2020 16th International Conference on Mobility, Sensing and Networking (MSN) | 978-1-7281-9916-0/20/$31.00 ©2020 IEEE | DOI: 10.1109/MSN50589.2020.00098

Subject-adaptive Skeleton Tracking with RFID

†Chao Yang, ‡Xuyu Wang, and †Shiwen Mao
†*Dept. of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201*
‡*Dept. of Computer Science, California State University, Sacramento, CA 95819-6021*
Email: czy0017@auburn.edu, xuyu.wang@csus.edu, smao@ieee.org

***Abstract*—With the rapid development of computer vision, human pose tracking has attracted increasing attention in recent years. To address the privacy concerns, it is desirable to develop techniques without using a video camera. To this end, RFID tags can be used as a low-cost wearable sensor to provide an effective solution for 3D human pose tracking. User adaptability is another big challenge in RF based pose tracking, i.e., how to use a well-trained model for untrained subjects. In this paper, we propose Cycle-Pose, a subject-adaptive realtime 3D human pose estimation system, which is based on deep learning and assisted by computer vision for model training. In Cycle-Pose, RFID phase data is calibrated to effectively mitigate the severe phase distortion, and High Accuracy Low-Rank Tensor Completion (HaLRTC) is employed to impute missing RFID data. A cycle kinematic network is proposed to remove the restriction on paired RFID and vision data for model training. The resulting system is subject-adaptive, achieved by learning to transform the RFID data into a human skeleton for different subjects. A prototype system is developed with commodity RFID tags/devices and evaluated with experiments. Compared with a traditional system RFID-Pose, higher pose estimation accuracy and subject adaptability are demonstrated by Cycle-Pose in our experiments using Kinect 2.0 data as ground truth.**

***Index Terms*—Radio-frequency Identification (RFID), Computer Vision (CV), Human pose estimation and tracking, Cycle-consistent adversarial network, Deep Learning.**

1. Introduction

With the rapid development of computer vision, tracking of human poses has become an important problem area in recent years, evolving from 2D [1] to 3D poses [2]. Although camera-based techniques have been shown effective for human pose tracking, such vision-based techniques frequently raise security and privacy concerns. For example, 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. To address this issue, several radio frequency (RF) sensing based schemes have been proposed, such as WiFi [4], [5], Frequency-Modulated Continuous Wave (FMCW) radar [6], and mmWave radar [7].

To this end, radio frequency identification (RFID) provides a promising solution for human pose estimation [8], [9]. Compared with existing contact-free RF sensing systems, RFID tags can be used as wearable sensors because of their small size. Furthermore, the interference caused by the multipath effect is much lower in the RFID system and the cost of RFID systems is lower than the advanced radar based systems. However, because of the low data rate (i.e., the sampling rate) in RFID systems, generating a joint confidence map for all the joints, as in other RF based systems, is highly challenging. Consequently, the existing RFID based pose tracking systems are focused on monitoring the movement of one particular limb using the phase data sampled from multiple tags [10], [11]. When multiple joints are moving simultaneously, the performance could be affected by the disturbance from other RFID tags (e.g., the mutual coupling effect) or inter-tag collisions.

Subject adaptability is another challenge for RF based human pose tracking. Different people have different skeleton forms, but most of the neural networks incorporated in RF based pose tracking systems are trained with a limited number of subjects [5], [9]. The untrained subjects could be considered as a new data domain in machine learning. When testing in the new domain, the performance of the trained model will be degraded. Transfer learning is a possible solution to address the new domain issues [12], [13], but the trained model needs to be updated by a light-weight training for the new domain. The light-weight training requires new vision data of the untrained subject, which, again, leads to privacy concerns. The domain discriminator proposed in recent works [14], [15] could address the domain-adaptive issue, which, however, only works for the classification problem so far. The model structure may not be suitable for data sequence estimation as in human pose tracking.

In this paper, we address the challenges in human pose estimation using RFID with a novel vision-aided, deep learning based solution. We propose the Cycle-Pose system to track the movements of multiple human limbs in real-time. In Cycle-Pose, RFID tags are attached to the target human joints. The movements of the tags are captured by the phase variations when the tags are interrogated by the reader. A vision-aided solution is proposed to help the deep learning model transform tag phase variations to human limb rotations, rather than localizing them with traditional tag localization techniques [16]. Furthermore, we also proposed a

978-1-7281-9916-0/20/$31.00 ©2020 IEEE
DOI 10.1109/MSN50589.2020.00098

novel *cycle consistent adversarial network model* to achieve subject adaptability. The proposed cycle kinematic network model is trained without the restriction of requiring paired RFID and vision data, such that the network can learn to transform RFID data into a human skeleton for *any* subject. Thus, the system achieves higher subject adaptability than traditional schemes when generating human pose for untrained subjects. In Cycle-Pose, the 3D human pose is reconstructed by estimated rotation angles of human limbs from RFID data and any given initial human skeleton in realtime. A specific benefit is that vision data will not be needed anymore in the inference stage, and thus the user's privacy can be well protected. The main contributions of this paper are summarized as follows.

- To the best of our knowledge, this is the first subject-adaptive 3D human pose estimation system using commodity RFID reader and tags, which can effectively track 3D human pose without vision data in the testing stage.
- We propose a cycle kinematic network model and train the network with self-supervision. The proposed model learns the transformation from RFID data to 3D skeleton for different subjects, to effectively achieve subject adaptability.
- We develop a prototype system with commodity RFID tags/devices and Kinect 2.0, to evaluate the system performance and compare it with the traditional technique RFID-Pose [9]. Our experimental study validates that the proposed Cycle-Pose system can effectively track the human pose for different subjects with subject adaptability.

