Now You See It: Convolutional Radio Object detection

Aditya Singh
Computer Science Department
Ashoka University
Sonepat, India

Dr. Debayan Gupta
Computer Science Department
Ashoka University
Sonepat, India

***Abstract*—Radio waves are electromagnetic radiation with wavelengths longer than the infrared light (3 KHz to 300 GHz). Radio waves are generated by transmitters (Wifi Access points, RTL-SDR) and received by radio receivers using antenna. High Frequency radio waves have been used to detect objects using RADAR and SONAR by processing the received attenuated reflected signals, the process usually involves extensive signal processing. Specifically, WiFi signals (900 MHz to 6 GHz) have been used to localise and identify suitable objects because of their high frequency (low wavelength) and ubiquitous availability in indoor areas. Artificial Neural networks (ANNs) are known to be good functional approximates. Moreover, Deep Neural Networks have been widely successful in extracting features from temporal and spatial translation invariant data (visual and audio). In this article, we study the adaptation of convolutional neural networks and machine learning techniques to the complex-valued temporal radio signal domain for detection of objects. We review and analyze the use of ANNs in signal processing, specifically for object (and human body) detection. We compare the results of previous research using ANNs vs. pure signal processing for entity detection and monitoring. We also present our results on using low frequency waves (<500 MHz) for simple object detection in a controlled environment. Our results indicate the importance of using high frequency signals for capturing the intricacies of reflecting objects in the reflected signals, and encourage the use of ANNs for RF-based entity detection.**

***Index Terms*—RTL-SDR, WiFi signals, ANN**

I. Introduction

The ability to detect and localise objects, specifically human bodies is useful for several applications. For instance, smart buildings and cities can optimise energy consumption based on the number of people in specific regions. Estimating the number of people in a region has even more applications for ensuring the proper practice of social distancing norms during a pandemic [1]. Moreover, human-activity recognition for monitoring the movement and behaviour of humans has applications in health monitoring, emergency detection (for elderly) and enhancing virtual reality experiences without additional hardware. Sleep stage tracking is another benefit of human body monitoring to better understand sleep patterns and diagnose sleep disorders.

Over the past decade deep learning has fueled advances in the domain of object classification, human body recognition, human pose estimation and human action recognition from visual data. Photographic images of an area can now be used to identify the number of people present in the area, and the proximity amongst them [1], [2]. Depth estimation of a recognised entity has also been shown to be possible through deep learning. However, these approaches towards human-body monitoring suffer from several intrinsic limitations-1) They require a number of cameras to be installed in an area to achieve an adequate performance and cover all blind-spots, this results in a high deployment cost. 2) They suffer from the same limitations humans suffer from, our eyes cannot perceive visual information in the dark or see through walls and occlusions. 3) Such methods for monitoring pose immense privacy issues. The existence of comprehensive facial recognition technology means that any monitoring done using visual data and without proper consent compromises the right to privacy of the public.

Fortunately, visible light is just one end of the frequency spectrum. Recent advances in wireless research have shown that certain Radio Frequency Signals can traverse through walls and occlusions, and are reflected by the human body (and any other electricity conducting material). If one is able to interpret the reflected signals, one can perform human-body and object recognition through walls and occlusions, and in the dark. This method operates in a similar fashion to Radar and Sonar, but at a much lower power. Radar and Sonar were the first systems to use RF reflections to detect and track objects. However, they mostly have military applications and focus on inanimate objects (planes,metallic objects,etc.). Radar literature that deals with object detection and localisation typically operates at a much higher frequency (Terahertz or millimetre and sub-millimetre waves), where the wavelength is comparable to the roughness of the surface (object becomes scatterer as opposed to a reflector), and hence radar is able to image the object and determine its location with high accuracy. However, such systems operate at very short distances and cannot deal with occlusions. Some Radar systems use centi-metre waves (carrier frequency around a few GHz, similar to what we consider in this article), however, they seem to have a significantly lower resolution as compared to modern RF-based systems that we review in this paper [3], [4].

