Making the Invisible Visible: Action Recognition Through Walls and Occlusions

Tianhong Li* Lijie Fan* Mingmin Zhao Yingcheng Liu Dina Katabi
MIT CSAIL

Abstract

Understanding people’s actions and interactions typically depends on seeing them. Automating the process of action recognition from visual data has been the topic of much research in the computer vision community. But what if it is too dark, or if the person is occluded or behind a wall? In this paper, we introduce a neural network model that can detect human actions through walls and occlusions, and in poor lighting conditions. Our model takes radio frequency (RF) signals as input, generates 3D human skeletons as an intermediate representation, and recognizes actions and interactions of multiple people over time. By translating the input to an intermediate skeleton-based representation, our model can learn from both vision-based and RF-based datasets, and allow the two tasks to help each other. We show that our model achieves comparable accuracy to vision-based action recognition systems in visible scenarios, yet continues to work accurately when people are not visible, hence addressing scenarios that are beyond the limit of today’s vision-based action recognition.

1. Introduction

Human action recognition is a core task in computer vision. It has broad applications in video games, surveillance, gesture recognition, behavior analysis, etc. Action recognition is defined as detecting and classifying human actions from a time series (video frames, human skeleton sequences, etc). Over the past few years, progress in deep learning has fueled advances in action recognition at an amazing speed [30, 40, 31, 18, 8, 11, 23, 17, 20]. Nonetheless, camera-based approaches are intrinsically limited by occlusions – i.e., the subjects have to be visible to recognize their actions. Previous works mitigate this problem by changing camera viewpoint or interpolating frames over time. Such approaches, however, fail when the camera is fixed or the person is fully occluded for a relatively long period, e.g., the person walks into another room.

Intrinsically, cameras suffer from: our eyes sense only visible light and hence cannot see through walls and occlusions. Yet visible light is just one end of the frequency spectrum. Radio signals in the WiFi frequencies can traverse walls and occlusions. Further, they reflect off the human body. If one can interpret such radio reflections, one can perform action recognition through walls and occlusions. Indeed, some research on wireless systems has attempted to leverage this property for action recognition [33, 39, 19, 37]. However, existing radio-based action recognition systems lag significantly behind vision-based systems. They are limited to a few actions (2 to 10), poorly generalize to new environments or people unseen during training, and cannot deal with multi-person actions (see section 2 for details).

In this paper, we aim to bridge the two worlds. We introduce, RF-Action, an end-to-end deep neural network that recognizes human actions from wireless signals. It achieves performance comparable to vision-based systems, but can work through walls and occlusions and is insensitive to lighting conditions. Figure 1 shows RF-Action’s performance in two scenarios. On the left, two people are shaking hands, yet one of them is occluded. Vision-based humans, suffer from: our eyes sense only visible light and hence cannot see through walls and occlusions. Yet visible light is just one end of the frequency spectrum. Radio signals in the WiFi frequencies can traverse walls and occlusions. Further, they reflect off the human body. If one can interpret such radio reflections, one can perform action recognition through walls and occlusions. Indeed, some research on wireless systems has attempted to leverage this property for action recognition [33, 39, 19, 37]. However, existing radio-based action recognition systems lag significantly behind vision-based systems. They are limited to a few actions (2 to 10), poorly generalize to new environments or people unseen during training, and cannot deal with multi-person actions (see section 2 for details).

In this paper, we aim to bridge the two worlds. We introduce, RF-Action, an end-to-end deep neural network that recognizes human actions from wireless signals. It achieves performance comparable to vision-based systems, but can work through walls and occlusions and is insensitive to lighting conditions. Figure 1 shows RF-Action’s performance in two scenarios. On the left, two people are shaking hands, yet one of them is occluded. Vision-based

Figure 1: The figure shows two test cases of our system. On the left, two people are shaking hands, while one of them is behind the wall. On the right, a person is hiding the dark and throwing an object at another person who is making a phone call. The bottom row shows both the skeletal representation generated by our model and the action prediction.
systems would fail in recognizing the action, whereas RF-Action easily classifies it as handshaking. On the right, one person is making a phone call while another person is about to throw an object at her. Due to poor lighting, this latter person is almost invisible to a vision-based system. In contrast, RF-Action recognizes both actions correctly.