In the following, we present the system overview in Section 2. The challenges and our proposed solutions are discussed in Section 3. Our prototype implementation and performance study are presented in Section 4. Section 5 summaries this paper.

2. System Overview

An overview of the proposed Cycle-Pose system is shown in Fig. 1, which consists of four main modules: (i) Data Collection, (ii) Data Preprocessing, (iii) Cycle Kinematic Network, and (iv) 3D Skeleton Generation.

(i) *Data Collection:* The Cycle-Pose system aims to generate a 3D human skeleton from collected RFID data. Both RFID data and vision data should be sampled for the training process. The RFID data is collected from 12 RFID tags attached to the human joints by three polarized antennas, and is used as input to the proposed cycle kinematic network. The vision data is sampled by Kinect 2.0 for the same subject and action simultaneously. Kinect 2.0 is a depth camera. It captures 3D human movements by both the RGB camera and infrared sensor. 3D movements of each human joint are generated by processing the Kinect data with MATLAB and stored in the form of 3D coordinates for offline training supervision and used as benchmark in the testing phase.

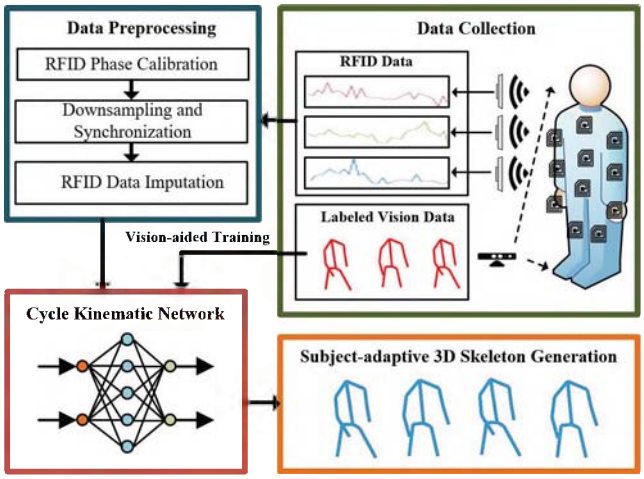

Figure 1. The Cycle-Pose system architecture.

(ii) *Data Preprocessing:* However, the collected raw RFID data cannot be directly used for 3D skeleton tracking. It should be calibrated first before analyzed by the proposed neural network model. The collected RFID phase is severely distorted by channel hopping and phase wrapping of the RFID system. It should be firstly calibrated to mitigate such distortion. After that, since the sampling rates of RFID device and Kinect 2.0 are highly different, the sampled RFID data is downsampled and synchronized with the vision data. Because of the slotted ALOHA-like transmissions in RFID systems, the phase data is not evenly sampled; there is at most one sample in each time slot and most RFID phase samples are missing. Thus, we employ the tensor completion technique, High Accuracy Low-Rank Tensor Completion (HaLRTC), to estimate the missing RFID data.

(iii) *User-adaptive 3D Skeleton Generation with the Cycle Kinematic Network:* We propose a cycle kinematic network to generate 3D pose data from calibrated RFID phases. Unlike monitoring only one particular limb's movements as in traditional RFID based pose tracking systems [10], [11], the proposed system estimates the 3D coordinates of all the joints simultaneously. Moreover, the proposed cycle kinematic network achieves subject adaptability, which is missing in prior systems [5], [9]. This is because the cycle kinematic network is trained with unpaired RFID data and vision data, which is sampled from a different moving subject. Thus, the trained network can achieve better adaptability when transforming RFID data to 3D coordinates for a different, untrained subject.

3. Challenges and Solutions

3.1. RFID Phase Data Calibration

The first challenge in human 3D skeleton generation from RFID data is the poor quality of RFID data. As discussed, the raw RFID data suffers from severe phase distortion and large amount of missing samples. Thus, the

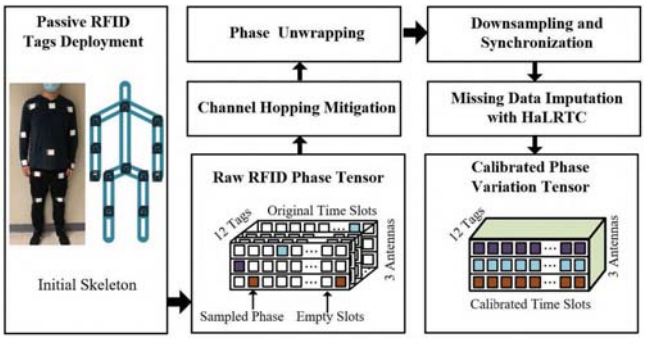

Figure 2. Flowchart of RFID data preprocessing.

RFID data should be well-calibrated before being used to train the neural network. Fig. 2 presents the flowchart of the proposed RFID preprocessing procedure. As shown in the figure, the passive tags are attached to 12 joints of the human body. RFID phase data is collected by the reader using the low-level protocol when interrogating the tags [17], [18].

