hat{\mathbf{\Psi}}$. ADMM literately updates multiple variables, i.e., $\boldsymbol{M}_i$, $\hat{\mathbf{\Psi}}$, and $\boldsymbol{Y}_i$ as follows:

1) $\boldsymbol{M}'_i = \arg\min(\boldsymbol{M}_i) : \boldsymbol{L}_\rho(\hat{\mathbf{\Psi}}, \boldsymbol{M}_i, \boldsymbol{Y}_i)$
2) $\hat{\mathbf{\Psi}}' = \arg\min(\hat{\mathbf{\Psi}}) : \boldsymbol{L}_\rho(\hat{\mathbf{\Psi}}, \boldsymbol{M}'_i, \boldsymbol{Y}_i)$
3) $\boldsymbol{Y}'_i = \boldsymbol{Y}_i - \rho(\boldsymbol{M}'_i - \hat{\mathbf{\Psi}}')$.

These functions converge when the update between two adjacent iteration is sufficiently small. Thus, the update threshold is set to determine whether $\hat{\mathbf{\Psi}}$ is successfully estimated or not. To balance the data imputation performance and the convergence rate of the algorithm, we set the convergence threshold to $1e^{-6}$ to make sure the data are effectively recovered with an acceptable convergence rate. Compared with other low-rank tensor completion algorithms, HaLRTC can solve the optimization problem (6) more accurately with a lower complexity. The entire tensor completion process in our system only takes less than 0.1 s to execute because the downsampling reduces the input tensor size. As illustrated in Fig. 6, all the missing data can be effectively estimated by HaLRTC. So the reconstructed tensor $\hat{\mathbf{\Psi}}$ can be used by the deep learning model for 3-D human pose estimation.

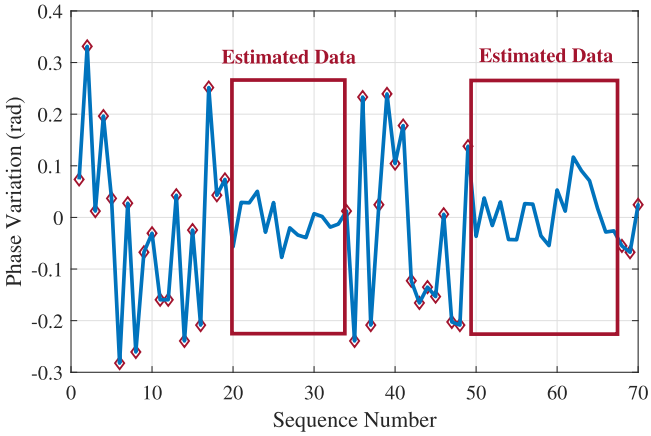


Fig. 6. Missing data are estimated by HaLRTC.

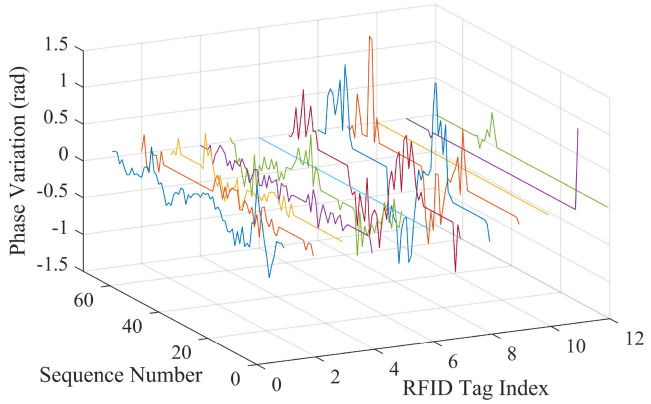


Fig. 7. Sparse RFID phase variation matrix collected from one antenna.

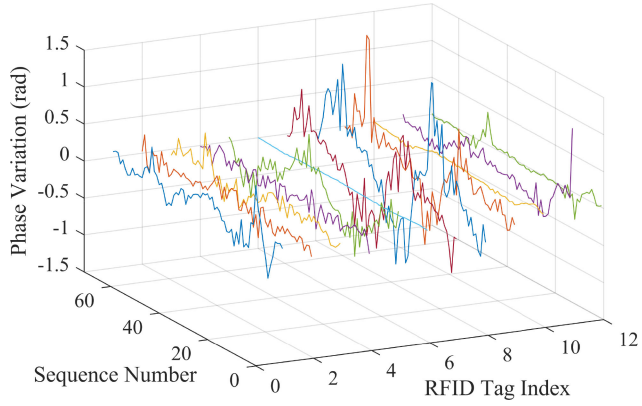


Fig. 8. Phase variation matrix completed by HaLRTC.

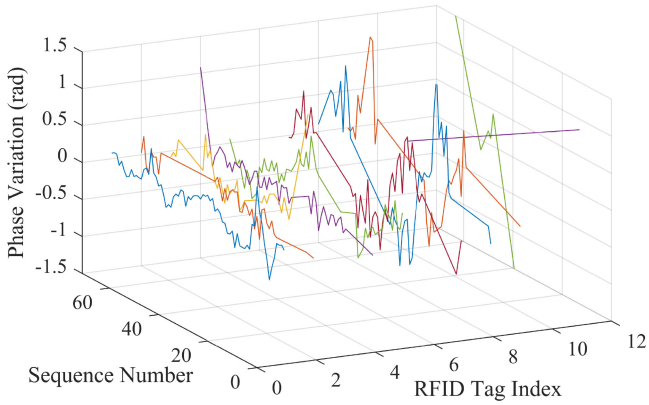


Fig. 9. Phase variation matrix completed by the bilinear interpolation method.

To evaluate the performance of the HaLRTC algorithm, we compare compare it with a conventional interpolation method, i.e., the bilinear interpolation technique. Fig. 7 shows one slice of phase variation data in tensor $\hat{\boldsymbol{\Psi}}$, which represents the synchronized phase variation data for all tags sampled by one antenna. As the figure shows, there are still many samples of value 0, indicating that most data are still missing after downsampling, especially for tags 6, 10, 11, and 12. Both HaLRTC and bilinear interpolation techniques are used to interpolate the miss samples, and the results are presented in Figs. 8 and 9, respectively. From Figs. 8 and 9, it can be seen that the phase variation data estimated by tensor completion shows high consistency among all tags, while sharp variations are generated by bilinear interpolation. Especially for the tags with high sparsity, e.g., tags 11 and 12, significant distortions have been introduced by bilinear interpolation, which will cause considerable skeleton estimation errors.