RF-Action is based on a multimodal design that allows it to work with both wireless signals and vision-based datasets. We leverage recent work that showed the feasibility of inferring a human skeleton (i.e., pose) from wireless signals [43, 45], and adopt the skeleton as an intermediate representation suitable for both RF and vision-based systems. Using skeletons as an intermediate representation is advantageous because: (1) it enables the model to train with both RF and vision data, and leverage existing vision-based 3D skeleton datasets such as PKU-MMD and NTU-RGB+D [26, 31]; (2) it allows additional supervision on the intermediate skeletons that helps guide the learning process beyond the mere action labels used in past RF-based action recognition systems; and (3) it improves the model’s ability to generalize to new environments and people because the skeleton representation is minimally impacted by the environment or the subjects’ identities.

We further augment our model with two innovations that improve its performance: First, skeleton, particularly those generated from RF signals, can have errors and mispredictions. To deal with this problem, our intermediate representation includes in addition to the skeleton a time-varying confidence score on each joint. We use self-attention to allow the model to attend to different joints over time differently, depending on their confidence scores.

Second, past models for action recognition generate a single action at any time. However, different people in the scene may be engaged in different actions, as in the scenario on the right in Figure 1 where one person is talking on the phone while the other is throwing an object. Our model can tackle such scenarios using a multi-proposal module specifically designed to address this issue.

To evaluate RF-Action, we collect an action detection dataset from different environments with a wireless device and a multi-camera system. The dataset spans 25 hours and contains 30 individuals performing various single-person and multi-person actions. Our experiments show that RF-Action achieves performance comparable to vision-based systems in visible scenarios, and continues to perform well in the presence of full occlusions. Specifically, RF-Action achieves 87.8 mean average precision (mAP) with no occlusions, and an mAP of 83.0 in through-wall scenarios. Our results also show that multimodal training improves action detection for both the visual and wireless modalities. Training our model with both our RF dataset and the PKU-MMD dataset, we observe a performance increase in the mAP of the test set from 83.3 to 87.8 for the RF dataset (no occlusion), and from 92.9 to 93.3 for the PKU-MMD dataset (cross subjects), which shows the value of using the skeleton as an intermediate common representation.

Contributions: The paper has the following contributions:

- It presents the first model for skeleton-based action recognition using radio signals; It further demonstrates that such model can accurately recognize actions and interactions through walls and in extremely poor lighting conditions using solely RF signals (as shown in Figure 1).
- The paper proposes “skeletons” as an intermediate representation for transferring knowledge related to action recognition across modalities, and empirically demonstrate that such knowledge transfer improves performance.
- The paper introduces a new spatio-temporal attention module, which improves skeleton-based action recognition regardless of whether the skeletons are generated from RF or vision-based data.
- It also presents a novel multi-proposal module that extends skeleton-based action recognition to detect simultaneous actions and interactions of multiple people.

2. Related Works

(a) Video-Based Action Recognition: Recognizing actions from videos has been a hot topic over the past several years. Early methods use hand-crafted features. For instances, image descriptors like HOG and SIFT have been extended to 3D [6, 22] to extract temporal clues from videos. Also, descriptors like improved Dense Trajectories (iDT) [35] are specially designed to track motion information in videos. More recent solutions are based on deep learning, and fall into two main categories. The first category extracts motion and appearance features jointly by leveraging 3D convolution networks [5, 30]. The second category considers spatial features and temporal features separately by using two-stream neural networks [32, 36].

(b) Skeleton-Based Action Recognition: Skeleton-based action recognition has recently gained much attention [12, 4]. Such an approach has multiple advantages. First, skeletons provide a robust representation for human dynamics against background noise [23]. Second, skeletons are more succinct in comparison to RGB videos, which reduces computational overhead and allows for smaller models suitable for mobile platforms [20].