3.1.1. Phase Distortion Mitigation. The collected phase value indicates the distance between the reader antenna and the tag [17]; so the sampled phases captures the movement of the RFID tags attached to the human body. According to the FCC regulation, the frequency of the channel is not fixed but hops among 50 different channels, which generates a different phase offset for each different channel. Consequently, the phase value is also determined by the current channel used for RFID interrogation. The phase ϕ_s on channel s can be written as [19]:

$$\phi_s = \text{mod}\left(\frac{2\pi 2Lf_s}{c} + \phi_s^0,\ 2\pi\right),\ s = 1, 2, ..., 50, \quad (1)$$

where L is the distance between the tag and antenna, c is the speed of light, and f_s and ϕ_s^0 represent the frequency and initial phase offset of channel s, respectively. According to (1), if we want to track the variation of the tag-to-antenna distance L, the impact of the channel phase offset ϕ_s^0 should be firstly mitigated. Fortunately, the phase offset ϕ_s^0 is a constant for each channel s. If we use the *variation* between two adjacent phase samples on the same channel, the identical channel phase offset in the two samples will be canceled. The phase variation η_s^n for each channel s is:

$$\eta_s^n = \text{mod}\left(\frac{4\pi(L_s^n - L_s^{n-1})f_s}{c},\ 2\pi\right), \quad (2)$$
$$s = 1, 2, ..., 50,\ n = 2, 3, ...,$$

where L_s^n represents the tag-to-antenna distance for the nth sample on channel s. It can be seen that the phase offset ϕ_s^0 is removed in (2), while the movement $L_s^n - L_s^{n-1}$ remains. The phase distortion caused by channel hopping is effectively mitigated this way.

In addition, phase distortion is also caused by the modulo operation in (1) and (2). Since the collected phase data ϕ is rounded to the range $[0, 2\pi]$ rad, sharp phase changes are usually generated by the modulo operation when the phase crosses 0 rad or 2π rad. Thus, the calculated phase variation should be unwrapped to remove such distortion. Given the 110Hz sampling rate used in the RFID system, we assume that the phase variation between two adjacent samples, η, should be no larger than π or smaller than $-\pi$. We use the following scheme to unwrap the sampled phase variation data η.

$$\eta' = \begin{cases} \eta - 2\pi\frac{\eta}{|\eta|}, & \text{if } |\eta| > \pi \\ \eta, & \text{otherwise.} \end{cases} \quad (3)$$

which automatically determines whether the value should be unwrapped by adding or subtracting 2π. After the unwrapping process, all sharp phase changes will be smoothed out, and the calibrated phase variation data can effectively represent the movements of the RFID tags now.

3.1.2. Data Imputation. In addition to distortion, missing phase samples is another challenge caused by the Slotted ALOHA-like transmission used in RFID communications. In each time slot, only *one* tag can send its EPC and low-level data to the reader. In the Cycle-Pose system, although we attach 12 tags to the human body, only one tag can be sampled at a time. In the input data tensor illustrated in Fig. 2, there is only one sample in each slice (12 tags $\times$ 3 antennas). Thus, the sparsity of the phase data tensor is $35/36$, which is way too high for pose estimation. In order to learn the relationship between RFID data and 3D skeleton data obtained from Kinect, we need to (i) deal with the high sparsity issue in the tensor data and (ii) synchronize the phase data (i.e., the input to the deep learning model) and the vision data (i.e., the labels). To solve these problems, we first downsample the RFID data from 110Hz to 30Hz to match the 30fps sampling rate of Kinect. Then, we synchronize the RFID and vision data based on the timestamps when the RFID and vision data samples are simultaneously collected for the same subject. Note that we cannot synchronize the data from different subjects because the RFID data and vision data are not sampled simultaneously in this case.

After synchronizing the two types of data, the input tensor to the deep learning model can be expressed as:

$$\boldsymbol{H}(:,:,t) = \begin{bmatrix} \eta_{t,1}^1 & \eta_{t,2}^1 & \cdots & \eta_{t,n_G}^1 \\ \eta_{t,1}^2 & \eta_{t,2}^2 & \cdots & \eta_{t,n_G}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \eta_{t,1}^{n_A} & \eta_{t,2}^{n_A} & \cdots & \eta_{t,n_G}^{n_A} \end{bmatrix},\ t = 1, 2, ..., N_t, \quad (4)$$

where t means the tth time slot, n_A and n_G are the numbers of antennas and tags, respectively, and $\eta_{t,n_G}^{n_A}$ represents the calibrated phase variation from tag n_G sampled by antenna n_A in time slot t. To address the high sparsity of the tensor, we leverage tensor completion to estimate the missing samples in $\boldsymbol{H}$. The algorithm used in the system is HaLRTC [19], which can achieve high accuracy in data imputation at a relatively high speed.

3.2. Skeleton Generation from RFID Data

Three-dimension human skeleton generation with RFID data is highly challenging also because of the extremely low sampling rate restriction in RFID systems. Most of the existing human pose tracking system is based on the confidence map generated from collected signals, such as camera [1], WiFi [4], and FMCW radar [6]. The human features is first captured and shown on the confidence map, so the human skeleton can be further constructed based on the map. However, this technique is not suitable for RFID based systems, because the sampling rate of the RFID system is much lower than that of video camera, WiFi, and FMCW radar. This is because the RFID system is designed to interrogate RFID tags one at a time, which means no matter how many tags are used in the system, the maximum data rate is fixed by one sample per time slot. If we want to generate a confidence map video with 10fps from RFID data with 110Hz sampling rate, only the 11 phase samples (i.e., sampled at the same time as one video frame or 11 pixels) can be used for map generation. Even if we reduce the map resolution to 100×100, transforming the 11 samples to 10,000 pixels is a severely *ill-posed problem*, which is challenging for training the deep learning model.