The superior performance of tensor completion in data imputation is mainly because the data are not evenly sampled in the RFID system. The sampled data from different tags usually have highly different sparsity (e.g., tag 1 versus tag 11 or 12 in Fig. 7). The traditional interpolation method is not suitable for this significant uneven sparsity situation. However, by solving the optimization problem (6), the missing samples can be interpolated based on the low rank components of the tensor data, which indicates the movement of the subject. In addition, the tensor completion process in our system only takes less than 0.1 s to execute because the downsampling has reduced the input tensor size. Thus, HaLRTC is a well-suited method for phase variation data imputation in RFID-Pose.

V. Challenges and Solutions: Human Pose Reconstruction With RFID Data

A. Challenges in RFID-Based Human Pose Tracking

Tracking multiple joints of a human subject simultaneously with RFID tags is highly challenging, because the data rate of RFID systems is extremely low comparing to other wireless systems. According to the RFID Gen2 protocol, the medium access control (MAC) in RFID system follows the slotted ALOHA protocol, which means only one tag can respond to the reader in each time slot. Such a transmission scheme makes the data rate of RFID much lower than other sensing systems such as video camera [1], WiFi [10], and FMCW radar [12]. In these RF-based skeleton tracking system, the human skeleton is extracted from the confidence map of the target joints, which is usually generated by a neural network. The RFID system's sampling rate is about 110 Hz for each antenna. In order to generate a 100×100 confidence map to localize the joint positions at a rate of 5 fps (frames/second), only 22 phase data samples can be obtained

for each frame. Recovering a map with 10 000 data samples with only 22 phase data samples is a severely *ill-posed problem*, which is extremely challenging to solve even with advanced deep learning techniques.

The above ill-posed problem implies that the confidence map method may not be suitable to estimate the 3-D pose of the human body. Consequently, the existing RFID-based techniques mostly focus on estimating the movement of a particular limb movement, such as the front arm, the front leg, and thighs [14], [15]. Although, theoretically, the entire body movement could be reconstructed by combing all the limb movements, these systems may not be effective for realtime human pose estimation, especially when multiple moving joints need to be tracked simultaneously.

In RF-wear [15], two RFID tag arrays are attached to the two adjacent limbs of the subject, which are then used to estimate the rotation angle of human limbs with good accuracy. However, when tracking multiple limbs simultaneously, every limb should be attached with an RFID array. In this scenario, there will be a large number of tags to be interrogated by the RFID reader. The severe mutual coupling effect and considerable intertag collisions will cause a lot of missing samples and some tags may even be hardly sampled by the reader. Similarly, in the RF-Kinect system [14], the rotation angle of one particular limb is estimated by the RF hologram technique [21]. Unfortunately, since the angle estimation is based on the probability distribution map built on the phase value of all attached tags, the accuracy of angle estimation could be affected when multiple tags are moving together. Moreover, the generation of the probability distribution map for each joint requires phase measurements for all possible rotation angles, which entail heavy calibration work.

Studying existing RFID-based pose tracking systems, we found that, although generating the skeleton confidence map is challenging, the rotation angles of all human limbs could be relatively easily estimated from the scarce RFID data. This is because, when the limb's length is known, the system only needs to generate three angle values to reconstruct the particular limb's movement. That is, only $3n$ angle values need to be estimated when tracking n joints, which is considerably less than the number of samples required for confidence map generation, and is highly suited for RFID-based sensing systems with constrained sampling rates. Accordingly, our goal is to estimate the rotation angle of each limb and leverage the forward kinematic technique to reconstruct the human skeleton with the estimated rotation angles.

B. Forward Kinematics

The technique to generate human 3-D pose from limb rotation angles is *Forward Kinematics*, which is widely used in robotics and 3-D animation [48]. An example of forward kinematic is shown in Fig. 10. The left-hand-side figure shows a human skeleton with a "T" pose, and the 12 joints with marked numbers are the target joints to track in our RFID-pose system. In forward kinematics, the 3-D position of a joint is generated by 1) the rotation angle of the limb connecting the two joints; and 2) the length of the limb, and 3) the position of its parent joint, which

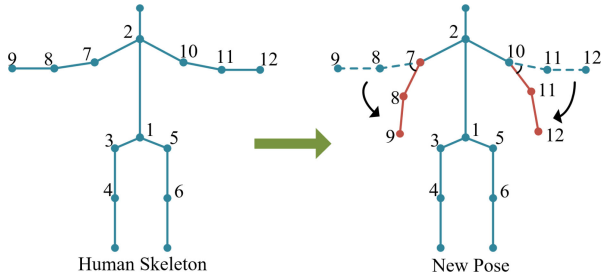


Fig. 10. Example of limb rotation in the human skeleton.

is defined as the rotation anchor. For example, in Fig. 10, the subject puts down his/her arms. Then joints 8, 9, 11, and 12 all move downward. Since joint 7 (i.e., the left shoulder) is the rotation anchor of the left upper arm, it is considered as the parent joint of joint 8 (i.e., the left elbow). The position of joint 8 can be calculated with the length of the upper arm and the 3-D rotation angle. Similarly, the locations of joints 9, 11, and 12 can be estimated from their corresponding parent joints 8, 10, and 11, and the 3-D rotation angles, respectively. Accordingly, once the initial skeleton is given (i.e., the original locations of all joints and the lengths of all limbs), each joint can be localized recursively based on the position of its parent joint and rotation angles.

The recursive rotation for the nth joint in time slot T can be expressed as

$$\vec{P}_n^T = \vec{P}_{\text{parent}(n)}^T + \mathbf{R}_n^T \vec{P}_{\text{relative}(n)}^0 \tag{9}$$

where $\vec{P}_n^T$ represents the position of joint n of time slot T, $\vec{P}_{\text{parent}(n)}^T$ denotes the position of joint n's parent joint, $\mathbf{R}_n^T \in SO(3)$ represents the corresponding rotation matrix ($SO(3)$ denotes the 3-D rotation group), and $\vec{P}_{\text{-}(n)}^0$ is the 3-D offset of joint n relative to its parent joint, given by

$$\vec{P}_{\text{relative}(n)}^0 = \vec{P}_n^0 - \vec{P}_{\text{parent}(n)}^0 \tag{10}$$

where $\vec{P}_n^0$ and $\vec{P}_{\text{parent}(n)}^0$ represent the positions of joint n and its parent joint in the initial 3-D skeleton, respectively. From (9), we can see that, with the initial skeleton data, all joint positions can be calculated by the corresponding rotation matrix $\mathbf{R}_n^T$.