Prior work on skeleton-based action recognition can be divided into three categories. Early work used Recurrent Neural Networks (RNNs) to model temporal dependencies in skeleton data [9, 31, 48]. Recently, however, the literature shifted to Convolutional Neural Networks (CNNs) to learn spatio-temporal features and achieved impressive performance [8, 23, 20]. Also, some papers represented skeletons as graphs and utilized graph neural network (GNN)
for action recognition [38, 13]. In our work, we adopt a CNN-based approach, and expand on the Hierarchical Co-occurrence Network (HCN) model [23] by introducing a spatio-temporal attention module to deal with skeletons generated from wireless signals, and a multi-proposal module to enable multiple action predictions at the same time.

(c) Radio-Based Action Recognition: Research in wireless systems has explored action recognition using radio signals, particularly for home applications where privacy concerns may preclude the use of cameras [37, 14, 29, 1]. These works can be divided into two categories: The first category is similar to RF-Action in that it analyses the radio signals that bounce off people’s bodies. They use action labels for supervision, and simple classifiers [37, 14, 29, 1]. They recognize only simple actions such as walking, sitting and running, and a maximum of 10 different actions. Also, they deal only with single person scenarios. The second category relies on a network of sensors. They either deploy different sensors for different actions, (e.g., a sensor on the fridge door can detect eating) [19, 39], or attach a wearable sensor on each body part and recognize a subject’s actions based on which body part moves [21]. Such systems require a significant instrumentation of the environment or the person, which limits their utility and robustness.

3. Radio Frequency Signals Primer

We use a type of radio commonly used in past work on RF-based action recognition [45, 24, 41, 47, 28, 33, 16, 42, 46, 44]. The radio generates a waveform called FMCW and operates between 5.4 and 7.2 GHz. The device has two arrays of antennas organized vertically and horizontally. Thus, our input data takes the form of two-dimensional heatmaps, one from the horizontal array and one from the vertical array. As shown in Figure 2, the horizontal heatmap is a projection of the radio signal on a plane parallel to the ground, whereas the vertical heatmap is a projection of the signal on a plane perpendicular to the ground (red refers to large values while blue refers to small values). Intuitively, higher values correspond to higher strength of signal reflections from a location. The radio works at a frame rate of 30 FPS, i.e., it generates 30 pairs of heatmaps per second.

As apparent in Figure 2, RF signals have different properties from visual data, which makes RF-based action recognition a difficult problem. In particular:

- RF signals in the frequencies that traverse walls have lower spatial resolution than visual data. In our system, the depth resolution is 10 cm, and the angle resolution is 10 degrees. Such low resolution makes it hard to distinguish activities such as hand waving and hair brushing.
- The human body is specular in the frequency range that traverse walls [2]. RF specularity is a physical phenomenon that occurs when the wavelength is larger than the roughness of the surface. In this case, the object acts like a reflector - i.e., a mirror - as opposed to a scatterer. The wavelength of our radio is about 5cm and hence humans act as reflectors. Depending on the orientation of the surface of each limb, the signal may be reflected towards our sensor or away from it. Limbs that reflect the signal away from the radio become invisible for the device. Even if the signals are reflected back to the radio, limbs with a small surface (e.g., hands) reflect less signals and hence are harder to track.
- Though RF signals can go through walls, their attenuation as they traverse a wall is significantly larger than through air. As a result, the signals reflected from a human body is weaker when the person is behind a wall, and hence the accuracy of detecting an action decreases in the presence of walls and occlusions.

4. Method

RF-Action is an end-to-end neural network model that can detect human actions through occlusion and in bad lighting. The model architecture is illustrated in Figure 3. As shown in the figure, the model takes wireless signals as input, generates 3D human skeletons as an intermediate representation, and recognizes actions and interactions of multiple people over time. The figure further shows that RF-Action can also take 3D skeletons generated from visual data. This allows RF-Action to train with existing skeleton-based action recognition datasets.