In this paper, we employ the forward Kinematic technique to tackle the ill-posed problem, which is widely used in robotics and 3D animation [20]. With a given initial skeleton (i.e., the original locations of all joints and the lengths of all limbs), forward kinematic can estimate the joint position based on the relative space rotation and its parent joint position. For example, when the right elbow position is given, the right-hand position of the subject can be calculated by the length of the front arm and the relative rotation between the hand and elbow. In the proposed cycle Kinematic network, a 3D rotation is represented in a *unit quaternion* format based on Ruler's rotation theorem, expressed as:

$$r + x\vec{\alpha} + y\vec{\beta} + z\vec{\gamma}, \tag{5}$$

where r, x, y, and z are real numbers, and $\vec{\alpha}$, $\vec{\beta}$, and $\vec{\gamma}$ are the quaternion units related to the three coordinates, respectively. Given the 3D position of a joint, represented as $a\vec{\alpha} + b\vec{\beta} + c\vec{\gamma}$, and a 3D rotation with unit quaternion $r + x\vec{\alpha} + y\vec{\beta} + z\vec{\gamma}$. The rotation matrix $\mathbf{\Omega}$ can be derived as:

$$\mathbf{\Omega} = \begin{bmatrix} 1-2(y^2+z^2) & 2(xy+zr) & 2(xz-yr) \\ 2(xy-zr) & 1-2(x^2+z^2) & 2(yz+xr) \\ 2(xz+yr) & 2(yz-xr) & 1-2(x^2+y^2) \end{bmatrix}. \tag{6}$$

The updated position of the joint, $a'\vec{\alpha} + b'\vec{\beta} + c'\vec{\gamma}$, will be calculated as:

$$\begin{bmatrix} a' \\ b' \\ c' \end{bmatrix} = \mathbf{\Omega} \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \tag{7}$$

With the forward kinematic technique, the current human pose is determined by the previous human pose and the 3D rotation for each body joint. To estimate the positions of the 12 human joints, only 48 parameters are required. Compared with the traditional approach of generating a 10,000-pixel map, the proposed technique effectively reduces the problem complexity and can improve the accuracy as well.

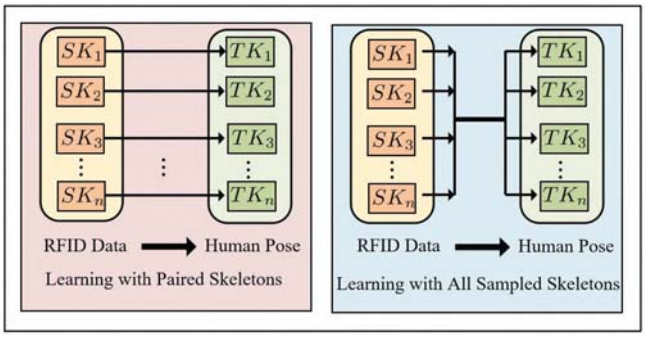

Figure 3. Different structures for pose generation training.

3.3. Dealing with Subject Adaptability

The forward Kinematic technique can effectively address the ill-posed problem. However, the initial skeleton of the subject is still needed, which limits the adaptability of the trained model to different untrained subjects. People have different skeleton forms. To make sure that the deep learning model can successfully generate 3D skeleton for different subjects, the training dataset should include all kinds of human skeletons, which leads to a significantly high cost on collection of labeled training data. If the network is trained with a limited amount of skeletons, the performance of the network could be poor when testing for a subject with a new skeleton not included in the training dataset [5]. This is because the traditional training process is performed with the same source initial skeleton and target initial skeleton, which is illustrated in the left-hand-side graph in Fig. 3. In the figure, SK_n represents the source initial skeleton for subject n in the RFID data, while TK_n represents the target initial skeleton in the vision data with $TK_n = SK_n$. The traditional training structure is focused on learning the relationship between 3D skeleton coordinates and the RFID data for the same skeleton. Thus, the training results is suitable for these n specific skeletons included in the training data, but the well-trained model may not perform well when it is used to test a new skeleton.

3.3.1. Cross-skeleton Learning Structure. To improve the subject adaptability of the learning model, the network should be designed to learn the relationship between different source and target skeletons, so that the system could effectively transform RFID data to 3D skeleton no matter the given subject skeleton is included in the training dataset or not. Thus, we propose a new network structure as illustrated in the right-hand-side graph of Fig. 3. As shown in the figure, the training is not only focused on learning with paired skeletons, but also leaning with different source

skeletons and target skeletons. For example, for a specific movement type such as kicking, all RFID data and vision data are utilized in training, no matter the movement data is sampled from the same subject or not. Thus, the network can learn how to transfer RFID data to human 3D pose with different initial skeletons, such as Sk_1, Sk_2, ..., and Sk_n. Since the network is not trained with a specific initial skeleton, the well-trained model can achieve higher subject adaptability compared with the traditional network structure.