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public

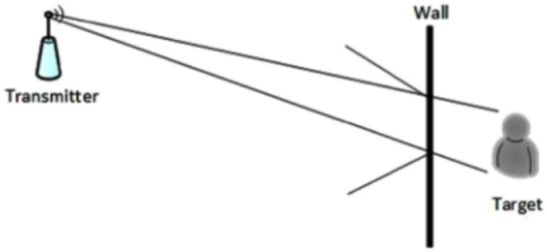

Fig. 1. Reflection of transmitted RF signal through wall

The use of wireless devices for human-body/object detection and monitoring can be broadly classified as (i) Device-based active and (ii) Device-free passive methods. Device-based active methods require the users (or object) to carry a wearable/attachable wireless communication device (might include a gyroscope, accelerometer, etc.) [5]–[7]. The device is then able to emit a signal with the relevant data which is used by the receiver to localise the entity and determine other useful properties (relative location, velocity, direction, orientation, etc). Such systems require significant instrumentation of the person/object and the environment in which it will be used, which limits their utility and robustness. Moreover, this method is not scale-able for detection of novel entities that enter the region being monitored and only works on entities equipped with the device. Due to these limitations, recently there has been a considerable interest in device-free methods, which do not have any prior requirement from the entities they aim to monitor/detect. Instead, such methods rely on the interaction of a transmitted wireless signal with the relevant entities in the area of interest. Systems like the Microsoft Kinect (Infrared waves- 300 GHz to 430 THz) have revolutionised the field of Human-computer interaction by enabling 3-D motion tracking without instrumenting the body of the user. However, they suffer from the same limitations that are encountered by visual-data based approaches, i.e they require the user to remain within the line-of-sight and they cannot track across walls. Thus, there has been a recent push in research to build device free methods that can overcome these restrictions. Typical WiFi signals operate in the frequency range of Radio waves (900 MHz to 6 GHz), and hence can exploit the ability of RF waves to penetrate through non-conductive material (wood, cement, brick), be reflected off of electrically conductive material (water, humans, metal). This factor, combined with the ubiquity of WiFi signals has resulted in an increasing interest in the development of efficient and practical methods for detection and monitoring of entities(humans/objects) using Radio signals in the WiFi frequency.

There has been encouraging success in interpreting the RF signal reflections from metal/water objects and human bodies to precisely determine the location and number of reflecting entities in indoor areas. The research has focused on either using extensive signal processing techniques, or Artificial Neural Networks (ANNs) to extract information about the reflecting entities from the received attenuated signals. In this paper we review the surge of machine learning for signal processing applications, in general and for monitoring the surrounding reflecting objects using RF signals in the WiFi frequency. We also review the results achieved by pure signal processing methods. We then present the results of our experiment on using Machine Learning techniques for detecting water based objects using low-frequency radio waves (433 MHz) emitted by a software defined radio. Our results highlight the importance of using high frequency radio waves, such as those operating on the WiFi frequency channel, for object detection, and encourage the future use of ANNs to achieve better results in this area.

II. Literature Reviewed

An RF signal is a wave whose phase is a linear function of the traveled distance. By sampling the signal, we can record both its amplitude and its phase. The sampled signal can be represented as a complex discrete function of time t as follows:

$$s_t = A_t e^{-j2\pi \frac{r}{\lambda} t}$$

where r is the distance traveled by the signal, λ is its wavelength, and A is its amplitude.

We review the recent research aimed at leveraging RF signals in the WiFi frequency for the detection of reflecting entities in indoor areas. This section is split into two parts, the first focuses on pure signal processing based research, and the second discusses machine learning based methods.

A. Signal Processing based methods

Work in this area utilizes the IEEE 802.15 .4 protocol, and transmit in the 2.4-8 GHz frequency band. Multiple nodes are places strategically to cover a large area. J. Wilson and N. Patwari use a Radio Tomography based approach to measure the attenuation in the transmitted signal through a medium [8]. The variance in received signal strength (RSS) is known to be related linearly to the power contained in non-static multi-path components. The variance of RSS relates to the location of movements of reflecting objects relative to node locations. They apply a Kalman filter to filter out noise and track movement coordinates from image data. Using a 34 node system to monitor a 700-sq feet house, they are able to track the movements of a single human body through occlusions with an average error of 1.5ft. However, this VRTI (Variance Radio Tomographic Imaging) based approach requires a setting up of the large node based network, and is only able to operate when changing multi-path power is greater than static multi-path power (condition for linearity of variance with power of non-static components).

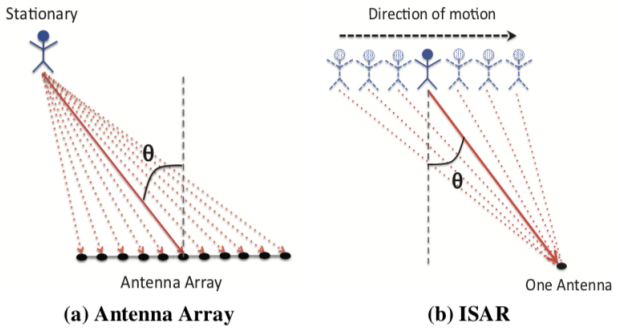

Fig. 2.

In contrast, F. Adib and D. Katabi employ a 3 antenna multi-input multi-output (MIMO) radar based approach. Such systems also use distributed devices (and require a lesser bandwidth compared to RTI) and instead measure the scattering of the transmitted signal by the object of interest to track it [9]. Their system, Wi-Vi, operates in the ISM band using typical WiFi hardware (Access points). A common obstacle in using such MIMO systems is the flash effect, since only a fraction of the RF signal is able to traverse through walls, the signal reflected off of the object/body reduces in power by three to five orders of magnitude. This makes it harder to register the minute variations from objects behind the wall using a low power system in compliance with FCC regulations (military systems are able to use Ultra wide band systems with 2 GHz of bandwidth). Wi-Vi is able to use the nulling effect of MIMO systems to eliminate static reflections (including the wall), essentially the signal transmitted from multiple antennas is encoded in such a way that it is nulled in a particular receiving antenna. They then employ a technique called inverse aperture synthetic radar (ISAR) to emulate an antenna array, thereby reducing the need for a large number of antennas by interpreting each moving human as an separate antenna array. Wi-Vi is able to achieve great accuracy for detecting 3-5 non-static humans through walls at a distance of 6-10 metres, it can also perform gesture recognition at a distance of 7 metres. However it does not operate through walls thicker than 8".