In the rest of this section, we will describe how we transform wireless signals to 3D skeleton sequences, and how we infer actions from such skeleton sequences - i.e., the yellow and green boxes in Figure 5. Transforming visual data from a multi-camera system to 3D skeletons can be done by extracting 2D skeletons from images using an algorithm like AlphaPose and then triangulating the 2D keypoints to generate 3D skeletons, as commonly done in the literature [15, 45].
4.1. Skeleton-Generation from Wireless Signals

To generate human skeletons from wireless signals, we adopt the architecture from [45]. Specifically, the skeleton generation network (the orange box in Figure 3) takes in wireless signals in the form of horizontal and vertical heatmaps shown in Figure 2 and generates multi-person 3D skeletons. The input to the network is a 3-second window (90 frames) of the horizontal and vertical heatmaps. The network consists of three modules commonly used for pose/skeleton estimation [45]. First, a feature network comprising spatio-temporal convolutions extracts features from the input RF signals. Then, the extracted features are passed through a region proposal network (RPN) to obtain several proposals for possible skeleton bounding boxes. Finally, the extracted proposals are fed into a 3D pose estimation sub-network to extract 3D skeletons from each of them.

4.2. Modality-Independent Action Recognition

As shown in Figure 3, the Modality-Independent Action Recognition framework uses the 3D skeletons generated from RF signals to perform action detection. The framework has three modules as follows: 1) An attention-based feature learning network, which extracts high-level spatio-temporal features from each skeleton sequence. 2) We then pass these features to a multi-proposal module to extract proposals — i.e., time windows that each corresponds to the beginning and end of an action. Our multi-proposal module consists of two proposal sub-networks: one to generate proposals for single person actions, and the other for two-person interactions. 3) Finally, we use the generated proposals to crop and resize the corresponding latent features and input each cropped action segment into a classification network. The classification network first refines the temporal proposal by performing a 2-way classification to determine whether this duration contains an action or not. It then predicts the action class of the corresponding action segment.

Next, we describe the attention module and the multi-proposal module in detail.

4.2.1 Spatio-Temporal Attention Module

We learn features for action recognition using a spatio-temporal attention-based network. Our model builds on the hierarchical co-occurrence network (HCN) [48]. HCN uses two streams of convolutions: a spatial stream that operates on skeleton keypoints, and a temporal stream that operates on changes in the locations of the skeleton’s keypoints across time. HCN concatenates the output of these two streams to extract spatio-temporal features from the input skeleton sequence. It then uses these features to predict human actions.

However, skeletons predicted from wireless signals may not be as accurate as those labeled by humans. Also, different keypoints may have different prediction errors. To make our action detection model focus on body joints with higher prediction confidence, we introduce a spatio-temporal attention module (Figure 4). Specifically, we define a learn-
able mask weight $W_m$, and convolve it with latent spatial features $f_s$, and temporal features $f_t$ at each step:

$$Mask = \text{Conv}(\text{concat}(f_s, f_t), W_m).$$

We then apply the $Mask$ on the latent features as shown in Figure 4. In this way, the mask could learn to provide different weights to different joints to get better action recognition performance. We also add a multi-headed attention module [34] on the time dimension after the feature extraction to learn the attention on different timestamps.

Our proposed attention module helps the model to learn more representative features, since the learnt mask leverages information provided by both the spatial stream and temporal stream, and the multi-headed attention helps the model to attend more on useful time instances. This spatio-temporal attention changes the original HCN design where the spatial and temporal path interact with each other only using late fusion. Our experiments show that the spatio-temporal attention module not only helps increase the action detection accuracy on skeletons predicted from wireless signals, but also helps increase the performance on benchmark visual action recognition datasets. This further shows the proposed attention module helps to combine spatial and temporal representations more effectively, and would lead to better feature representations.

4.2.2 Multi-Proposal Module

Most previous action recognition datasets have only one action (or interaction) at any time, regardless of the number of people present. As a result, previous approaches for skeleton action recognition cannot handle the scenario where multiple people perform different actions simultaneously. When there are multiple people in the scene, they simply do a max over features extracted from each of them, and forward the resulting combined feature to output one action. Thus, they can only predict one action at a time.

