A Survey Of RF-Based Action And Skeleton Recognition

Technical Report

Zhengyuan Jiang*
Department of Information Security
University of Science and Technology of China
Hefei
jzy2018@mail.ustc.edu.cn

October 16, 2021

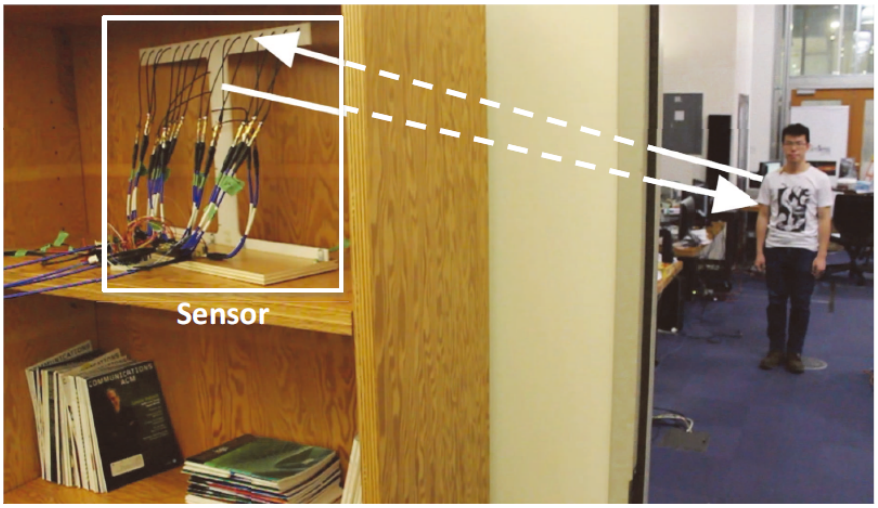
Abstract

Usually, people understand others' actions and interactions depending on vision. It is quite interesting, that wireless methods, can recognize people's movements without actually seeing them. When it is too dark in the room, or when a person is behind the wall, it will be difficult to see clearly in such conditions. Under these circumstances, radio frequency(RF) signals can be useful, because they can travel through the wall and reflect off the human body, and finally be received by antenna array. This passage lists four papers about 3D skeleton or actions recognition based on RF signals.

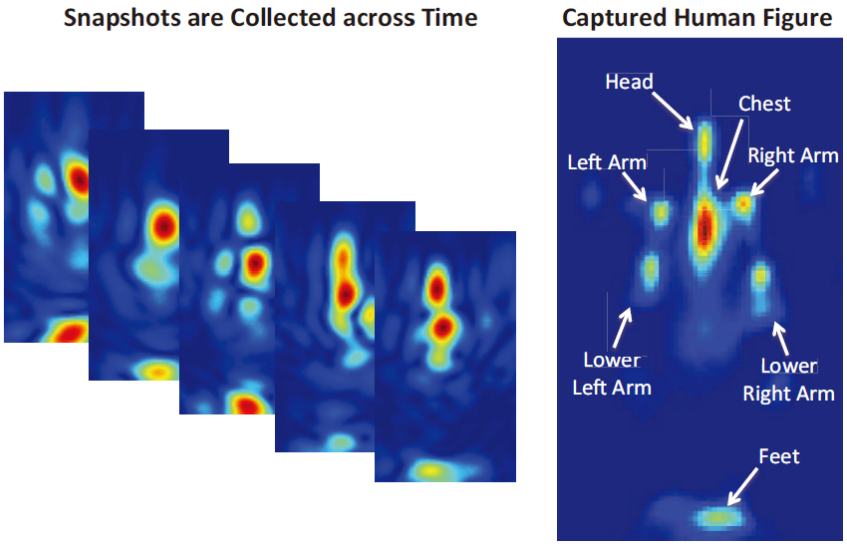
Keywords Wireless sensing · IoT · Posture recognition · Machine Learning