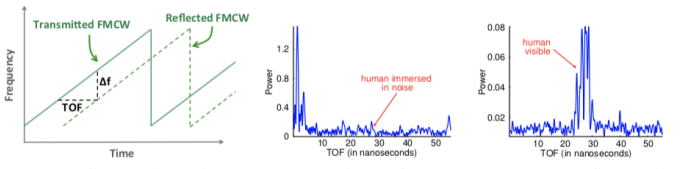

Fig. 3.

Building from Wi-Vi, Adib et. all further propose WiTrack 2.0 which is a multi antenna MIMO system (5 antennas and 5 recievers) [10]. They address the multi-path multi user problem by obtaining Time of Flight (related to power density, as TOF depends on how much power is being transmitted back after reflection) measurements from different vantage points using a Frequency Modulated Carrier Wave (FMCW). FMCW is a technique that allows a radio device to measure the depth of an RF reflector. An FMCW device transmits a frequency chirp i.e., a periodic RF signal whose frequency linearly increases in time, as shown in the figure. The chirp reflects off objects in the environment and travels back to the device after the time-of-flight. The device can measure the time-of-flight and use it to infer the depth of the reflector. To do so, the device leverages the linear relationship between time and frequency in chirps. Specifically, it measures the time-of-flight (and its associated depth) by measuring the frequency shift between the transmitted and received signal. Following this they introduce a Successive Silhouette Cancellation approach to localise each human. This technique starts by localising the closer body first (highest relative power in its received reflection), then eliminating its impact on the received signal before localising further users. This process goes on iteratively until all the users are localised. For localising static users WiTrack 2.0 detects the movement caused by breathing. By processing the reflected signals at multiple time scales it is able to localise both extremes of movements (breathing and walking). WiTrack 2.0 operates at 5.4-7.25 GHz (receiving every 2.5 milliseconds) and is able to localise up to 5 static and non-static users in a range of 10 metres through walls with a median accuracy of 11.7 cm in each of the x-y dimension.

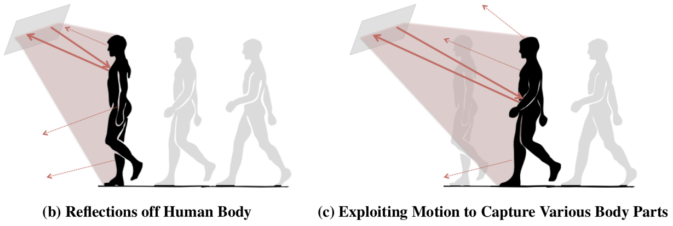

Fig. 4. Reflection of transmitted RF signal through wall

Furthermore, they are able to exploit the fact that the reflecting limbs vary in a moving person to stitch together the captured reflection (across time) and reconstruct a coarse skeleton of a human body [11]. They use a similar, albeit larger setup (more antennas and receivers) to WiTrack 2.0, and this system, called RF-capture, is able to perform at par with the Kinect, without its limitations, and additionally even distinguish between users (within a range of 10 metres). However, RF-capture is limited to a single person performing a single action, which is to walk towards the device. Further, it cannot simultaneously localize multiple key points on the human body.

Cushman et. al. are able to perform through wall human body detection within a range of 10 metres employing

considerably less hardware [12]. They use a National Instrument Universal Software Radio peripheral devices (USRP) and a single antenna for receiving, the system operates in the ISM bands 2.4-2.5 GHz and 4.9-5.85 GHz. The data sent by the transmitter is compared with the received I/Q data and the dissimilarities are exploited to classify the reflecting non-static entity. After de-noising the dissimilarities, they are interpreted on a time-frequency waveform (Short-Time Fourier Transform or Gabor transform), on which peaks are observed to detect a moving object/body. A disadvantage of this method is that it does not infer anything about the location or state of the entity, but is only able to detect whether there is movement or not. Nevertheless the portable nature of the setup could prove to be beneficial for military purposes.

PASSIVE RADAR-

B. Machine Learning based methods

The surge in digital data, combined by the availability of large computation power has allowed Artificial Neural Networks to emulate several tasks earlier associated exclusively with human intelligence. The ability of Convolutional Neural Networks to provide feature translation invariance has further propelled the use of ANNs for learning from large unstructured data in fields of computer vision, natural language processing and voice recognition. Deep Neural networks have proved to be a viable approach for voice processing on raw time-domain waveforms, reducing the tedious process of feature engineering [13], [14]. The field of Radio communications present a unique signal processing domain with a number of interesting challenges and opportunities for the machine learning community.