According to Euler's rotation theorem, a 3-D rotation can be represented as a *unit quaternion* in the system with format

$$\ell + x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}. \tag{11}$$

In the unit quaternion ℓ, x, y, and z are real numbers, and $\boldsymbol{i}, \boldsymbol{j}$, and $\boldsymbol{k}$ are quaternion units. Given a 3-D position vector represented as $a\boldsymbol{i} + b\boldsymbol{j} + c\boldsymbol{k}$ and a 3-D rotation with unit quaternion $r_\ell + r_x\boldsymbol{i} + r_y\boldsymbol{j} + r_z\boldsymbol{k}$. The rotation matrix $\mathbf{R}$ is derived as

$$\mathbf{R} = \begin{bmatrix} 1-2({r_y}^2+{r_z}^2) & 2(r_x r_y + r_z r_\ell) & 2(r_x r_z - r_y r_\ell) \\ 2(r_x r_y - r_z r_\ell) & 1-2({r_x}^2+{r_z}^2) & 2(r_y r_z + r_x r_\ell) \\ 2(r_x r_z + r_y r_\ell) & 2(r_y r_z - r_x r_\ell) & 1-2({r_x}^2+{r_y}^2) \end{bmatrix}. \tag{12}$$

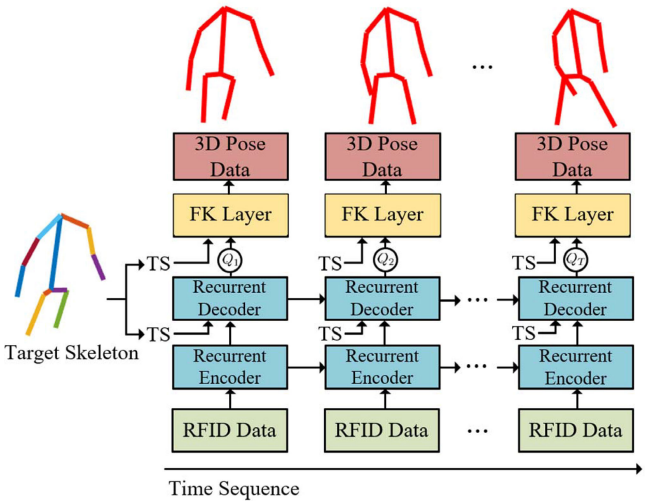


Fig. 11. Deep kinematic neural network incorporated in RFID-Pose.

The new position vector, after the 3-D rotation, can be calculated as

$$\begin{bmatrix} a' \\ b' \\ c' \end{bmatrix} = \mathbf{R} \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \tag{13}$$

The rotation matrix $\mathbf{R}$ is used in the forward kinematic (FK) layer of the learning model in the RFID-Pose system, which is to reconstruct the human 3-D pose with the initial skeleton and the corresponding spatial rotations.

C. Deep Kinematic Neural Network

To reconstruct 3-D human pose, we leverage a deep kinematic neural network to learn the features of RFID phase variation collected when the subject is moving. The structure of the learning model is illustrated in Fig. 11. The offline training goal is to learn the relationship between the RFID phase variation and the rotation of the human limbs. The 3-D pose ground truth obtained from Kinect is in the form of 3-D coordinates for the human joints. The initial target skeleton is required for each training dataset to transform the estimated rotation angle to the 3-D positions through forward kinematic.

As Fig. 11 shows, the deep kinematic neural network is mainly composed of two parts, i.e., the recurrent autoencoder and the forward kinematic layer. The recurrent neural network (RNN) is suitable for learning the features of phase variation sampled in a time sequence, while the Autoencoder is a simple but effective learning model to extract the features of RFID phase data [29], [30]. The input training data are the RFID phase variation sequence and the 3-D pose data sequence, which are synchronized after data preprocessing (see the previous section).

The recurrent autoencoder consists of two key parts, an encoder and a decoder. In each time slot, the features in the input RFID phase data are first extracted by the recurrent encoder and stored in the hidden layers, which consist of 256 gated recurrent units (GRU). Because of the recurrent structure, the hidden layer outputs in the previous time slot are also fed to the following Encoder. Thus, the recurrent encoder can extract feature of the RFID phase data from both the current time slot and previous time slots. Then the recurrent decoder is leveraged to transfer the extracted feature stored in the encoder hidden layer to 3-D rotation data. Since the limb length data are required for the 3-D rotation estimation from extracted RFID feature, the initial human skeleton should be added as another input to the decoder. Moreover, the recurrent structure also feeds the previous hidden layer outputs to the current decoder for learning the features in the output data sequence. The unit quaternion Q_T for each joint is obtained by normalizing the recurrent decoder output.

Next, with the initial skeleton and Q_T, the forward kinematic layer leverages the rotation matrix $\mathbf{R}$ to generate 3-D coordinates for the subject, which are in the same format as the Kinect ground truth data. With the error calculated between the estimated pose and the ground truth, the weights in the recurrent autoencoder will be trained by using error backpropagation.

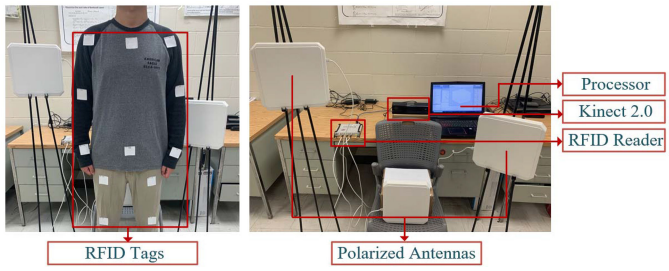


Fig. 12. Illustration of the system setup for 3-D pose estimation.

VI. Implementation and Evaluation

A. System Implementation

To evaluate the performance of the RFID-pose system, we develop a prototype system with an off-the-shelf Impinj R420 reader equipped with three S9028PCR polarized antennas. The RFID tags used for tracking human joint movements are ALN-9634 (HIGG-3). The vision data used for training supervision and test accuracy evaluation are collected with an Xbox Kinect 2.0 device. The sampling rate of the RFID phase data are around 110 Hz, and the frame rate of the Kinect 2.0 is 30 fps. All data are downsampled to 7.5 Hz after preprocessing and synchronization. The length of the RFID input tensor N_T is set to 30 during the experiments, which represents 4 s motion data.