Unfortunately, training with different SKs and TKs is challenging because there is a significant variance in training data between two different subjects. Although performing the same movement, different subjects could have very different speeds and scales, as illustrated in Fig. 4. Their limbs have different lengths and it is also hard to guarantee that each RFID tag will be attached at exactly the same location. Fig. 4 shows the skeleton obtained by Kinect when two different subjects perform the same action (i.e., arm waving), sampled at the same frame rate. As shown in the figure, both the hand moving speed and the latitude of the arm are very different between the two data sequences shown in the first row (for Subject 1) and the second row (for Subject 2).

The considerable difference shown in Fig. 4 indicates that the network should not be directly trained with the position loss between estimated pose and vision pose for different subjects. A *self-supervised network* should be designed for cross-skeleton learning with unpaired initial skeletons.

3.3.2. Cycle Kinematic Network. Cycle-Consistent Adversarial Network is an advanced neural network structure, which is proposed to solve the image-to-image translation with unpaired training datasets [21]. The cycle consistent network has also been employed to solve the temporal video alignment problem of two different video streams [22]. The cycle consistent network can generate fake input data from the output data, so the network can be trained with self-supervision between the real input data and the generated fake input data. Thus, the requirement on paired, labeled data is lower than that in traditional neural networks.

Observing the strength of cycle consistent adversarial network, we propose a Cycle Kinematic Network to deal with the challenges in training for cross-skeleton learning, which is presented in Fig. 5. As shown in the figure, the RFID phase variation sequence feature is first extracted by a recurrent encoder termed recurrent encoder-Forward (or, *recurrent encoder-F*). With additional input of the source initial skeleton of the subject, the recurrent decoder-Forward (or, *recurrent decoder-F*) translates the human movement features captured by RFID phase variations to *unit quaternion*, which represents the 3D rotation of the subject's joints. Then, the unit quaternion is employed by the forward Kinematic algorithm to generate 3D human skeleton with a given target initial skeleton. The cycle consistent network is used to recover the RFID data from the estimated quaternion, which is also constructed by the recurrent encoder-Backward and decoder-Backward (or, *recurrent encoder-B* and *recurrent decoder-B*). If the translation from RFID phase variation

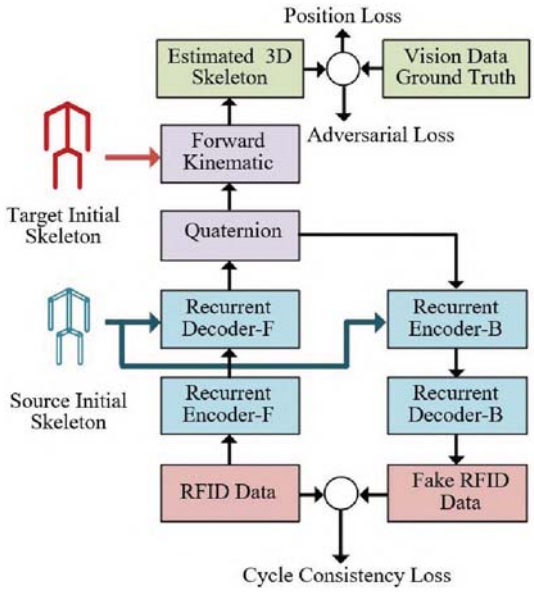

Figure 5. Overview of the proposed cycle kinematic network model.

to 3D limb rotation exists, the inverse transformation could be performed using recurrent autoencoder-B. With the fake RFID data, the network can be trained with self-supervision, and the ground truth of vision data does not need to be strictly paired with the input RFID data.

3.3.3. Loss Function for Training. The loss function used for the training the proposed cycle kinematic network is composed of three parts as illustrated in Fig. 5, including the position loss, the adversarial loss, and the cycle consistency loss. When the training step is set to K, we define the calibrated RFID phase variation sequence as $F_{1:K}$ and the reconstructed fake RFID data as $\hat{F}_{1:K}$. The estimated skeleton by the neural network is represented as $\hat{V}_{1:K}$, and the vision data sequence used for supervision is denoted by $V_{1:K}$. The position loss between the estimated 3D skeleton and the ground truth is calculated with the estimated skeleton and the vision skeleton ground truth as:

$$Loss_p = ||\hat{V}_{1:K} - V_{1:K}||_2^2. \tag{8}$$

However, for the unpaired training data collected from different skeletons, $Loss_p$ also includes the error caused by the asynchronous training dataset. We can calculate the cycle consistency loss as:

$$Loss_c = ||\hat{F}_{1:K} - F_{1:K}||_2^2. \tag{9}$$

Since the cycle consistency loss is calculated with the fake RFID data obtained by the cycle consistent network, the influence of unpaired data can be mitigated by merging the cycle consistency loss and position loss as:

$$Loss_{all} = Q_p \cdot Loss_p + Q_c \cdot Loss_c, \tag{10}$$

with suitable positive coefficients Q_p and Q_c, satisfying $Q_p + Q_c = 1$. With the loss function of the generator

