Review: the Human Pose Estimation using Radio Frequency

Atul Kumar, Diya Patel, Arundhati Sharma, Baibhav Ranjan, Sandeep Vyas

Abstract— Machine learning and artificial intelligence(AI) has made a lot of advancement in the technologies available. These modifications in the technology have led many researchers of Computer Science and Artificial Intelligence (AI) Laboratory(CSAIL) at Maassachusetts institute of technology to develop an idea which provides a technique that is able to construct a human-like stick figure of a person standing behind a wall. They examined and evaluated the plan of RF-Capture and tools which could analyze data generated through radio signals. Using computer vision as a technique to build a mechanism that is capable of seeing through walls is like achieving a great milestone in this field. The accurate human pose can be estimated even through occlusions and walls using computer vision that was never possible before. Even the results of the experiment were unexpected. It performed beyond what the scientists thought. Moreover, it is not required to provide visual data to model to predict posture, hence there is no need of attaching a device.

Index Terms— Machine learning, Artificial intelligence, Human pose estimation.

1 Introduction

Localizing and tracking the motion of the people in the past years has been at boom due to security and medical reasons using wireless signals. This is now possible using the newly developed neural network model RF-Pose. RF-Pose uses AI technique and tools to sense people's movements through walls. The project senses change in radio frequencies when a human comes into view, and uses an AI trained with both images of humans in certain poses and the corresponding reflected radio frequencies from their bodies to tell what someone is doing. While normal humans can't spot through walls, the AI system was designed using images and RF changes with no visible barrier but was then able to pick out people's poses when a wall was placed between them and the system [1].

The image seen in figure 1 illustrates this ability quite well, where stick-legs are generated on a human whose torso is visible through a window.

To capture a human figure whether occluded or from behind a wall RF-pose first emits wireless signals that can penetrate wall but not a human body. These signals that bounce back from the human's body are analyzed by the neural network and are used to reconstruct the stick figure. Moreover, it does not require the person to attach any sensor to the body or to wear any device.

The applications are as follows [2]:

- It can detect a person through occlusions and walls.
- It can determine a person's movement from behind a wall.
- It can even trace handwriting of a person in air through walls.

The neural network incorporated is used to guide wireless devices to observe posture and action of people even from the other side of the wall. The power of the transmitted signal is 10000 times weaker as compared to a standard cell-phone. These are the radio signals that reflect back after striking human's bodies. Hence these signals are used to create a dynamic stick figure and heat map as shown in figure 2, which moves its body exactly how a person moves [3].

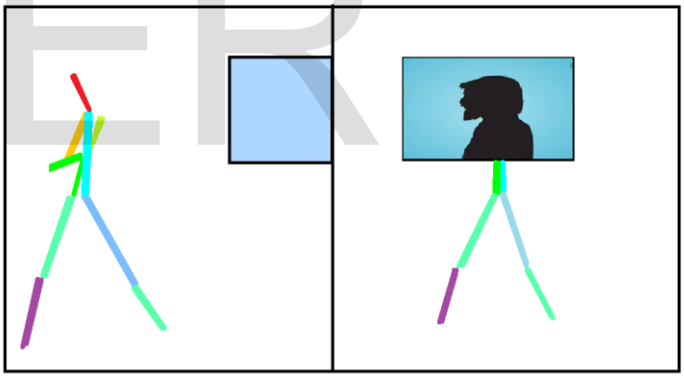

Fig. 1 Stick figure obtained from the neural model.

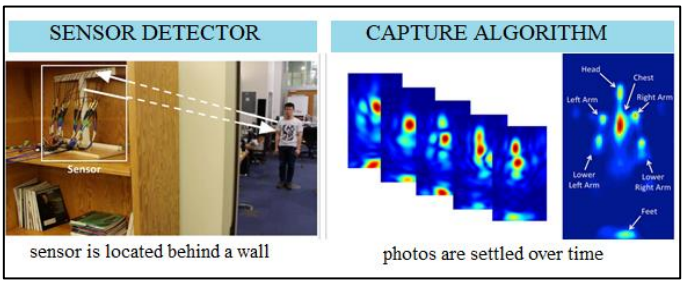

Fig. 2 Device Setup and Heat map (Reproduced from [2]).

2 Working

2.1 Dataset

The data was collected in the public environment and was recorded and used with the consent of the person. It was col-

- *Atul Kumar and Baibhav Ranjan is currently pursuing bachelor degree program in Electronics and Communication Engineering at JECRC, Jaipur, India, E-mail: atulsrivastwa11@gmail.com*
- *Diya Patel and Arundhati Sharma is currently pursuing bachelor degree in Computer Science engineering at JECRC, Jaipur, India.*
- *Dr. Sandeep Vyas is currently working as Associate Professor in Electronics and Communication Engineering Department at JECRC Jaipur, India.*

lected when hundreds of people were doing daily indoor activities.

1. Synchronized wireless and vision data.
2. Web camera + RF sensor.
3. Synchronized (images + RF data) with an average error of 7ms.

To generalize the model it was tested and trained in different places and on different people. It was tested when a hundred new people were doing their work [3].

3 METHOD

3.1 Computer Vision

Computer vision techniques are used to acquire, interpret and discern images and videos. Recent advancement in digital imaging technology and the ubiquity of digital cameras has made Computer Vision solutions cost-effective and practical [4].

Pose estimation using computer vision-

A. Top-down method: - detect bodies in the pictures then apply pose estimator to collect the key points [5].

B. Bottom-up method: - detect key-points in the pictures and then apply post-processing [6-7].

The bottom-up approach was used in the RF-pose. Specifically, the model follows the cross model student-teacher network.