The setup of the system is illustrated in Fig. 12. As the figure shows, we attach RFID tags to the 12 joints of the human body, which are the pelvis, neck, left hip, left knee, right hip, right knee, left shoulder, left elbow, left wrist, right shoulder, right elbow, and right wrist. To each joint, *one* passive RFID tag is attached to monitor the joint movement. The head and feet are omitted in our prototype system because of the limited scanning range of the RFID antenna used. The antennas are placed at different altitude positions to ensure that the antennas can interrogate all the tags. If we want to scan all the joints from head to feet, more antennas should be used in the system. However, the pose with the 12 joints is sufficient to monitor human behavior in most cases.

Fig. 13. Illustration of two example poses. (Left) Standing still. (right) Walking.

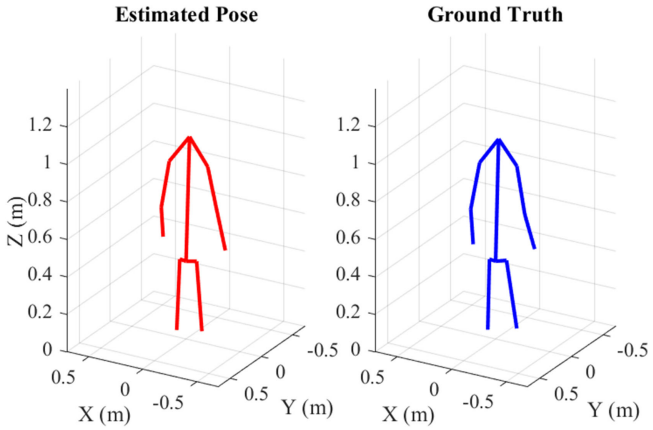


Fig. 14. Pose estimation when the subject is standing still.

An MSI laptop with a Nvidia GTX 1080 GPU and an Intel Core i7-6820HK CPU is used as the processor for data training and signal processing. The frequency used by the prototype system hops among 50 channels from 902 to 928 MHz, and it remains on a channel for 0.2 s.

B. Performance Evaluation and Results

1) Overall Accuracy for Different Motions: We train the proposed deep kinematic neural network with different types of motions. The first type of motions is simple motion, which is only involved with the movement of a single limb. The second type of motions is complicated motion, which is composed of movements of the entire body, such as body twisting, deep squat, boxing, and walking. Two examples of the motions are illustrated in Fig. 13. The left-hand-side figure shows a subject simply standing still, and the right-hand-side figure shows the subject is walking. The estimation results for these two examples are presented in Figs. 14 and 15, respectively, where the estimated pose is marked with red lines, and the Kinect obtained ground truth is marked with blue lines. We also present the estimation results for other complicated motions, including squat, twisting, and kicking, in Figs. 16, 17, and 18, respectively. From these figures, we can see that the estimated poses are all highly close to the ground truth collected by Kinect. These example results show that the RFID-Pose system can adequately estimate the 3-D human pose whether the subject is moving or not.

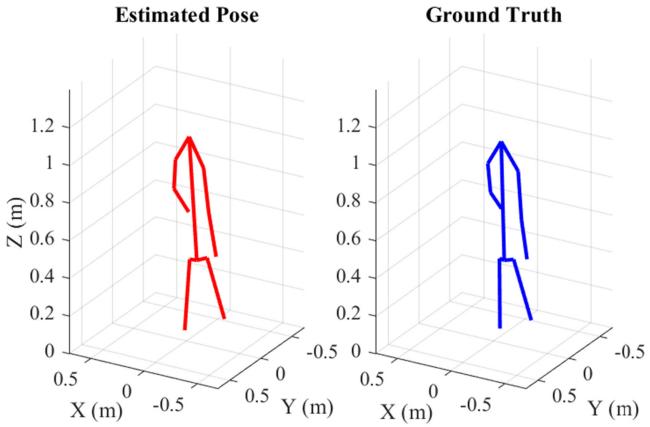


Fig. 15. Pose estimation when the subject is walking.

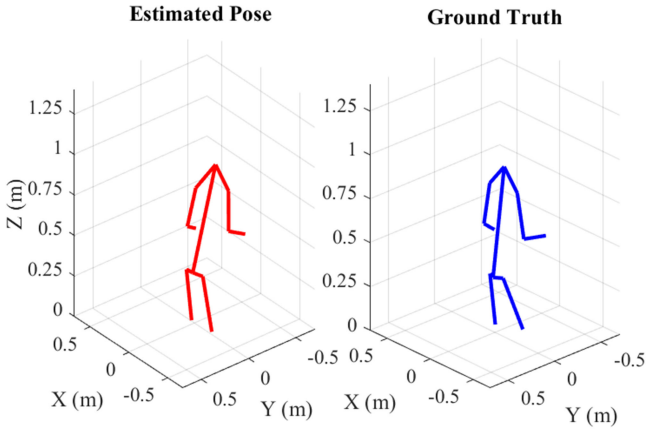


Fig. 16. Pose estimation when the subject is squatting.

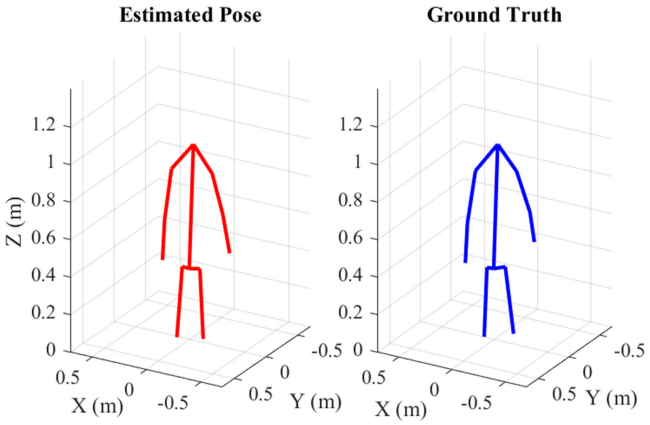


Fig. 17. Pose estimation when the subject is twisting.