A recent application of ANNs in signal processing has been radio modulation recognition. It has been shown that relatively simple CNNs have been able to outperform algorithms with decades of expert feature searches for radio modulation recognition [15]. Timothy et. al. were able to achieve a state of the art accuracy by training a CNN on raw I/Q samples (WiFi enabled SDR operating at 900 MHz) normalised to unit variance, collected across 11 different modulations (8 digital and 3 analog). Even for high signal to noise ratio, the CNN achieved a better accuracy over a larger coverage area. They were able to further better the accuracy by leveraging deep residual and Long short-term memory architectures which have previously performed better on visual and time-series data [16]. Furthermore, Mingmin et. al. were able to identify sleep stages and sleep quality across multiple domain environments by using a conditional adversarial deep neural network trained on raw RF reflections obtained using an Electroencephalography(EEG) based sleep monitor [17]. In addition, CNNs have also been successfully trained on raw power spectral density (PSD) features of EEG signals emitted by the brain to determine what action the user is thinking about. It is known that the brain emits specific signals when one imagines a particular action being performed. Perez-Zapata et. al. were able to use this detail to differentiate between thoughts about hand movements, feet movements and tongue movements [18].

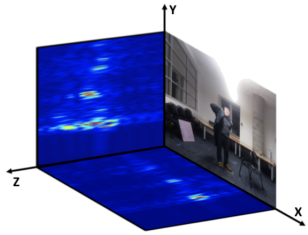

Fig. 5. RF heatmaps and an RGB image recorded at the same time.

Building from RF-capture, Mingmin et. al. introduced RF-pose3D which is able to provide a significant improvement in RF-based sensing by incorporating ANNs [19]. RF-pose3D takes as input the 4D RF signal captured by an FMCW radio (window of 3 seconds), similar to the radio used by RF-capture and WiTrack (5.4-7.2 GHz) [10], [11]. The labels corresponding to the 4D RF signals are obtained by triangulating the positions of keypoints of multiple humans in 3-D space across time using a multi-camera system (video+location of keypoints of humans in the video). A traditional CNN however, cannot perform 4D convolutions, and operates only up to 3D convolutions (tailored for visual data). To address this challenge, RF-pose3D decomposes the 4D convolutions into a combination of 3D convolutions performed on two planes and the time axis. They provide a rigorous proof for the equivalence of their decomposition to 4D convolutions. Likewise, CNN feature-label pairs are also decomposed to operate on two planes. The system has two arrays of antennas organized vertically and horizontally. As shown in Figure 5, 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. The model operates using a regular softmax loss and is able to predict the location of each keypoint in space as the voxel with the highest score. This model is able to accurately track the motion and presence of a single person, to scale this to multiple people, the authors employ a rather clever technique. They design a separate region proposal network (RPN, designed as a regular CNN) which generates potential people regions as 2D bounding boxes on the horizontal plane for the CNN to operate on. The RPN operates on the output of the CNN (in the horizontal plane) from an intermediate layer (feature map). The inspiration behind this is twofold, firstly, since the original CNN is able to perform well, it is safe to assume that the features in the intermediate layers are able to condense the information and remove irrelevant information effectively. Secondly, in case one person is occluding another one (and hence hiding their reflections), performing 4D convolutions and combining the information across space time will allow

the RPN to detect a temporarily occluded person, and return that information as its output to the CNN for training and prediction. Furthermore, it is viable to operate solely on the horizontal place for region proposal as it is reasonable to assume that a person will not be standing on another one. The two networks (CNN and RNN) are thus jointly trained as shown in the figure.

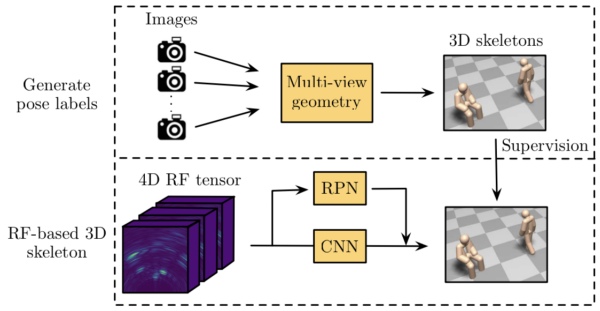

Fig. 6. The labeled samples are used to train the model in the bottom graph. The model can be divided into two components: a region proposal network (RPN) that zooms in on RF data from one individual, and a CNN that extracts the 3D skeleton from the proposed region.

The resulting network is able to accurately detect the 3D skeleton of multiple people and simple actions (walking, sitting, standing) over a range of 40 feet across multiple environments. Li at. al. further augment this system to detect multiple actions by performing multi-modal training using the obtained 3D skeleton from RF-pose3D and existing vision based action recognition datasets [20]. The spatial-temporal attention module which is trained on the multi-modal data is then able to perform action recognition on visual and RF based data. Further, they found that transferring knowledge related to action recognition across modalities is able to empirically improve performance regardless of whether the skeletons are generated from RF or vision based data. An additional advantage of incorporating the knowledge of human labelled vision based 3D skeleton datasets is that they are able to leverage its knowledge about interactions between different people to recognise simple interactions from RF-data as well. This system, called RF-action is able to accurately recognise actions and interactions (29 single actions and 6 interactions) of multiple humans through walls and in the dark (up to 40 feet) and represents a significant improvement in action recognition capabilities.