1 Related works

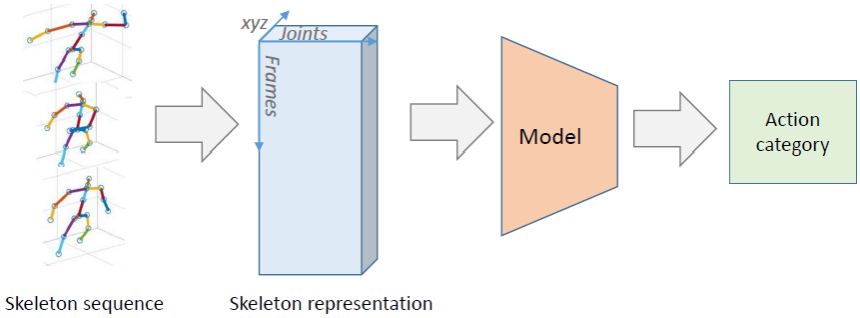
In 2015, a group of MIT[1] present RF-Capture, a system that captures the human figure through a wall. This RF-Capture system tracks the people's limbs and body parts to show a coarse skeleton. The sensor is placed behind a wall. It emits low-power radio signals. The signals traverse the wall and reflect off different objects in the environment, including the human body.



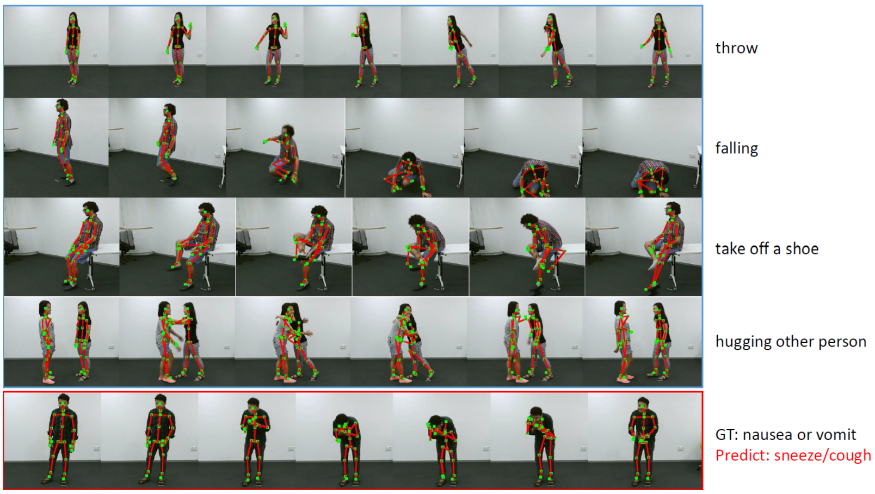
They also use Imaging and Reconstruction Algorithms, which is related to is past work in the Graphics and Vision community on specular object reconstruction. Finally, they successfully capture the human figure by analyzing multiple reflection snapshots across time and combining their information to recover the various limbs of the human body. More importantly, they claim the possibility of capturing human figures by wireless methods and related mathematical background, which I will explain in next part.