The overall accuracy of human pose estimation is presented in the form of cumulative distribution function (CDF) of estimation errors in Fig. 19. The mean error of all the 12 joints for each time slot T is calculated as follows:

$$\epsilon(T) = \frac{1}{12}\sum_{n=1}^{12} ||\hat{P}_n^T - \dot{P}_n^T|| \tag{14}$$

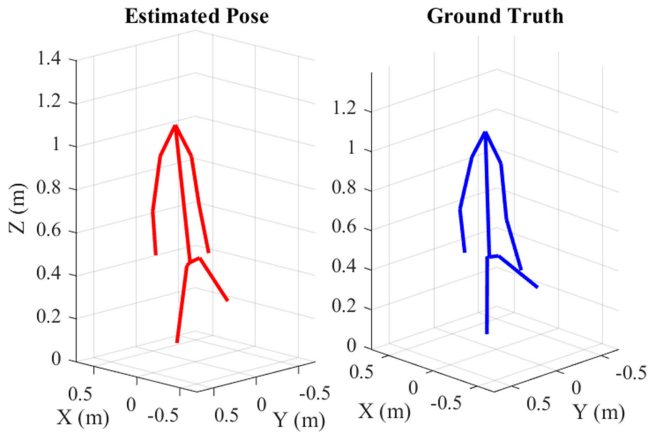


Fig. 18. Pose estimation when the subject is kicking.

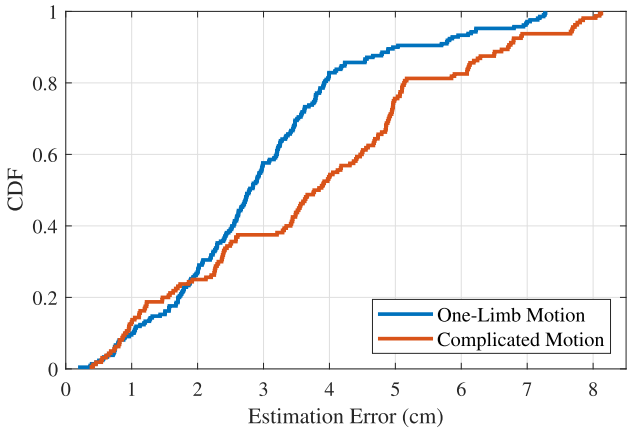


Fig. 19. Overall pose estimation accuracy in forms of CDF of estimation errors.

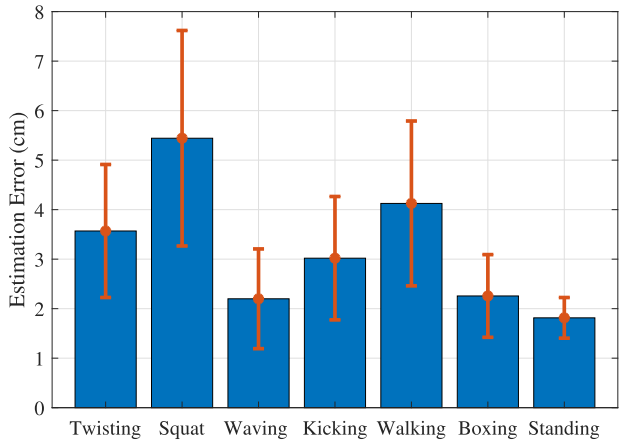


Fig. 20. Estimation errors for different types of motions.

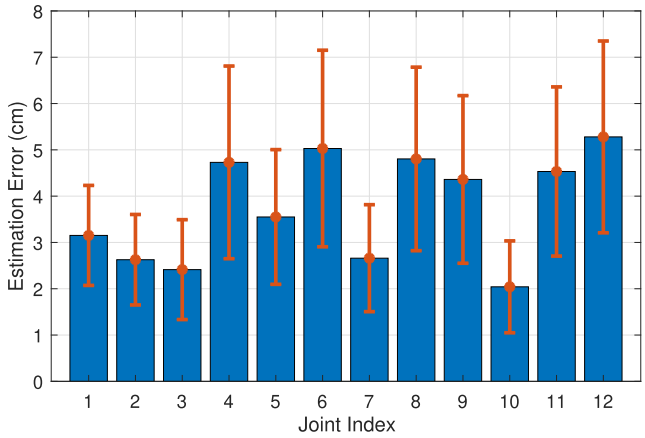


Fig. 21. Estimation errors for different joints.

where $\hat{P}_n^T$ denotes the estimated position and $\dot{P}_n^T$ is the ground truth position collected by the Kinect in the 3-D space for joint n at time T; and $||\hat{P}_n^T - \dot{P}_n^T||$ is the Euclidean distance between these two 3-D vectors. From the CDF curves, we can see that the median estimation error is 2.83 cm for the single-limb motion test and 3.75 cm for the complicated motion test. The results show that the estimation accuracy of the entire body motion is lower than one-limb motion, because more moving joints need to be reconstructed in the former case. However, RFID-Pose still achieves very high accuracy for all the complicated motions, and the largest error among all the tests is 8.12 cm, which is smaller than the maximum estimation error reported in the existing RFID pose estimation system (i.e., 10 cm) [14]. The estimation results validates that the proposed RFID-Pose system can estimate the joints position more accurately and can effectively reconstruct the pose of the entire moving body through RFID phase data.

2) Accuracy for Different Motions: To evaluate the estimation performance for different motions, we plot the accuracy for all the specific movements in Fig. 20, including body twisting, squat, waving hands, kicking, walking, boxing, and standing still. As the figure shows, the pose estimation accuracy is different for different motions, where the highest accuracy 1.81 cm is achieved when the human is in a stable state (i.e., standing still). This is because no joint is moving when the subject stands still, and thus no joint movements need to be estimated in this case.

We also notice that the squat and walking motions have worse estimation accuracy than others, which are 5.44 and 4.12 cm, respectively. The pelvis joint position variation is the main cause for the limited performance. Note that our network is designed for learning the spatial rotation of each joint relative to the parent joint. As a root joint of the human skeleton, the pelvis position estimation does not benefit from the forward kinematic layer. Thus, the pelvis joint's position is not as accurate as the rotation angle for each human limb, which also leads to higher errors in all other joints. That is the reason for the lower accuracy when the pelvis joint frequently varies during the monitoring process. Nevertheless, the error 5.44 cm is still acceptable for most pose-based applications, such as video gaming and motion recognition.