Saandeep Depatla and Yasamin Mostofi tackled a more specific problem of crowd counting using received signal strength measurements (RSSI) [21]. They use two WiFi nodes placed on opposite ends of the room, one transmits constantly at a frequency of 2.4 GHz and the other node records the received signal strength. They propose that the inter-event times, corresponding to the dip events of the received signal, are fairly robust to the attenuation through walls and carry information about the total number of moving people in an area. They are able to model the received signal as a superposition of the effects of the number of people present in an area. They then derive the Probability Mass Function of the inter-event times based on this model, and use it to estimate the number of people using a maximum likelihood (ML) estimator. Their system is able to accurately identify up to 20 people in large classrooms and hallways with a high accuracy.

Yousefi et. al. incorporate deep learning algorithms for behaviour recognition using Channel State information (CSI) [22]. The main advantage of using CSI over RSS measurements is that in RSS, the changes caused by reflecting objects are averaged out over all the WiFi bandwidth being used and hence cannot capture the change at certain frequencies. By obtaining the CSI for each sub-carrier (which faces a narrowband channel) however, one can quantify the diversity in observed channel dynamics. They feed in as input raw CSI amplitude data into an RNN LSTM, which has been shown to achieve state of the art performance on time-varying tasks (and requires no feature engineering). The LSTM is able to easily outperform other machine learning techniques that require feature engineering (Random Forest, Hidden Markov Model) for the task of individual activity recognition (6 activities) of a single user within 3 metres. They use a single receiver and transmitter (5 GHz) system (placed on opposite ends of the room) and suggest that this approach can be extended to multiple people by using a MIMO radar setting. Thier system, however, is not adaptable to multiple environments and requires further training of the LSTM to work across different areas.

Zhihao et. al. propose Find-it as an improvement to Ultra-Wideband techniques for human detection [23]. As we discussed earlier, WiTrack 2.0 and Wi-Vi use Ultra-Wideband to generate FMCW at the cost of 1.7 GHz bandwidth (a huge limitations as such huge bandwidths are not available everywhere) [9], [10]. Find-it operates using narrow band RF-signals generated using an USRP SDR. The setup consists of two transmitters and one receiver (placed side by side) and uses the nulling method employed in [9] and [10] to eliminate reflections from static objects. They additionally magnify the differences between static and moving objects by applying a time-frequency transform (STFT) and lower the bandwidth to 1 MHz to decrease computation and deployment cost. They are able to use the STFT graphs to detect moving bodies with an accuracy of 90%. The Doppler effect is commonly used to recognize the relative direction of movement. However, the sample rate for Find-it (1MHz) is too high to disambiguate slight Doppler shifts. So, the researchers, using the same idea as [22], they resample the received signals and extract corresponding channel state information (CSI). By doing time-frequency transform for CSI signals and using their phase information, they obtain slight Doppler shifts, and are able to know that the moving human is near or far away from the receiver. Furthermore, to adapt Find-it across multiple

environments they cluster the data obtained by the SFFT into two categories, label them (human,no-human) and feed it to a Support Vector Machine (SVM) Classifier. The SVM is then able to detect the presence of a human in multiple environments in the range of a few metres (not specified).

In the following section we take inspiration from the above approaches, specifically deep learning, and try to use low frequency RF reflections to detect an electrically conducting object.

III. Experiment

We operate using a Raspberry Pi and a dipole antenna as the transmitter and a RealTek Software Defined Radio (RTL-SDR) to receiver and demodulate the the signal. We use the python library pyrtlsdr to interface the sdr with a macbook pro (64-bit, 3.1 GHz Intel Core i5). The data is collected in raw Inphase-Quadrature format over time, as also done in [12], [15], [19], [20]. I/Q Data consists of I and Q represented as two separate variables (a vector of length two) in the form of the complex number I + Qi (I-real,Q-Imaginary). I/Q Data as the signal representation is much more precise than just using a series of samples of the momentary amplitude of the signal because it is able to capture the complete function of the signal as a corkscrew (helix, spiral, coil spring) in three dimensions.

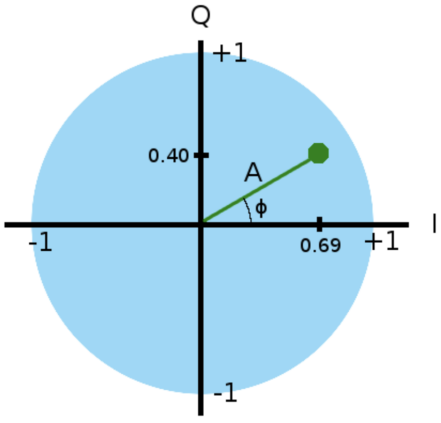

Fig. 7. I is the current momentary amplitude of the signal (i.e. the Real signal) Q is the momentary amplitude of the signal phase shifted -90 degrees.

I/Q data can thus be processed using Matlab or Scipy python libraries to visualise how the received power of the signal varies (RSS) in the time-frequency domain by applying SFFT. The Short-Time Fourier Transform, STFT, determines the sinusoidal frequency and phase content of local sections of a signal as it changes over time. It divides a longer time signal in to shorter segments with the same length and then computes the Fourier transform on each shorter segment separately. It can then be used to quantify how the power of the signal varies with time. In this project we use the

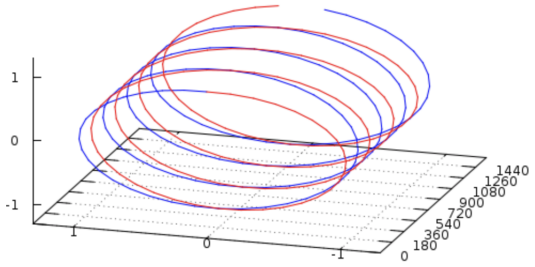

Fig. 8. The I/Q signals in 3D

Welch's periodogram (matlab) to compute the power and PSD since it has shown to provide better results when used with ANNs [18]. Further, we also use the SFFT to obtain heatmaps (Spectrograms) as they have been successfully used for RF object detection with deep learning [19], [20], [22].