However, in our dataset, when there are multiple people in the scene, they are free to do any actions or interact with each other at any time. So there are many scenarios where multiple people are doing actions and interacting simultaneously. We tackle this problem with a multi-proposal module. Specifically, denote $N$ to be the number of people appearing at the same time. Instead of performing max-pooling over $N$ features, our multi-proposal module outputs $N + \binom{N}{2}$ proposals from these $N$ features, corresponding to $N$ possible single-person actions and $\binom{N}{2}$ possible interactions between each two people. Our multi-proposal module enables us to output multiple actions and interactions at the same time. Finally, we adopt a priority strategy to prioritize interactions over single person actions. For instance, if there are predictions for ‘pointing to something’ (single person) and ‘pointing to someone’ (interaction) at the same time, our final prediction would be ‘pointing to someone’.

4.3. Multimodal End-to-end Training

Since we want to train our model in an end-to-end manner, we can no longer use $\arg\max$ to extract 3D keypoint locations, as in past work on RF-based pose estimation [45]. Thus, we use a regressor to perform the function of $\arg\max$ to extract the 3D locations of each keypoint. This makes the model differentiable and therefore the action label can also act as supervision on the skeleton prediction model.

Our end-to-end architecture uses 3D skeletons as an intermediate representation which enables us to leverage previous skeleton based action recognition datasets. We combine different modalities to train our model in the following manner: for wireless signal datasets, gradients back propagate through the whole model, and they are used to tune the parameters of both the skeleton prediction model and the action recognition model; for previous skeleton-based action recognition datasets, the gradients back propagate till the skeleton, and they are used to tune the parameters for the action recognition module. As shown in the experiments section, this multi-modality training significantly increases the data diversity and improves the performance of our model.

5. Experiments

5.1. Dataset

Since none of the available action detection datasets provide RF signals and the corresponding skeletons, we collect
Figure 5: Qualitative Results. The figure shows RF-Action’s output in various scenarios. The top two rows show our model’s performance in visible scenes. The bottom two rows show our model’s performance under partial/full occlusions and poor lighting conditions. The skeletons shown are the 2D projection of the intermediate 3D skeletons generated by our model.

...
Ground Truth Labels. To perform end-to-end training of our proposed RF-Action model, we need two types of ground truth labels: 3D human skeletons to supervise our intermediate representation, and action start-end time and category to supervise the output of our model. The 3D skeletons are triangulated using AlphaPose and the multi-view camera system described earlier. As for actions’ duration and category, we manually segment and label the action of each person using the multi-view camera system.

5.3. Qualitative Results

Figure 5 shows qualitative results that illustrate the output of RF-Action under a variety of scenarios. The figure shows that RF-Action correctly detects actions and interactions, even when different people perform different actions simultaneously, and can deal with occlusions and poor lighting conditions. Hence, it addresses multiple challenges for today’s action recognition systems.

5.4. Comparison of Different Models

We compare the performance of RF-Action to the state-of-the-art models for skeleton-based action recognition and RF-based action recognition. We use HCN as a representative of a top performant skeleton-based action detection system in computer vision. It currently achieves the best accuracy on this task. We use Aryokee [33] as a representative of the state-of-the-art in RF-based action recognition. To our knowledge, this is the only past RF-based action recognition system that performs action detection in addition to classification.

Table 1 shows the results for testing on visible scenes and through-wall scenarios. This shows that the additional supervision from the skeletons, as well as RF-Action neural network design, are important for delivery of accurate performance using RF data.

5.5. Comparison of Different Modalities

Next, we investigate the performance of RF-Action when operating on RF-based skeletons versus vision-based skeletons. We train RF-Action on the training set, as before. However, when performing inference, we either provide it with the input RF signal from the test set, or we provide it with the visible ground truth skeletons obtained using our camera system. Table 2 shows the results for different input modalities. The table shows that for visible scenes, operating on the ground truth skeletons from the camera system leads to only few percent improvements in accuracy. This is expected since the RF-skeletons are trained with the vision-based skeleton as ground truth. Further, as we described in our experimental setting, the camera-based system uses 10 viewpoints to estimate 3D skeletons while only one wireless device is used for action recognition based on RF. This result demonstrates that RF-based action recognition can achieve a performance close to a carefully calibrated camera system with 10 viewpoints. The system continues to work well in through-wall scenarios though the accuracy is few percents lower due to the signal experiencing some attenuation as it traverses walls.