3) Accuracy for Different Joints: The estimation error for each of the 12 joints is presented in Fig. 21. The joint index map is shown in Fig. 10. From joint 1 to joint 12, the joints are: Pelvis, neck, left hip, left knee, right hip, right knee, left shoulder, left elbow, left wrist, right shoulder, right elbow, and right wrist. As the figure shows, RFID-Pose achieves high estimation accuracy for joints 1, 2, 3, 5, 7, and 10, where the estimation errors are

TABLE II
PERFORMANCE EVALUATION FOR DIFFERENT SUBJECTS

Subject Index	*Estimation Error*
Subject 1 (trained)	3.72cm
Subject 2 (trained)	4.55cm
Subject 3 (trained)	3.58cm
Subject 4 (untrained)	5.32cm
Subject 5 (untrained)	8.17cm

all lower than 3.55 cm. The estimation errors for the other joints are all higher than 4.36 cm. This is because the joints in the first group are on or close to the human torso, while the other joints are on the limbs (i.e., arms and legs). The relatively worse limbs tracking performance is mainly due to two reasons. First, since the joints of the limbs are tracked based on the torso joints with the forward kinematic technique, the estimation errors of the parent joints on the torso will be accumulated and affect the accuracy of tracking the limb joints. However, the pelvis localization in each time slot is independent, and the estimation error of the pelvis in previous time slots will not be accumulated in the present time slot. Second, since human limbs usually move at a larger extent than the torso joints, there are usually fewer RFID samples for these joints, which leads to a higher estimation error. However, notice that even the wrist estimation error, the highest one, is lower than 5.28 cm. Such results prove that the RFID-Post system can accurately estimate the human pose with the vision-aided technique.

C. More Experiments Under Different Scenarios

In addition to evaluating the overall accuracy, we conduct several additional experiments to test the system performance under different scenarios, including different subjects, different environments, and different standing positions in front of the antennas. We also discuss the generalization issue based on the experimental results.

1) Different Subjects: We conduct experiments with five different subjects to examine the impact of different initial skeletons. The training dataset includes three different subjects, while the other two subjects are not trained but for testing only. The mean estimation errors are presented in Table II. As the table shows, the estimation errors for all the trained subjects are lower than 4.55 cm, which means the system can estimate the human skeleton for different subjects. However, when the trained system is used to test the untrained subjects, i.e., subjects 4 and 5, the performance becomes worse but still acceptable. Furthermore, we find that the accuracy for subject 4 is higher than subject 5 because the initial skeleton of subject 4 is similar to trained subject 2. It implies that the performance of testing untrained subjects could be improved when the network is trained with more subjects with different skeleton patterns.

2) Different Environments and Standing Positions: The influence of different environments and standing positions are also investigated. The experiments are conducted in four different environments, including two different locations in the lab, a corridor, and a living room. The first three environments are illustrated in Fig. 22. As the figure shows, the first two locations

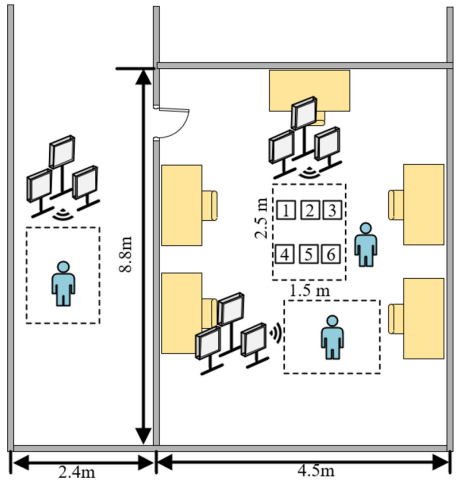


Fig. 22. Different deployment environments and standing positions.

TABLE III
PERFORMANCE EVALUATION UNDER DIFFERENT ENVIRONMENTS

Testing Environments	*Estimation Error*
Computer Lab-1	3.83cm
Computer Lab-2	3.90cm
Corridor	4.03cm
Living Room	3.75cm

TABLE IV
PERFORMANCE EVALUATION FOR DIFFERENT STANDING POSITIONS

Position Index	*Estimation Error*
Position 1 (Trained)	4.53cm
Position 2 (Trained)	3.82cm
Position 3 (Trained)	4.75cm
Position 4 (Untrained)	8.38cm
Position 5 (Untrained)	5.71cm
Position 6 (Untrained)	9.14cm

are selected in the same lab but have highly different deployments, to introduce different environmental interference. The other two locations are selected in the corridor and living room, respectively, which also suffers from quite different multipath effects. As Table III shows, the estimation error in different environments changes from 3.75 to 4.03 cm, which means the influence of the environments is limited. This is because the received RFID signal is dominated by the line-of-sight component; the other reflected signals are very weak. Thus, the multipath effect from the environment is not strong and does not affect much the performance of RF-Pose.

The interference of different stand positions is also investigated in our experiments. As illustrated in Fig. 22, we compare the system performance for six different positions in the 2.5×1.5 m scanning area in the Lab scenario. Data collected in positions 1, 2, and 3 are used to train the system, while the data collected in positions 4, 5, and 6 are only used for testing. The estimation errors are presented in Table IV. As the table shows, the estimation errors for the three untrained positions 4, 5, and 6 are all higher than 5.71 cm, while the errors for the three trained positions 1, 2, and 3 are all lower than 4.75 cm. The results show

that the estimation accuracy degrades when the subject stands in an untrained position, especially the untrained position near the border of the scanning area. Fortunately, the high accuracy for the trained standing positions shows that the accuracy of untrained positions could be improved by adding more training data sampled from different training positions. Due to limited scanning range of the polarized antennas, six different standing positions for training are sufficient to combat the influence of untrained standing positions.

3) Remarks on Generalization: Since the initial subject skeleton is needed in the training process, the performance of the proposed system could be affected when testing the subject with an untrained subject or the subject is tested in a different standing position/environment. In RFID-Pose, the initial skeleton is also necessary to address the ill-posed problem caused by the low data rate of RFID systems. This article is mainly focused on the fundamental problem of transferring sparse RFID data to 3-D human skeleton. However, the experiment results shown in Tables II and IV also demonstrate that the generalization issue could be mitigated by extending the training dataset for different subjects and standing positions. We will further tackle the generalization problem of RFID-based pose monitoring systems in our future work.

VII. Conclusion

In this article, we proposed a vision-aided, realtime 3-D pose estimation and tracking system named RFID-Pose. A preprocessing module was proposed to effectively mitigate the influence of phase distortion and missing samples in the RFID data. The proposed system then leveraged a deep kinematic network to estimate human postures in realtime from RFID phase data, which was trained with the assistance of CV data as labels collected by Kinect 2.0. The RFID-pose system was prototyped with commodity RFID devices. Its high accuracy and realtime operation were demonstrated in our experimental study using Kinect 2.0 as a benchmark.

References

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 7291–7299.

[2] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 623–630.