A spectrogram is a visual representation of the Short-Time Fourier Transform. Chunks of the input signal are taken and local Fourier Transform is applied on each chunk. Each chunk has a specified width, an associated frequency distribution and a Fourier Transform is applied to this chunk. For each chunk that is centred at a specific time point in the time signal, we obtain a bunch of frequency components. The collection of all of these frequency components at each chunk is plotted all together and results in a spectrogram.

The spectrogram is then a 2D visual heat map where the horizontal axis represents the time of the signal and the vertical axis represents the frequency axis. In the visualised image the darker colours correspond to lower magnitude of that frequency component (power), and the lighter colours correspond to a higher magnitude in the frequency component.

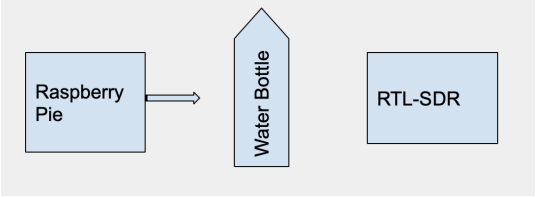

Fig. 9. Pi emitting at 433 MHz narrowband and SDR recieving the reflected signals (placed about a metre apart)

A. Setup and Data collection

We try to differentiate between the presence of a water bottle (tap water, hence conducting material), a metal bottle vs. no object being placed between the transmitter and receiver. The data is collected in an isolated room with no other non-static reflecting material inside. The setup is illustrated in figure 9, the receiver and transmitter are placed about a metre apart. The Raspberry Pi board is programmed to emit a constant low power RF [11], [18] signal centred at 433 MHz (which is

the highest frequency our SDR can receive) with a bandwidth of 8 MHz. The signal is received and sampled by the SDR at 2.6 MS/s [12], [15], [22], with each recorded sample being 0.5 seconds long. We collect 1000 such samples for each case, namely "water bottle present" and "No object present".

B. Preprocessing and CNN

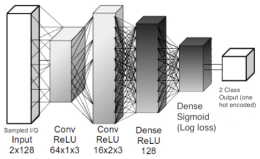

Fig. 10. The I/Q signals fed to CNN1

Convolutional Neural Networks are able to extract feature maps from non-processed data obtaining high level abstractions over input data by using trained filters. The weights of these filters are corrected during the training process by stochastic gradient descent or any other variation of the gradient descent algorithm. We preprocess the data and feed it to to a Convolutional Neural Network similar to the one used in [15] [18] [19], emulating the VGG architecture which has been able to achieve state of the art accuracy on vision recognition tasks, but scaled down to correspond to the size of our dataset. We train against several larger candidate neural networks and find no architecture dependent difference in the obtained results. Moreover, CNN performance have known to not be improved by the number of deep layers after a point. For the sake of recording results we adopt the architecture shown in figure 10 with 256 neurons in layer 1, 128 neurons in layer 2 and 256 neurons in layer 3, sigmoid being the activation of the last layer for multi-class predictions. 40 % of the neurons are dropped after the final convolutional layer to avoid overfitting. Relu activations are used in each layer to introduce non-linearity, a glorot uniform kernel initialiser is used to initialise the convolutional layers and the network is trained using a categorical cross entropy loss function with an adam optimiser. We feed the CNN with three representations of the received signal.

Raw I/Q data-

We provide the network with the raw time series radio signal by treating the complex I/Q data as dimension of 2 real valued I/Q inputs to the CNN, this approach has been most commonly used for feature learning in signal processing. It has been found to be a promising technique for feature learning on large time series data. We generate 10,000 samples of length 1024 for each class ('bottle','no-bottle') with some overlap to augment the dataset .

Power Spectral Density-

We generate the power spectral density function for each 0.5 second sample and calculate the power values at different frequencies in time. We obtain a 20*100 length vector capturing the variation of PSD values for 433 ± 1MHz. This method has also been used in [18]. Moreover, power of the reflected RF signal has been used widely to detect the object. We obtain 1,000 samples for each class which are used to train the CNN. Power spectral Density plots for each case are shown in figure 11.

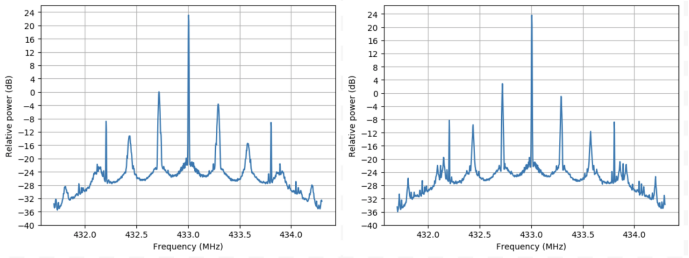

Fig. 11. PSD of received signals, on the left is PSD graph without the object, the right is the PSD graph with object. No visually observable difference is observed.

