MilliPose: Facilitating Full Body Silhouette Imaging from Millimeter-Wave Device

Aakriti Adhikari
Department of Computer Science and Engineering
University of South Carolina
United States
aakriti@email.sc.edu

Sanjib Sur
Department of Computer Science and Engineering
University of South Carolina
United States
sur@cse.sc.edu

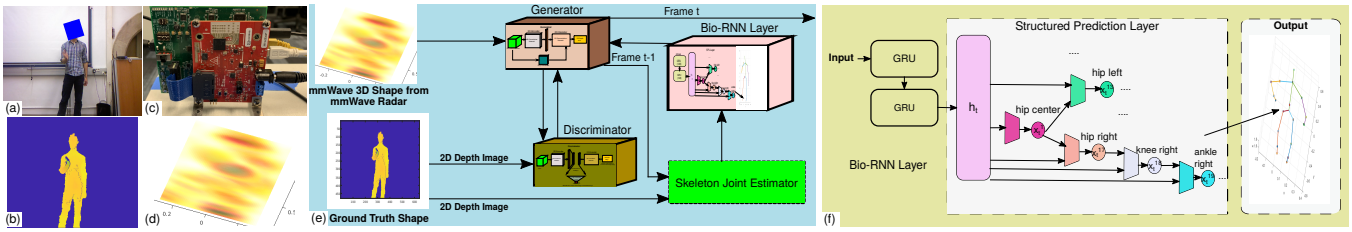

Figure 1: (a) RGB image of a human pose; (b) Its depth image; (c) 77–81 GHz millimeter-wave (mmWave) device; (d) 3D mmWave image of the human pose; (e) *MilliPose* system architecture; (f) Expanded view of the Bio-RNN layer.

ABSTRACT

This paper proposes *MilliPose*, a system that facilitates full human body silhouette imaging and 3D pose estimation from millimeter-wave (mmWave) devices. Unlike existing vision-based motion capture systems, *MilliPose* is not privacy-invasive and is capable of working under obstructions, poor visibility, and low light conditions. *MilliPose* leverages machine-learning models based on conditional Generative Adversarial Networks and Recurrent Neural Network to solve the challenges of poor resolution, specularity, and variable reflectivity with existing mmWave imaging systems. Our preliminary results show the efficacy of *MilliPose* in accurately predicting body joint locations under natural human movement.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Computing methodologies** → **Machine learning approaches**.

KEYWORDS

Millimeter-Wave; Conditional Generative Adversarial Networks; Recurrent Neural Network

ACM Reference Format:
Aakriti Adhikari and Sanjib Sur. 2021. MilliPose: Facilitating Full Body Silhouette Imaging from Millimeter-Wave Device. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp-ISWC '21 Adjunct), September 21–26, 2021, Virtual, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3460418.3479281

UbiComp-ISWC '21 Adjunct, September 21–26, 2021, Virtual, USA

ACM ISBN 978-1-4503-8461-2/21/09.
https://doi.org/10.1145/3460418.3479281

1 INTRODUCTION

Vision-based motion capture systems, such as VICON, Kinect, stereo cameras, *etc.*, can analyze human motions, identify gestures, track limb positions and rotations, and monitor gait, posture, and activities. However, these systems rely on optical cameras, IRs, and LiDARs, which yield poor performance under occlusion, low visibility, and low light conditions [5]. More importantly, they are privacy-invasive and are often undesired by users at their home or office. Millimeter-wave (mmWave) technology in 5G-and-beyond systems enable through-obstruction imaging and are not limited by low light conditions. Besides, the low-resolution, silhouette images generated by mmWave systems comply with the user's privacy while still ensuring accurate motion tracking. Such systems enable many applications, such as assisting at-home fitness training, activity monitoring of patients post-surgery, detecting fall, measuring gait velocity and stride length, tracking the behavior of the elderly to ensure their well-being, and ensuring the safety of construction workers under low visibility conditions (night, rain, or fog).