5.6. Action Detection

In Figure 6 we show a representative example of our action detection results on the test set. Two people are enrolled in this experiment. They sometimes do actions independently, or interact with each other. The first row shows the action duration for the first person, the second row shows the action duration for the second person, and the third row shows the interactions between them. Our model can detect both the actions of each person and the interactions between them with high accuracy. This clearly demonstrates that our multi-proposal module has good performance in scenarios where multiple people are independently performing some actions or interacting with each other.

5.7. Ablation Study

We also conduct extensive ablation studies to verify the effectiveness of each key component of our proposed ap-
Our model can achieve much higher performance. This is because our dataset includes a lot of inaccurate joints. We also conduct experiments on the NTU-RGB+D [31] dataset. Unlike PKU-MMD and RF-MMD which allow for action detection, this dataset is valid only for action classification. The table shows that our attention module is useful in this case too.

### Multi-Proposal Module
We propose a multi-proposal module to enable multiple action prediction at the same time. We evaluate our model’s performance with or without the multi-proposal module. As shown in Table 4, the added multi-proposal module significantly increase the performance. This is because our dataset includes a lot of instances when people are performing different actions at the same time. Our model will get very poor accuracy at these scenarios with single-proposal, while with multi-proposal our model can achieve much higher performance.

<table>
<thead>
<tr>
<th>Methods</th>
<th>RF-MMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-Proposal</td>
<td>87.8</td>
</tr>
<tr>
<td>Single-Proposal</td>
<td>65.5</td>
</tr>
</tbody>
</table>

Table 4: Benefits of Multi-proposal Module. The table shows adding multi-proposal module largely improves the performance on RF-MMD.

### Multimodal Training
As explained earlier, the use of skeletons as an intermediate representation allows the model to learn from both RF datasets and vision-based skeleton datasets. To illustrate this advantage, we perform multimodal training by adding PKU-MMD’s training set into the training of our RF-Action model. More specifically, we use our dataset to train the whole RF-Action end-to-end model, and use the PKU-MMD dataset to train RF-Action’s activity detection model. These two datasets are used alternatively during training. As shown in Table 5, comparing the detection results with the model trained on either dataset separately, we find multimodal training can increase the model performance since it introduces more data for training and thus can get better generalization ability.

### End-to-End Model
RF-Action uses an end-to-end model where the loss of action recognition is back propagated through the skeleton generation network. Here we show that such an end-to-end approach improves the skeleton itself. Table 6 reports the average error in skeleton joint location, for two systems: our end-to-end model and an alternative model where the skeleton is learned separately from the action —i.e., the action loss is not propagated through the skeleton generation network. The table shows that the end-to-end model not only improves the performance of the action detection task, but also reduces the errors in estimating the location of joints in RF-based skeletons. This is because the action detection loss provides regularization for 3D skeletons generated from RF signals.

<table>
<thead>
<tr>
<th>Methods</th>
<th>mAP</th>
<th>Skeleton Err. (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>end-to-end</td>
<td>87.8</td>
<td>3.4</td>
</tr>
<tr>
<td>separate</td>
<td>84.3</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Table 6: mAP and intermediate 3D skeleton error on testing data with and without end-to-end training.

### Conclusion
This paper presents the first model for skeleton-based action recognition using radio signals, and demonstrates that such model can recognize actions and interactions through walls and in extremely poor lighting conditions. The new model enables action recognition in settings where cameras are hard to use either because of privacy concerns or poor visibility. Hence, it can bring action recognition to people’s homes and allow for its integration in smart home systems.
References


[27] Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. Actionflownet: Learning motion representation for...