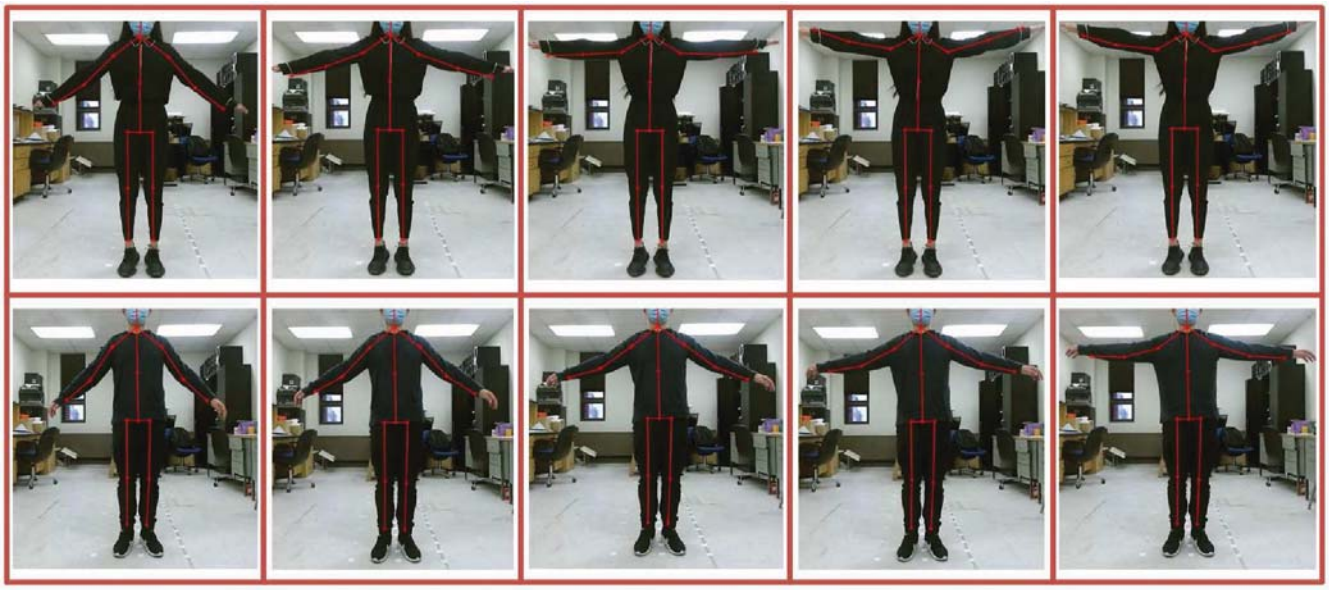

Figure 4. Labeled pose data sampled by Kinect for different subjects. The first row is for Subject 1 and the second row is for Subject 2.

$Loss_{all}$, the network can be effectively trained whether the RFID data and vision data are sampled from the same subject or not. In this paper, we set $Q_p = 0.6$ and $Q_c = 0.4$.

The adversarial loss is used to determine if the network is well trained or not, which is represented as a realism score calculated by a discriminator network D [20], as:

$$S_D = D(\hat{V}_{2:K} - \hat{V}_{1:K-1}, V_{2:K} - V_{1:K-1}). \quad (11)$$

The equation shows that the input of the discriminator is not the position loss but the variation between the previous frame and the current frame in V and $\hat{V}$, respectively. Although V and $\hat{V}$ are unpaired data sequences, the discriminator can determine if the movements performed by the two subjects are the same or not. This is because, for the same movement type, the variations of all the joints between two adjacent data sequences are still similar, no matter the two subject movements are synchronized or not. We set a realism score threshold to balance the discriminator and the generator (i.e., recurrent encoder-F and recurrent decoder-F). When the generator can successfully fool the discriminator, the network will effectively transform RFID data to 3D skeleton data.

4. Implementation and Evaluation

4.1. Prototype System Implementation

To evaluate the performance of Cycle-Pose, we develop a prototype system with an off-the-shelf Impinj R420 reader equipped with three S9028PCR polarized antennas. The RFID tags used in Cycle-Pose are ALN-9634 (HIGG-3). The vision data, used for training supervision as well as the ground truth for test accuracy evaluation, is collected with an Xbox Kinect 2.0 device.

We attach 12 RFID tags to the human body joints as shown in Fig. 6, including the left shoulder, left elbow, left wrist, right shoulder, right elbow, right wrist, neck, pelvis, left hip, left knee, right hip, and right knee. Head and feet are omitted in our prototype system because of the limited scanning range of the RFID antenna used in Cycle-Pose. More antennas can be added to scan the entire body, but the skeleton constructed with the 12 joints is sufficient to monitor human activities in most cases. With the three antennas placed at different altitude positions, every RFID tag can be scanned by at least one of the antennas.

As Fig. 6 shows, RFID phase data is collected when the subject is standing in front of the antennas and performing specific motions repeatedly. Different motions are sampled for training the cycle kinematic network with two different types. The first type includes simple motions, which only involve single-limb movement. The other type includes compound motions, composed of movements of the entire human body, such as boxing, walking, body twisting, and deep squatting.