However, designing such systems capable of silhouette imaging of humans is challenging for two key reasons. (1) *Poor Resolutions*: The mmWave imaging resolution is extremely low compared to the optical cameras. Due to the operating frequency and limited bandwidth, the point spread of mmWave imaging systems is a very wide 2D sinc function along azimuth and elevation. This eliminates the higher frequency components like edges, limbs, and joints of humans, and the resultant image may often look like sparse blobs [4, 6]. (2) *Specularity and Variable Reflectivity*: The mmWave reflections are highly specular, so the signals that only fall normal on the surface are reflected towards the imaging systems [6]. Furthermore, human bodies can absorb most of the mmWave signals, and due to the presence of clothing, different parts of the body would reflect

signals differently. Hence, the output will be a distorted image that is imperceptible with no adequate information about body parts.

We propose *MilliPose*, a machine-learning based system that enables high-quality imaging of humans by overcoming these fundamental challenges. *MilliPose* has two design components: (1) *MilliPose* is inspired by the existing works in enhancing low resolution visual images to high resolution using conditional Generative Adversarial Networks (cGAN) [4, 6] and aims to improve the mmWave image resolution; and (2) Since humans follow well-established rules of biomechanics with limited degrees of freedom of body joints [2], *MilliPose* learns such rules through several visual examples and a Recurrent Neural Network (RNN) and feedbacks the cGAN to accurately recover the missing parts and high spatial frequency information. We have prototyped *MilliPose* on Tensorflow 2.1 by building the RNN module and preliminarily evaluated its efficacy in predicting the joint locations. These predicted joint locations can then be fed back to the silhouette reconstruction module to improve the quality of mmWave shapes of the human body.

2 *MILLIPOSE* DESIGN

Figure 1(e) shows the overall network architecture of *MilliPose*. *First*, *MilliPose* trains a cGAN framework by showing several examples of mmWave images and its corresponding ground-truth depth images from depth sensors; the cGAN module learns the association between the mmWave image to the ground-truth shape. *Next*, *MilliPose* leverages the RNN module to learn the 3D body pose, and from the pose estimated from cGAN and skeleton joint estimator [1] in the previous frame, it predicts the next pose and feedbacks the cGAN module to generate better human shapes; the RNN and cGAN modules are trained simultaneously. *Finally*, during the runtime, when *MilliPose* has been trained appropriately, the model can accurately recover the full silhouette under poor resolution, low reflectivity, and specularity and without the ground-truth shape.

cGAN for Silhouette Generation: Since mmWave images from traditional signal processing suffers from low-resolution and missing high-frequency information, we propose to use a cGAN module, similar to [4, 6]. From hundreds of controlled human motion examples, we can train cGAN by showing millions of 3D images obtained from the mmWave radar, such as [7], and their corresponding ground-truth shapes as a point cloud data (PCD) obtained from depth sensors, such as Kinect. cGAN framework can learn the association between the 3D mmWave images to the 2D ground-truth shapes via a Generator G and a Discriminator D. Besides, the cGAN module takes feedback from the RNN module to correct itself from the predicted pose learned from human biomechanics.

Bio-RNN for Silhouette Quality Improvement: While the cGAN module learns to improve the quality of images per frame, it doesn't consider the inter-frame dependency for human motions. We believe, the quality of the images could be further improved by considering two core intuitions: (1) Human motion happens by the collective working of joints in a hierarchy, and each joint has a limited degrees of freedom and range of motions; and (2) The joints do not change their 3D pose significantly in consecutive frames.

To leverage these intuitions, we propose an auto-regressive model, Bio-RNN, that learns the inter-frame dependency and assists cGAN. Auto-regressive models take in previous pose information at each time step to predict the next pose. The Bio-RNN layer is based on a two-layered Gated Recurrent Unit (GRU) network, and takes as input the previous recurrent state and the features describing the previous pose to predict the next pose. Each GRU layer consists of 128 hidden units, and their states h_t are learned from visual examples. Furthermore, we pass h_t into a Structured Prediction Layer (SPL), similar to [1]. SPL (Figure 1[f]) models the joint hierarchy and predicts each joint individually with separate smaller networks comprised of dense units. Each joint network receives information about its configuration and its immediate parent explicitly and via h_t. SPL provides two benefits: (1) It integrates structural prior in the form of hierarchical architecture, where each joint is modeled by a different network, allowing the network to learn independently; and (2) Each parent propagates its prediction to its child's network, allowing for more precise local predictions. The SPL comprises 20 smaller networks (Figure 1[f]), and each network consists of 64 hidden units followed by LeakyReLU activation and linear projection for absolute joint location predictions.