Cross-model-teacher-student-network: It transfers richer information on dense key points of confidence maps. In particular, it transfers knowledge or information learned in a data modality i.e., the teacher network into the other i.e., the student network [4].

3.2 Cross Model Supervision

The design and training the network was a challenge because training a model requires labeled data but it is infeasible for a human being to annotate the radio signals with key-points.

The wireless device uses the lower frequencies to track people through walls. RF signal, 1000 times lower than Wi-Fi range can traverse occlusion such that it could trace or track the human body which provides us the key to make stick figure.

Here the cross model-student-teacher-network-model was very useful [4]. Teacher network provides cross model supervision. Student network provides RF-based pose estimator. It first modeled the information of the human pose and then transferred this knowledge using RF signals and synchronized images or pictures as a bridge. The key point confidence maps are predicted by the teacher network which takes the image data as input. These predicted maps generated as output are then used by the student network as input. These maps act as cross-model supervisor for the student network. RF signals are then used by student network in addition to those maps to learn and ultimately predict key point confidence maps from. The 2Dpose estimation network is used as the teaching network. While the student network gets trained for predicting fourteen key point confidence maps which correspond to the below mentioned anatomical parts of a person's body [8]:

- Head
- Elbows
- Wrists
- Neck
- Hips
- Shoulders
- Knees
- Ankles

Radio signals played a major role in teaching the model, the wireless model was fed with the information of these radio signals and the images of a human performing free climbing, walking, sitting, running, etc. During the training, a web camera was attached to a wireless sensor and visual string were synchronized. The model then extracted information of the pose from the visual string and used it as a supervisory signal for the wireless stream [9].

Fig. 3. Stick figure extracted as output.

This model is trained with 70% data and tested with 30% of the data. Training and testing data were from different environments.8-camera the system was used to provide ground-truth, which further helps in building 3-D human poses [3].

Fig. 4. Different environments in the data (Reproduced from [3]).

After the training is completed there is no use of attaching a web camera or any other device to capture images. It rather uses only the radio signals that bounce off a human's body. Hence it performs pose estimation even if a person is in a different room or fully occluded. This was a huge success because an artificial neural network model has never been observed to perform like this. The performance was recorded as shown in Table 1.

TABLE 1
ACCURACY OBTAINED IN BOTH CASES [3]

	vision system	RF pose
Visible scenes	68.80%	62.40%
Through walls	0%	58.10%

To decrease the error we use sequences of frames to make the network look at, in place of taking only a single frame as the input. To make the standard invariant to translations in both space and time, we use 10 layers of spatiotemporal convolutions. Pytorch is used for the implementation.

4 Applications

It works well for the multi-person pose estimation as well.

- Medical scope: RF-Pose estimation technique could be put to monitor diseases or illness such as multiple sclerosis, Parkinson's, and muscular dystrophy, which can help doctors to adjust medications accordingly.
- Security: It can help in security issues, as it can detect the person behind the wall causing any bayonet activity.
- Gaming: New versions of video games could also be created by using this technique.
- Military Uses: It can help soldiers in their mission and even in search and rescue operation of the victims during natural calamities.
- A key advantage of this technique is that sufferers or patients neither have to wear sensors nor have to charge their devices.

5 Limitations

A. Inter-person occlusion.
B. The performing distance of radio is fully dependent upon its transmission power.
C. This identification process considers only one activity that is, walking. The other activities are left for future research.

6 Conclusion

Earlier X-ray vision had a limited scope up to some fields but now since the discovery of the seeing through walls technique, it has opened up new research opportunities and also provides a new sensing modality that is different from visible light, that can model people from behind the walls and prepare a 2D projection in the form of heat maps. This technique can boost vision systems with powerful capabilities that will be very helpful in future related research work.

References

[1] J. S. Cook," RF-Pose Sees Through Walls with AI", https://blog.hackster.io/rf-pose-sees-through-walls-with-ai-259b6af36af6, 2018.
[2] H. Mao, F. Durand, F. Adib, D. Katabi, and C.Y. Hsu, "Capturing the Human Figure Through a Wall," ACM Transactions on Graphics (TOG), Volume 34, Issue 6, 2015.
[3] Y. Tian, M. A. Alsheikh H. Zhao, A. Torralba, D.Katabi,T. Li, and M .Zhao, "Through-Wall Human Pose Estimation Using Radio Signals," IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, pp. 7356-7365, 2018.
[4] A. Torralba ,Y. Aytar, and C. Vondrick, "Soundnet: Learning sound representations from unlabeled video", 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 2016.
[5] C. L. Rmpe , H. Fang, and S. Xie, "Regional multi-personpose estimation", IEEE International Conference on Computer Vision (ICCV), 2017.
[6] Y. Sheikh ,Z. Cao, S.E. Wei, and T. Simon, "Realtime multi-person 2D pose estimation using part affinity fields", Computer Vision and Pattern Recognition, pp. 7291-7299, 2017.
[7] B. Schiele,L. Pishchulin, M. Andriluka, B. Andres, P. V. Gehler, S. Tang, and E. Insafutdinov, " Deepcut: Joint sub-set partition and labeling for multi person pose estimation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016.
[8] C. L. Zitnick, T.-Y. Lin , P. Perona. S. Belongie, D. Ramanan, J. Hays, P. Dollar, and M. Maire, "Microsoft COCO: Common objects in context", European Conference on Computer Vision, pp. 740-755, 2014.
[9] H. Zheng , Y. Zhu, and, B. Y. Zhao, "Reusing 60 ghz radios for mobile radar imaging", Proceedings of 21st Annual International Conference on Mobile Computing and Networking (MobiCom 2015), pp. 103-116, 2015.