[3] Tom's Guide, "Millions of wireless security cameras are at risk of being hacked: What to do," 2020. Accessed: Jun. 20, 2020). [Online]. Available: https://www.tomsguide.com/news/hackable-security-cameras

[4] P. A. Laplante and J. F. DeFranco, "Software engineering of safety-critical systems: Themes from practitioners," *IEEE Trans. Rel.*, vol. 66, no. 3, pp. 825–836, Sep. 2017.

[5] S. Siboni *et al.*, "Security testbed for Iternet-of-Things devices," *IEEE Trans. Rel.*, vol. 68, no. 1, pp. 23–44, Mar. 2019.

[6] M. Noor-A-Rahim, M. Khyam, G. M. N. Ali, Z. Liu, D. Pesch, and P. H. Chong, "Reliable state estimation of an unmanned aerial vehicle over a distributed wireless IoT network," *IEEE Trans. Rel.*, vol. 68, no. 3, pp. 1061–1069, Sep. 2019.

[7] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Trans. Rel.*, vol. 62, no. 2, pp. 434–443, Jun. 2013.

[8] X. Yang, K. Tang, and X. Yao, "A learning-to-rank approach to software defect prediction," *IEEE Trans. Rel.*, vol. 64, no. 1, pp. 234–246, Mar. 2014.

[9] M. Liu, L. Miao, and D. Zhang, "Two-stage cost-sensitive learning for software defect prediction," *IEEE Trans. Rel.*, vol. 63, no. 2, pp. 676–686, Jun. 2014.

[10] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-WiFi: Fine-grained person perception using WiFi," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seoul, Republic of Korea, Oct. 2019, pp. 5452–5461.

[11] W. Jiang *et al.*, "Towards 3D human pose construction using WiFi," in *Proc. ACM MobiCom'20*, London, UK, Sep. 2020, Art. no. 23.

[12] M. Zhao *et al.*, "Through-wall human pose estimation using radio signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7356–7365.

[13] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs," *IEEE Sensors J.*, vol. 20, no. Sep., pp. 10 032–10 044, Sep. 2020.

[14] C. Wang, J. Liu, Y. Chen, L. Xie, H. B. Liu, and S. Lu, "RF-Kinect: A wearable RFID-based approach towards 3D body movement tracking," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, Mar. 2018, Art. no. 41.

[15] H. Jin, Z. Yang, S. Kumar, and J. I. Hong, "Towards wearable everyday body-frame tracking using passive RFIDs," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, Dec. 2018, Art. no. 145.

[16] C. Yang, X. Wang, and S. Mao, "SparseTag: High-precision backscatter indoor localization with sparse RFID tag arrays," in *Proc. 16th Annu. IEEE Int. Conf. Sens., Commun., Netw.*, Boston, MA, pp. 1–9, Jun. 2019.

[17] X. Wang, X. Wang, and S. Mao, "RF sensing for Internet of Things: A general deep learning framework," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 62–69, Sep. 2018.

[18] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Elsevier Comput. Vis. Image Underst.*, vol. 192, no. 3, Mar. 2020, Art. no. 102897.

[19] J. Zhang, S. Periaswamy, S. Mao, and J. Patton, "Standards for passive UHF RFID," *ACM GetMobile*, vol. 23, no. 3, pp. 10–15, Sep. 2019.

[20] "Low level user data support," Impinj, Seattle, WA, USA, Appl. note, 2019. [Online]. Available: https://support.impinj.com/hc/en-us/article_attachments/200774268/SR_AN_IPJ_Speedway_Rev_Low_Level_Data_Support_20130911.pdf

[21] L. Yang, Y. Chen, X.-Y. Li, C. Xiao, M. Li, and Y. Liu, "Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, Maui, HI, Sep. 2014, pp. 237–248.

[22] J. Wang and D. Katabi, "Dude, where's my card? RFID positioning that works with multipath and non-line of sight," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 51–62, Oct. 2013.

[23] L. Shangguan and K. Jamieson, "The design and implementation of a mobile RFID tag sorting robot," in *Proc. 14th Annu. Int. Conf. Mobile Syst.*, Singapore, Jun. 2016, pp. 31–42.

[24] Y. Ma, N. Selby, and F. Adib, "Minding the billions: Ultra-wideband localization for deployed RFID tags," in *Proc. 23rd Annu. Int. Conf. Mobile Comput. Netw.*, Snowbird, Utah, Oct. 2017, pp. 248–260.

[25] Y. Hou, Y. Wang, and Y. Zheng, "Tagbreathe: Monitor breathing with commodity RFID systems," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst.*, Atlanta, GA, USA, Jun. 2017, pp. 404–413.

[26] R. Zhao, D. Wang, Q. Zhang, H. Chen, and A. Huang, "CRH: A contactless respiration and heartbeat monitoring system with COTS RFID tags," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw.*, Hong Kong, China, pp. 1–9, Jun. 2018.

[27] C. Wang, L. Xie, W. Wang, Y. Chen, Y. Bu, and S. Lu, "RF-ECG: Heart rate variability assessment based on cots RFID tag array," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 2, Jun. 2018, Art. no. 85.

[28] C. Yang, X. Wang, and S. Mao, "AutoTag: Recurrent vibrational autoencoder for unsupervised apnea detection with RFID tags," in *Proc. IEEE Global Commun. Conf.*, Abu Dhabi, UAE, pp. 1–7, Dec. 2018.

[29] C. Yang, X. Wang, and S. Mao, "Unsupervised detection of apnea using commodity RFID tags with a recurrent variational autoencoder," *IEEE Access J.*, vol. 7, no. 1, pp. 67 526–67 538, Jun. 2019.

[30] C. Yang, X. Wang, and S. Mao, "Unsupervised drowsy driving detection with RFID," *IEEE Trans. Veh. Technol.*, vol. 69, no. 8, pp. 8151–8163, Aug. 2020.

[31] C. Yang, X. Wang, and S. Mao, "Respiration monitoring with RFID in driving environments," *IEEE J. Sel. Areas Commun.*, to be published.

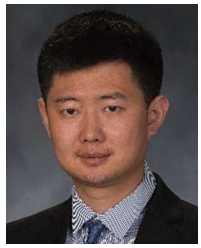
[32] C. Zhao, Z. Li, T. Liu, H. Ding, J. Han, W. Xi, and R. Gui, "RF-Mehndi: A fingertip profiled RF identifier," in *Proc. IEEE Conf. Comput. Commun.*, Paris, France, Jun. 2019, pp. 1513–1521.