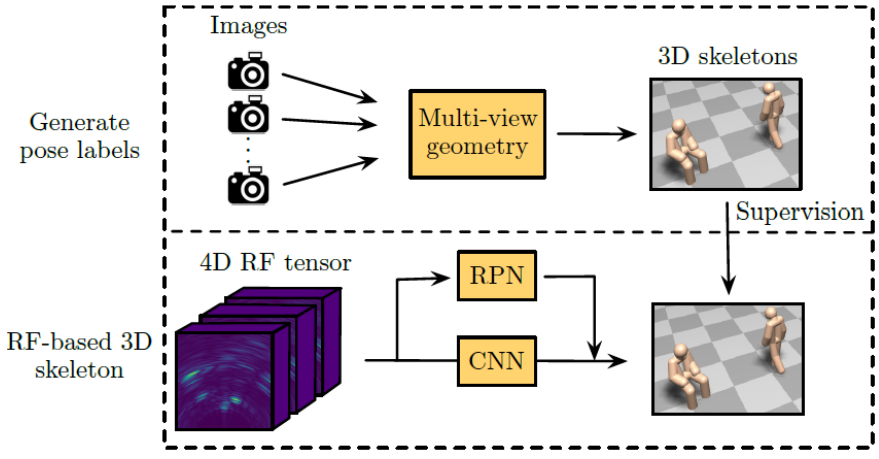
Since than, and due to the availability of large-scale datasets, Skeleton-based human action recognition has recently drawn increasing attentions. In 2018, a group of Hikvision[2] propose a new method to recognize people's action based on skeleton sequence as input. They import LSTM(Long-Short Term Memory) to analyze skeleton sequence. LSTM is an artificial recurrent neural network (RNN) architecture, which are well-suited to classifying, processing and making predictions based on time series data.



Besides, they divide this task into two aspects: the intra-frame representation for joint co-occurrences and the inter-frame representation for skeletons' temporal evolutions. They employ the CNN model for learning global co-occurrences from skeleton data, which is shown superior over local co-occurrences. And they design a end-to-end hierarchical feature learning network, where features are aggregated gradually from point-level features to global co-occurrence features. Finally, their approach consistently outperforms other state-of-the-arts on action recognition and detection benchmarks like NTU RGB+D, SBU Kinect Interaction and PKU-MMD.

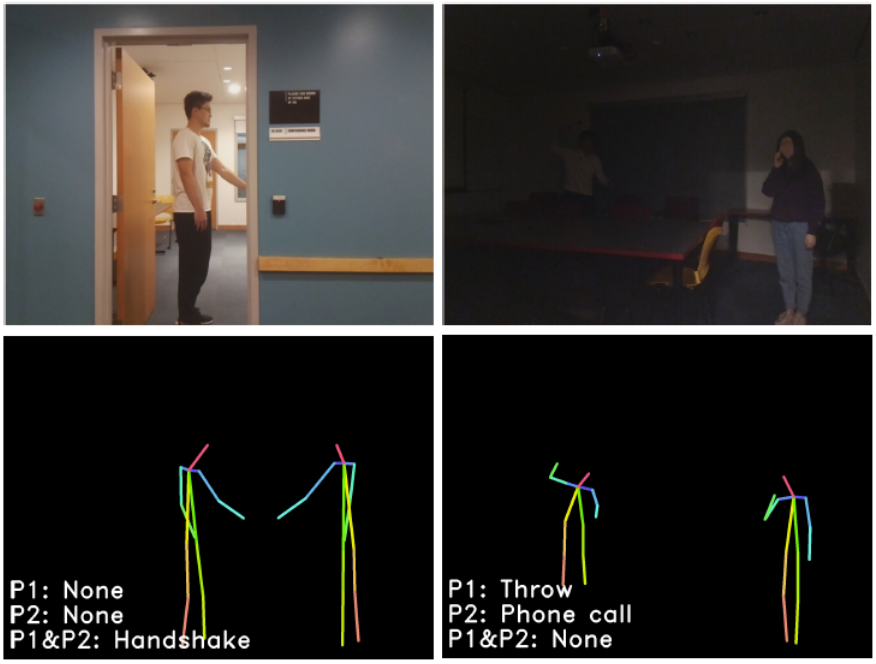


In 2018, a group in MIT[4] introduces RF-Pose3D, the first system that infers 3D human skeletons from RF signals. They divide this task into three parts. First, they receive RF signals with a multi-antenna FMCW radio, and use deep CNNs to analyze the 4-dimensional function of space and time of RF signals. Second, RF signals are converted into abstract domain, where RF information is concentrated, and the signals from different individuals are separated in the abstract domain. This part allows them to distinguish different individuals and deal with multi-person pose estimation. Third, they design a training system, and 12 cameras are used to form this system. They construct a multi-view geometry optimization problem, and restore the 3D skeleton from 12 angles (2D skeleton) to train the RF-Pose3D. After the training, the camera system was no longer needed.