4.2. Experimental Results and Analysis

To evaluate the performance of Cycle-Pose, we compare it with the traditional neural network model used in RFID-Pose [9], which only trains the network with paired RFID and vision data. The dataset used for training and testing is the same for both models, and the overall accuracy is shown in Fig. 7 and Fig. 10. The overall estimation error $\mathcal{E}_{all}$ used in our evaluation is calculated between the estimated 3D pose data and the ground truth vision data as:

$$\mathcal{E}_{all} = \frac{1}{12}\sum_{n=1}^{12} ||\hat{P}_n - \dot{P}_n||, \quad (12)$$

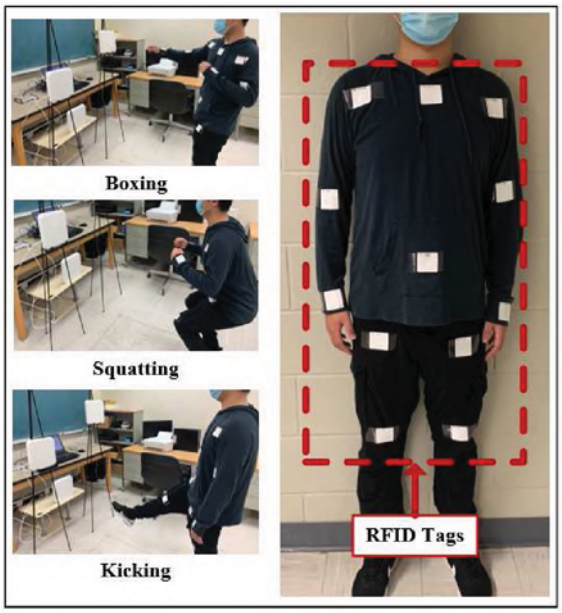

Figure 6. RFID tag deployment and motion sampling.

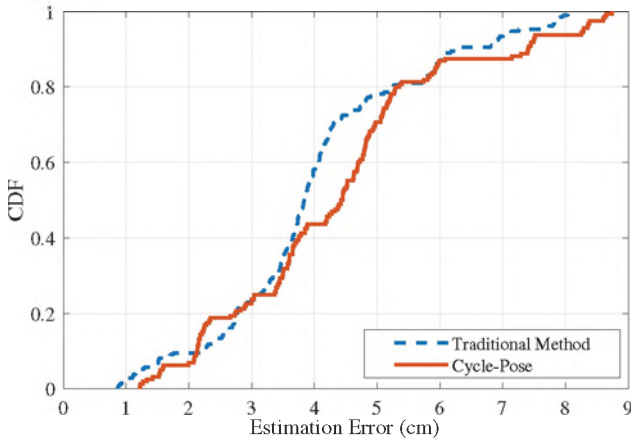

Figure 7. Overall accuracy when testing with trained subjects.

where $\hat{P}_n$ denotes the estimated position of joint n, $\dot{P}_n$ is the ground truth position collected by Kinect for joint n in the 3D space, and $||\hat{P}_n - \dot{P}_n||$ is the Euclidean distance between the two 3D coordinates.

Fig. 7 presents the cumulative distribution functions (CDF) of the estimation errors for both methods when the testing subjects are involved in the neural network training. The CDF curves show that the median estimation error for the traditional network is 3.83cm, while the median error for Cycle-Pose is 4.44cm. The maximum error for Cycle-Pose is 8.64cm, which is slightly higher than that of the traditional method (i.e., 8.09cm). These results show that the accuracy of Cycle-Pose is slightly lower than the traditional method when testing a trained skeleton. This is because the Cycle-Pose system not only learns the translation from RFID data to 3D skeleton, but also learns the transformation

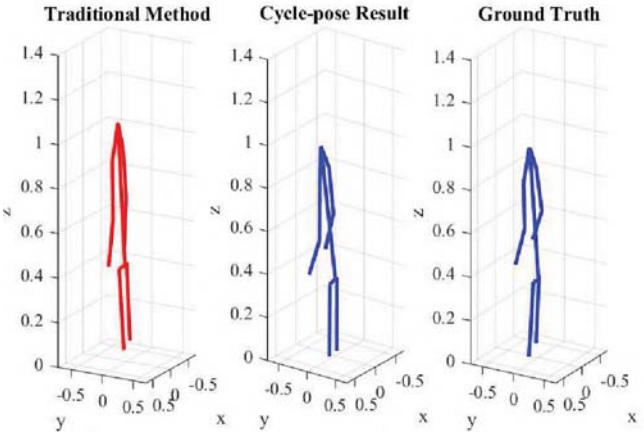

Figure 8. Comparison results when the untrained subject is squatting.

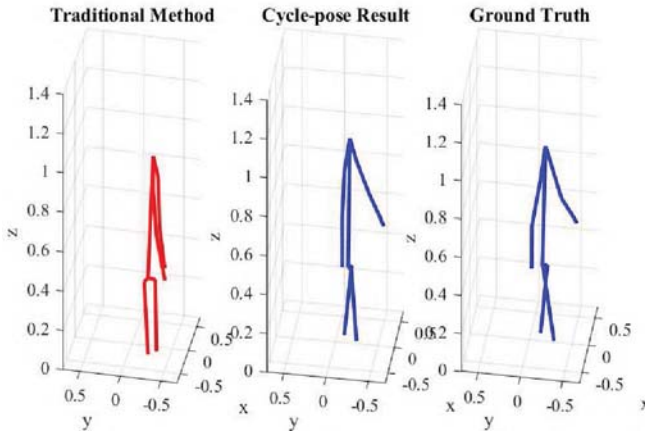

Figure 9. Comparison results when the untrained subject is walking.

from different source skeletons to target skeletons. The additional learning task affects the system performance for specific skeletons. However, the decrease of the accuracy is acceptable in most skeleton tracking applications, such as video gaming and human motion recognition.

The strengths of the Cycle-Pose system become obvious when testing with untrained subjects. Figs. 8 and 9 illustrate the comparison between two networks when an untrained subject is squatting and walking, respectively. From the figures, it can be seen that the human poses reconstructed by the Cycle-Pose system are highly similar to the corresponding ground truth, while the skeletons generated by the traditional method show higher estimation errors.