[33] J. Wang, J. Xiong, X. Chen, H. Jiang, R. K. Balan, and D. Fang, "TagScan: Simultaneous target imaging and material identification with commodity RFID devices," in *Proc. 23rd Annu. Int. Conf. Mobile Comp. Netw.*, Snowbird, UT, USA, Oct. 2017, pp. 288–300.
[34] T. Wei and X. Zhang, "Gyro in the air: Tracking 3D orientation of batteryless internet-of-things," in *Proc. 22nd Annu. Int. Conf. Mobile Comput. Netw.*, New York City, NY, USA, Oct. 2016, pp. 55–68.
[35] P. Li, Z. An, L. Yang, and P. Yang, "Towards physical-layer vibration sensing with RFIDs," in *Proc. IEEE Conf. Comp. Commun.*, Paris, France, Jun. 2019, pp. 892–900.
[36] J. Guo, T. Wang, Y. He, M. Jin, C. Jiang, and Y. Liu, "Twinleak: RFID-based liquid leakage detection in industrial environments," in *Proc. IEEE Conf. Comput. Commun.*, Paris, France, Apr. 2019, pp. 883–891.
[37] X. Wang, J. Zhang, Z. Yu, S. Mao, S. Periaswamy, and J. Patton, "On remote temperature sensing using commercial UHF RFID tags," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 715–10 727, Dec. 2019.
[38] J. Zhang *et al.*, "RFHUI: An intuitive and easy-to-operate human-UAV interaction system for controlling a UAV in a 3D space," in *Proc. Int. Conf. Mobile Ubiquitous Syst.: Comput., Netw. Services*, New York City, NY, USA, Nov. 2018, pp. 69–76.
[39] J. Zhang *et al.*, "RFHUI: An RFID based human-unmanned aerial vehicle interaction system in an indoor environment," *Elsevier Digit. Commun. Netw. J.*, vol. 6, no. 1, pp. 14–22, Feb. 2020.
[40] J. Zhang *et al.*, "Robust RFID based 6-DoF localization for unmanned aerial vehicles," *IEEE Access J.*, vol. 7, no. 1, pp. 77 348–77 361, Jun. 2019.
[41] R. Mitra, N. B. Gundavarapu, A. Sharma, and A. Jain, "Multiview-consistent semi-supervised learning for 3D human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 2020, pp. 6907–6916.
[42] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1347–1355.
[43] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, Feb. 2012.
[44] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, no. 1/2, Jul. 2010, Art. no. 4.
[45] M. Zhao*et al.*, "RF-based 3D skeletons," in *Proc. ACM Conf. ACM Special Interest Group Data Commun.*, Budapest, Hungary, Aug. 2018, pp. 267–281.
[46] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.
[47] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," 2013, *arXiv:1009.5055*. [Online]. Available: https://arxiv.org/abs/1009.5055
[48] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8639–8648.

Chao Yang (Student Member, IEEE) received the B.S. degree in electrical engineering from Yanshan University, Qinhuangdao, He'bei, China, in 2015, and the M.S. degree in electrical and computer engineering (ECE) in 2017 from Auburn University, Auburn, AL, USA, where he is currently working toward the Ph.D. degree in ECE with a focus on radio frequency (RF) sensing problems.

His current research interests include health sensing, indoor localization, Internet of Things, and wireless networks.

Mr. Yang was the corecipient of the IEEE Global Communication Conference (GLOBECOM) 2019 Best Paper Award.



Xuyu Wang (Member, IEEE) received the B.S. degree in electronic information engineering and the M.S. degree in signal and information processing, from Xidian University, Xi'an, China, in 2009 and 2012, respectively, and the Ph.D. degree in electrical and computer engineering from Auburn University, Auburn, AL, USA, in 2018.

He is an Assistant Professor with the Department of Computer Science, California State University, Sacramento, CA, USA. His research interests include indoor localization, deep learning, and big data.

Dr. Wand was the corecipient of the Second Prize of Natural Scientific Award of Ministry of Education, China, in 2013, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, Best Paper Award of IEEE Global Communication Conference (GLOBECOM) 2019, Best Journal Paper Award of IEEE Communications Society Multimedia Communications Technical Committee, in 2019, Best Demo Award of IEEE International Conference on Sensing, Communication and Networking (SECON) 2017, and Best Student Paper Award of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2017.

Shiwen Mao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Polytechnic University, Brooklyn, NY, USA (now New York University Tandon School of Engineering), in 2004.

He joined the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA, as an Assistant Professor, in 2006. He held the McWane Endowed Professorship from 2012 to 2015 and the Samuel Ginn Endowed Professorship from 2015 to 2020. Currently, he is a Professor and the Earle C. Williams Eminent Scholar, and the Director of the Wireless Engineering Research and Education Center at Auburn University. His research interests include wireless networks, multimedia communications, and smart grid.

Dr. Mo was the recipient of the IEEE ComSoc Technical Committee on Communications Switching and Routing (TC-CSR) Distinguished Technical Achievement Award, in 2019, the IEEE ComSoc Multimedia Communications Technical Committee (MMTC) Distinguished Service Award, in 2019, Auburn University Creative Research & Scholarship Award, in 2018, the 2017 IEEE ComSoc Internet Technical Committee (ITC) Outstanding Service Award, the 2015 IEEE ComSoc TC-CSR Distinguished Service Award, the 2013 IEEE ComSoc MMTC Outstanding Leadership Award, and NSF CAREER Award, in 2010. He is a corecipient of the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the IEEE ComSoc MMTC 2018 Best Journal Paper Award, the IEEE ComSoc MMTC 2017 Best Conference Paper Award, the Best Demo Award from IEEE International Conference on Sensing, Communication and Networking (SECON) 2017, the Best Paper Awards from IEEE Global Communication Conference (GLOBECOM) 2019, 2016, and 2015, IEEE Wireless Communications and Networking Conference (WCNC) 2015, and IEEE ICC 2013, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is an Associate Editor-in-Chief for IEEE/CIC China Communications, and an Area Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE INTERNET OF THINGS JOURNAL, IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY, and *ACM GetMobile*. He is an Associate Editor for IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE MULTIMEDIA, and IEEE NETWORKING LETTERS. He is a Distinguished Speaker (2018–2021) and was a Distinguished Lecturer (2014–2018) of the IEEE Vehicular Technology Society. He was the TPC Co-Chair of IEEE International Conference on Computer Communications (INFOCOM) 2018 and is the TPC Vice-Chair of IEEE GLOBECOM 2022.