Heatmaps/Spectrograms-

We apply SFFT to each sample to convert it into a spectrogram representation in the frequency-time domain (50% overlap in each chunk). We obtain 1,000 samples for each case ('bottle','no bottle'). Combined spectrogram for four samples (each case) is shown in figure 12. The image is cropped and resized to 224*224, and then the RGB values are converted to floating point integers and normalised (divided by 255.0), the resulting floating point matrices of size 244*244*3 are fed into the network.

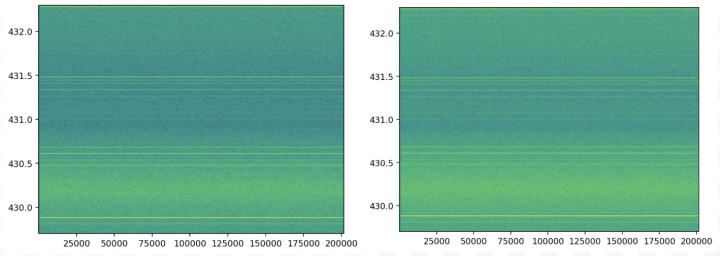

Fig. 12. Heatmap of received signals, frequency(in MHz)-Time(ms). On the left is spectrogram without the object, the right is the spectrogram with object. No discernible difference is observed.

Our goal of detecting whether the bottle is present or not is a multiclass problem, hence we use a one-hot-encoded vector of length 2 as the label for each sample ([1,0]$-$ >no object,[0,1]$-$ >object present). Thus the final layer output of the network has two neurons. For each type of training data, 10% of the data is kept for testing and the rest is used for training, furthermore 10% of the training data is used for validation while training.

Test set accuracy

Data	Accuracy
Raw I/Q	0.44 (overfitting)
PSD	0.51 (overfitting)
Spectrograms	0.61 (overfitting)
Spectrograms (CNN2)	0.96

C. Observations

We observe that our model is extremely prone to over fitting. It is not able differentiate between the classes at all and overfits on a single class. We also test our data using a shallow multi-layered perceptron and an SVM which give us worse results (the advantage of deep learning is the elimination of feature engineering, which is required in MLP and SVM). In addition, there are no discernible difference between the two cases, in contrast to the observations of some of the researchers we reviewed. This suggests that the data we have collected is extremely noisy, and contains very little information about the object we are trying to detect. We speculate that the received signal contains very little information about what is placed (or not placed) in front of the transmitter. The most glaring reason for this could be the relatively low frequency of our emitted signal. All the experiments we have studies operate in the range of 1-8GHz, which corresponds to a wavelength in centimetres (30-3 cm). A low wavelength is very desirable as only if the wavelength is comparable to the size of the object, will the signal collide sufficiently with the reflecting material and capture its intricacies in the wave that is bounced off (scattered). We operate using a frequency of 433 MHz, which corresponds to a wavelength of roughly one metre. The size of the water bottle is barely one-tenth of a metre (its width is around 8 cm), hence it is safe to assume that minimal interaction occurs between our transmitted signal and the water bottle kept in front. The change in the received wave due to the presence of the object is also then minimal, and so is the information it contains about the object.

We do however achieve some success while operating on the spectrograms using a larger deep network (CNN2) with softmax outputs and Binary Cross-entropy loss (even with CNN2, results obtained using raw I/Q samples and PSD data are the same). Softmax outputs are more suitable for this multiclass problem as softmax enforces some dependency upon the predictions. The network architecture is visualised using keras in the appendix. We observe that the network is able to locate a global minima on the loss surface and create a generalised decision boundary that achieves a high accuracy on the test set. The graphs for loss and accuracy are also shown below in figure 13 and 14. We also visualise (and analyse) the feature maps of the convolutional layers on a sample from the test set in the Appendix. Furthermore, the confusion matrix is also shown in figure 15, and we obtain a clean diagonal. The CNN is thus able to successfully differentiate between the presence and absence of a water bottle from the spectrogram representation of the received signal. However, this accuracy is achieved once in every alternate instance of training (overfitting still occurs), which further supports our speculation about the extreme noise and very little relevant information in data which makes it harder for the CNN to discover useful features. We also collected data in the presence of an empty metal bottle as opposed to a water bottle, no interesting results were observed using its data and hence we refrain from presenting them here.
We hope to carry out experiments while operating with a larger frequency and multiple antennas in the future.