They also use some useful tricks, and get a good results eventually. We can discuss their tricks in the next section.

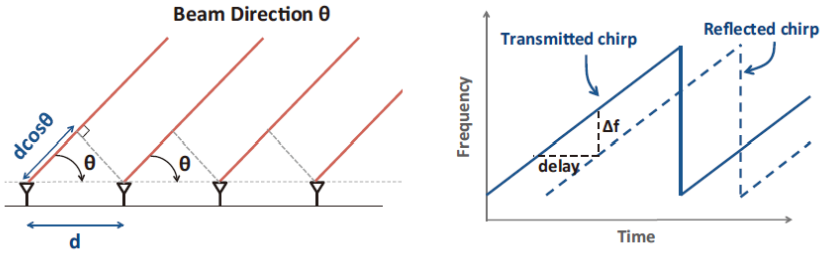
In Sept 2019, this group[3] propose a action recognition system based on their previous work, and it can work through walls and occlusions.



In view of the situation that the traditional image/vision recognition fails due to too dim light/wall obstruction, the 3D human skeleton is generated as an intermediate representation based on CNN, which can learn and train from both visual and RF data sets. The spatiotemporal attention module is introduced to process the skeleton generated by RF signals, and the multi-proposal module is introduced to recognize the actions and interactions of many people. More details will be listed in the next section.

2 Methods

In paper[1], They state that antenna arrays can be used to identify the spatial direction from which the RF signal arrives. This process leverages the knowledge of the phase of the received signals to beamform in post-processing.

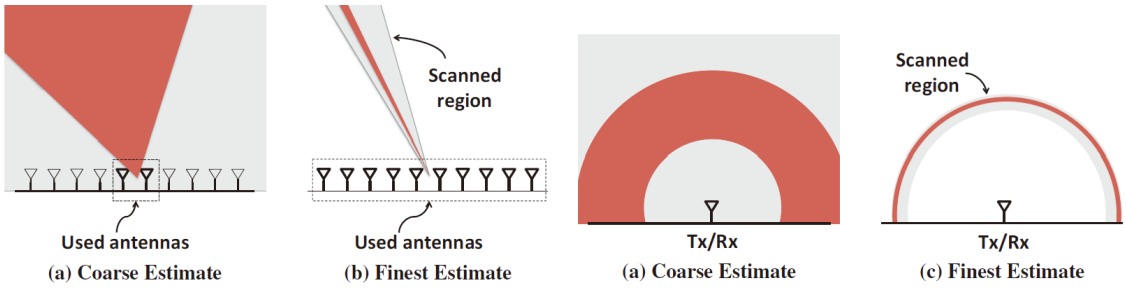


(a) Antenna Array **(b) Frequency Chirp**

And Frequency Modulated Carrier Wave (FMCW) is a technique that allows a radio device to measure the depth of an RF reflector. Combining these two, a RF-Capture system can be designed to capture the human figure. Then two one-dimensional antenna arrays (horizontal and vertical) can focus on the direction (,), and the spatial spherical coordinates (r, ,) can uniquely determine the position of voxels in space. Mathematically, the power arrived from voxels (r, ,) can be calculated as follows:

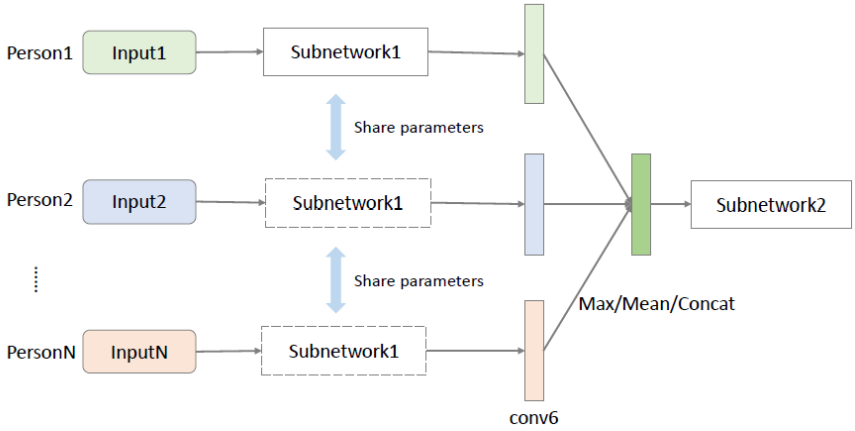
$$P(r,\theta,\phi) = \left| \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{t=1}^{T} s_{n,m,t} e^{j2\pi \frac{kr}{c} t} e^{j\frac{2\pi}{\lambda} \sin\theta (nd\cos\phi + md\sin\phi)} \right|,$$

This paper also propose a idea called Coarse-to-Fine 3D Scan. since much of the 3D space is empty, it would be highly inefficient to scan every point in space. Thus, RF-Capture uses a coarse-to-fine algorithm that first performs a coarse resolution scan to identify 3D regions with large reflection power. It then recursively zooms in on regions with large reflected power to refine its scan. For Angular scan, their RF-Capture starts with a small array of few antennas, and uses more antennas only to refine regions that exhibit high reflection power. For Depth scan, they start by using a small chunk of bandwidth which gives them coarse depth resolution. Then, they refine their estimate by adding more bandwidth to achieve finer resolution, but use that bandwidth only to scan regions of interest.



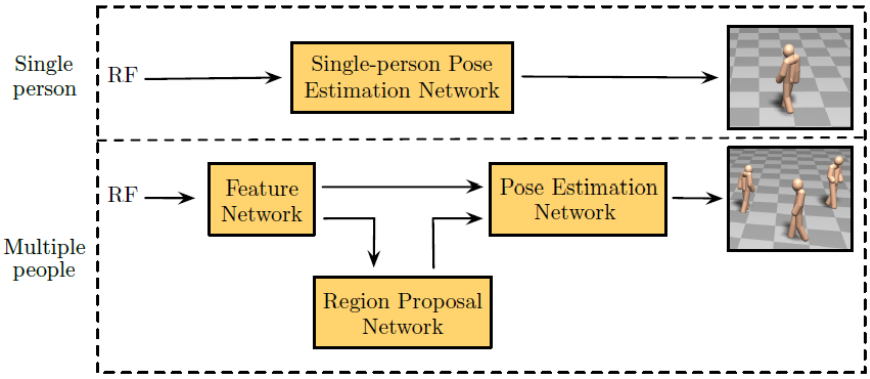
(a) data in/out rate (b) flow situation

In paper[2], it is innovative point for them to use CNN for action recognition. CNN is one of the most powerful and successful neural network models, which has been widely applied in image classification, object detection, video classification. Further, they propose a Hierarchical Co-occurrence Network (HCN) framework, which is designed to learn the joint co-occurrences and the temporal evolutions jointly in an end-to-end manner.