Loss Function: The Bio-RNN layer's loss function is calculated as the sum of Euclidean distance between the predicted and ground-truth joint locations, and is defined as: $L_{joint} = \sum_{i=1}^{n} \sqrt{\sum |C_p^i - C_t^i|^2}$, where C_p and C_t are the predicted and ground-truth joint locations. We consider 20 key joint locations of the human body [3] for the loss function estimation. To train all the network components simultaneously, we use a combination of the joint loss and the traditional cGAN loss [4]. The combined loss function can then be defined as: $L_{total} = \lambda_1 * L_{cGAN} + \lambda_2 * L_{joint}$, where (λ_1, λ_2) are the hyper-parameters that balances the cGAN loss and body joint loss.

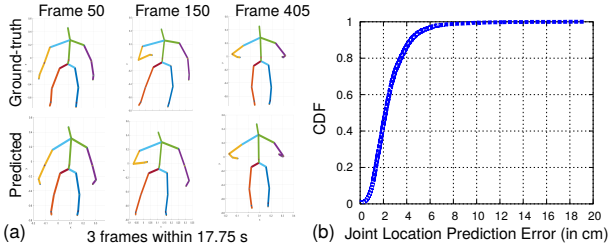

Figure 2: (a) Human pose in 3 example frames; (b) CDF of the error between predicted and ground-truth joint locations.

3 PRELIMINARY RESULTS

We preliminarily evaluate the efficacy of the Bio-RNN model by training it with open-sourced human pose dataset [3]. We select 104,880 frames for training, 26,220 frames for validation, and 500 frames for testing. Each frame in the dataset comprises 20 joints and the corresponding 3D coordinates. We find that Bio-RNN performed much better with Adam optimizer, a learning rate of 10^{-4}, minibatch size of 64, and epochs of 1000. Figure 2(a) shows 3 example frames (spanning less than 20 seconds) of ground truth and predicted joint locations. Furthermore, Figure 2(b) shows the CDF of joint location prediction error; the median error is 2.1 cm only, which shows the effectiveness of *MilliPose* in predicting the 3D poses accurately.

4 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we propose *MilliPose* to enable mmWave imaging for human body by combining machine-learning and human biomechanics knowledge. We design and prototype the Bio-RNN network that considers inter-frame dependency to predict 3D joint location. In the future, we plan to design and prototype cGAN network similar to [6] to generate the human silhouette and jointly train it with the Bio-RNN layer to overcome the poor resolution, specularity, and variable reflectivity problems in existing mmWave imaging systems. We will experiment with multiple volunteers in both home and office settings to evaluate *MilliPose* end-to-end. We believe, *MilliPose* can be a key solution towards enabling a wide range of applications involving human motions, that respects user's privacy and works well under low visibility and through-obstruction.

ACKNOWLEDGMENTS

We sincerely thank the reviewers for their comments. This work is partially supported by the NSF under grant CNS-1910853 and MRI-2018966.

REFERENCES

[1] Aksan, Emre and Kaufmann, Manuel and Hilliges, Otmar. 2019. Structured Prediction Helps 3D Human Motion Modelling. In *IEEE/CVF ICCV*.
[2] Duane Knudson. 2007. *Fundamentals of Biomechanics.* Springer.
[3] Escalera, Sergio, et al. 2013. Multi-Modal Gesture Recognition Challenge 2013: Dataset and Results. In *ACM ICMI*.
[4] Guan, Junfeng and Madani, Sohrab and Jog, Suraj and Gupta, Saurabh and Hassanieh, Haitham. 2020. Through Fog High-Resolution Imaging Using Millimeter Wave Radar. In *IEEE/CVF CVPR*.
[5] K. Bartol and D. Bonjanici and T. Petkovici and N. D'appuzo and T. Pribanici. 2020. A Review of 3D Human Pose Estimation From 2D Images. In *Int. Conf. and Exhibition on 3D Body Scanning and Processing Technologies*.
[6] Regmi, Hem and Saadat, Moh Sabbir and Sur, Sanjib and Nelakuditi, Srihari. 2021. ZigZagCam: Pushing the Limits of Hand-Held Millimeter-Wave Imaging. In *ACM HotMobile*.
[7] TI. 2020. IWR1443 Single-Chip 76-GHz to 81-GHz MmWave Sensor Evaluation Module. https://www.ti.com/tool/IWR1443BOOST