The accuracy results are presented in Fig. 10. As the CDF results show, the median estimation error of the Cycle-Pose system is 4.88cm, while the median error of the traditional system is 7.66cm, which are both higher than that in Fig. 7. The traditional network is only trained by paired RFID and vision data for the same subject. The training domain is restricted to the specific initial skeletons. When testing with an untrained subject with a different initial skeleton, the traditional network exhibits poorer subject adaptability

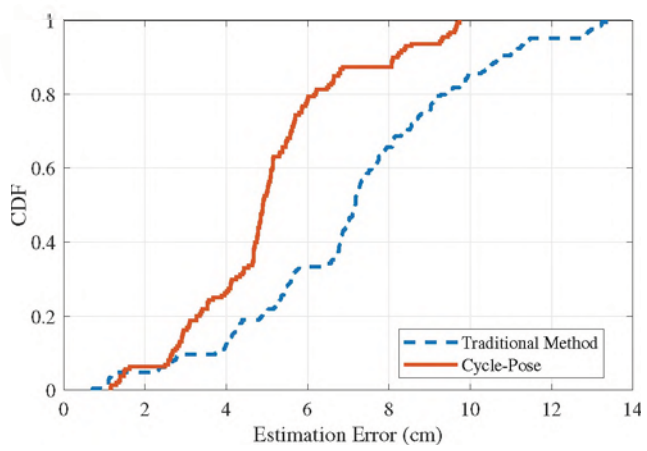

Figure 10. Overall accuracy when testing with untrained subjects.

than the Cycle-Pose system. In summary, although the accuracy of the Cycle-Pose system is slightly lower than that of the traditional RFID pose tracking technique when testing with a known subject, the proposed model achieves high subject adaptability when testing untrained subjects.

5. Conclusions

In this paper, we proposed a subject-adaptive, realtime 3D pose estimation and tracking system named Cycle-Pose. A preprocessing module was proposed to effectively mitigate the influence of phase distortion and missing RFID data samples. The proposed system then leveraged a novel cycle kinematic network to estimate human postures in realtime from RFID phase data, which was trained with unpaired RFID and vision data sampled from different subjects. The Cycle-Pose system was implemented with commodity RFID tags/devices and compared with a traditional RFID based technique RFID-Pose in our experimental study. Its high subject adaptability ability and accuracy were demonstrated in our comparison experimental study using Kinect 2.0 as ground truth.

Acknowledgments

This work is supported in part by the US National Science Foundation (NSF) under Grants ECCS-1923163 and CNS-1822055, and through the Wireless Engineering Research and Education Center at Auburn University.

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 CVPR 2017*, Honolulu, HI, July 2017, pp. 7291–7299.

[2] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Proc. IEEE CVPR 2010*, San Francisco, CA, June 2010, pp. 623–630.

[3] Tom's Guide, "Millions of wireless security cameras are at risk of being hacked: What to do," 2020 (accessed Aug. 28, 2020). [Online]. Available: https://www.tomsguide.com/news/hackable-security-cameras

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

[5] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D human pose construction using WiFi," in *Proc. ACM MobiCom'20*, London, UK, Sept. 2020, pp. 1–14.

[6] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proc. IEEE CVPR 2018*, Salt Lake City, UT, June 2018, pp. 7356–7365.

[7] 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. 17, pp. 10 032–10 044, Sept. 2020.

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

[9] C. Yang, X. Wang, and S. Mao, "RFID-Pose: Vision-aided 3D human pose estimation with RFID tags," *IEEE Transactions on Reliability*, revised.

[10] 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, pp. 1–28, Mar. 2018.

[11] 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, pp. 1–23, Dec. 2018.

[12] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[13] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. ACM ICML 2007*, Corvalis, OR, June 2007, pp. 759–766.

[14] W. Jiang *et al.*, "Towards environment independent device free human activity recognition," in *Proc. ACM MobiCom 2018*, New Delhi, India, Sept. 2018, pp. 289–304.

[15] F. Wang, J. Liu, and W. Gong, "Multi-adversarial in-car activity recognition using RFIDs," *IEEE Trans. Mobile Comput.*, in press.

[16] C. Yang, X. Wang, and S. Mao, "SparseTag: High-precision backscatter indoor localization with sparse RFID tag arrays," in *Proc. IEEE SECON 2019*, Boston, MA, June 2019, pp. 1–9.

[17] M. Lenehan, "Application note - low level user data support," Feb. 2019. [Online]. Available: https://support.impinj.com/hc/en-us/articles/202755318-Application-Note-Low-Level-User-Data-Support

[18] C. Yang, X. Wang, and S. Mao, "Unsupervised drowsy driving detection with RFID," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8151–8163, Aug. 2020.

[19] C. Yang, X. Wang, and S. Mao, "Respiration monitoring with RFID in driving environments," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 2, Feb. 2021, DOI: 10.1109/JSAC.2020.3020606.

[20] R. Villegas, J. Yang, D. Ceylan, and H. Lee, "Neural kinematic networks for unsupervised motion retargetting," in *Proc. IEEE CVPR 2018*, Salt Lake City, UT, June 2018, pp. 8639–8648.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV 2017*, Venice, Italy, Oct. 2017, pp. 2223–2232.

[22] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proc. IEEE CVPR 2019*, Long Beach, CA, June 2019, pp. 1801–1810.