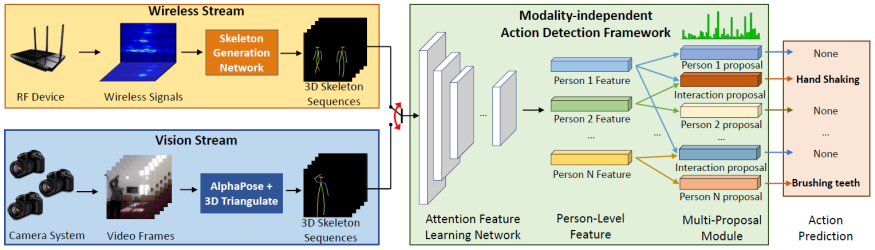


In paper[4], according to the characteristics of RF signals, the group solves the 4D convolution into a combination of 3D convolution performed on two planes and time axis, and used deep convolution neural network for training and learning. This model decomposition that allows us to reduce the complexity from O(XYRT) to O(XRT + YRT). They first prove that 4D RF tensor is planar decomposable. Then they prove that for a layer in a CNN, if its input is planar decomposable, its output is also planar decomposable. Thus, they can stack many convolution layers creating a deep CNN while maintaining decomposability. Finally, they prove that the computation of the loss function and the process of detecting which class has the maximum score are both decomposable when given a decomposable tensor as input. This last step means that they can train the network while operating on its decomposed version –i.e., the two 2D planar tensors and the time axis. This completes the model decomposition.

For multi-person recognition, they follow the divide-and-conquer paradigm by first detecting people regions and then zooming into each region to extract 3D skeleton for each individual. This leads to the design of a new neural network module called region proposal network(RPN), which generates potential people regions. The CNN model is split into feature network (FN) and pose estimation network (PEN). Feature network extracts abstract and high-level feature maps from raw RF signals. Based on these features maps, we first detect potential person regions with RPN. For each region detected by RPN, we zoom into the corresponding region on the feature maps, crop the features and feed them into our pose estimation network.



In paper[3], RF-Action detects human actions from wireless signals. It first extracts 3D skeletons for each person from raw wireless signal inputs (yellow box). It then performs action detection and recognition on the extracted skeleton sequences (green box). The Action Detection Framework can also take 3D skeletons generated from visual data as inputs (blue box), which enables training with both RF-generated skeletons and existing skeleton-based action recognition datasets.



Further, to train the model in an end-to-end manner, they no longer use argmax to extract 3D key point locations, as in past work on RF-based pose estimation[4]. Thus, they use a regressor to perform the function of the argmax to extract the 3D locations of each key point. This makes the model differentiable and therefore the action label can also act as supervision on the skeleton prediction model. Their end-to-end architecture uses 3D skeletons as an intermediate representation to leverage previous skeleton based action recognition datasets, and they combine different modalities to train the model.

3 Discussion

Each of these papers propose a excellent method for RF signal based 3D reconstruction or pose recognition, but they also has some limitations. For instance, in paper[1], they do not import Machine Learning method, thus the recognition performance is not very good and clear. In paper[2], they import RNNs to recognize action from posture sequence. Nevertheless, this will make their method more computational expensive and time consuming. In paper[4], most of experimental data comes from static postures in the office building (for example, walking, sitting and standing), ignoring the dynamic and changeable actions such as running and jumping, and it is not known whether the real-time performance can meet the requirements. Moreover, the distance is required to be within 40 feet. In paper[3], when watching demo, I find that some lines, such as arms, of the 3D skeleton as the intermediate representation sometimes suddenly appeared shift upwards or downwards. Moreover, this was still the case when the characters did not change their movements greatly. Because the recognition model paid more attention to the representative features due to the introduction of the attention module, it could still recognize them, but it might not be very good if people have larger movement range.

It is worth wondering that whether we can simplify the system and use less antennas. For example, if we can achieve almost good results using only one antenna array and several scattered antennas, instead of a 2D antenna array, it may be a useful work.

References

[1] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand. Capturing the human figure through a wall. *ACM Transactions on GraphicsVolume 34Issue 6November 2015 Article No.: 219*, pages 1–13, 2015.

[2] C. Li, Q. Zhong, D. Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv:1804.06055*, 2018.

[3] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi. Making the invisible visible: Action recognition through walls and occlusions. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 872–881, 2019.

[4] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba. Rf-based 3d skeletons. *SIGCOMM '18: Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, page 267–281, August 2018.