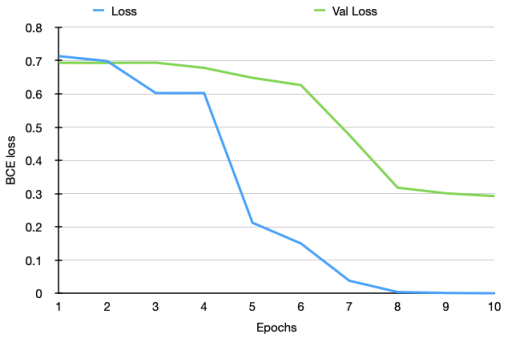

Fig. 13. Loss vs Epoch graph for CNN2 (spectrogram)

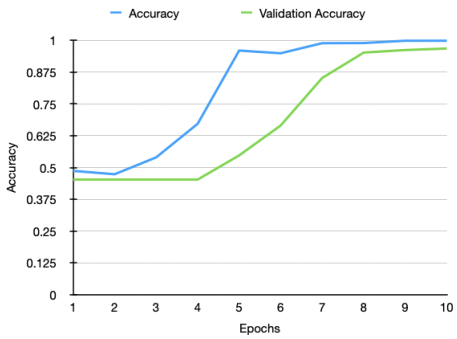

Fig. 14. Accuracy vs Epoch graph for CNN2 (spectrogram)

IV. Conclusion

The ability of RF waves to pass through walls, operate in the dark and be reflected off of human bodies (and other electrically conducting objects) presents an exciting opportunity to enhance the existing tracking technologies. Further, the ubiquity of WiFi signals makes RF-signals in the WiFi channel a natural choice for this purpose. They also provide an added advantage of providing privacy as opposed

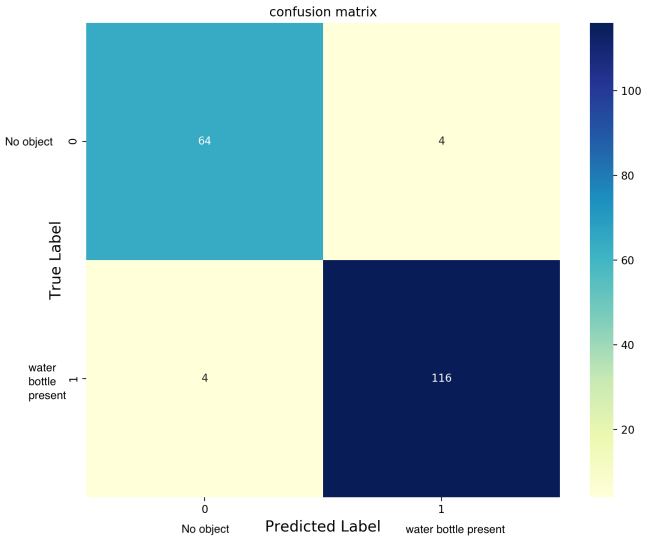

Fig. 15. Confusion matrix for predicting whether there is a water bottle in front of the transmitter (plotted for the test set)

to tracking based on visual data. The problem of manually designing a universal mapping from the RF reflections to other objects is an intractable task as such a mapping has to take care of reflection properties, constraints on movement, other reflecting objects and the surrounding environment. Neural Networks have provided immense success in learning complex mappings from training data. In this paper we reviewed manual methods of creating such a mapping, and methods that used ANNs. It is evident that the use of ANNs is not only beneficial in existing signal processing tasks, but also in solving more complex problems using signal data. We also experiment using low frequency RF-waves, and although our ANNs are generally not able to perform well on low frequency RF reflections (due to no real fault of their own), we are able to achieve some success in detecting a reflecting object (water bottle) in a closed environment even from extremely noisy data. This further highlights the effectiveness of CNNs for learning patterns from raw data. In future we hope to experiment using high frequency waves and recurrent neural network models, specifically LSTMs, which have achieved tremendous success on time-varying data. We believe it is possible to enable efficient, domain-independent detection and monitoring of reflecting objects over large areas using RF-waves and ANNs, and that this ability can benefit the society greatly.

ACKNOWLEDGMENT

I would like to thank Professor Debayan Gupta, for giving us the opportunity to explore this area of research, and for his endless support and guidance throughout.

REFERENCES

[1] J. D. Nichols, L. L. Bailey, N. W. Talancy, E. H. Campbell Grant, A. T. Gilbert, E. M. Annand, T. P. Husband, J. E. Hines et al., "Multi-scale occupancy estimation and modelling using multiple detection methods," Journal of Applied Ecology, vol. 45, no. 5, pp. 1321–1329, 2008.

[2] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008, pp. 1–4.

[3] CHETTY, K., SMITH, G., AND WOODBRIDGE, K. 2012. Through-the-wall sensing of personnel using passive bistatic wifi radar at standoff distances. IEEE Trans. Geoscience and Remote Sensing.

[4] COOPER, K. B., DENGLER, R. J., LLOMBART, N., BRYLLERT, T., CHATTOPADHYAY, G., SCHLECHT, E., GILL, J., LEE, C., SKALARE, A., MEHDI, I., ET AL. 2008. Penetrating 3-d imaging at 4-and 25-m range using a submillimeter-wave radar. Mi-crowave Theory and Techniques, IEEE Transactions on.

[5] Matthew Keally, Gang Zhou, Guoliang Xing, Jianxin Wu, and Andrew Pyles. Pbn: towards practical activity recog- nition using smartphone-based body sensor networks. In Proceedings of the 9th ACM Conference on Embedded Net- worked Sensor Systems, pages 246–259. ACM, 2011.

[6] J. Weppner and P. Lukowicz, "Bluetooth based collaborative crowd density estimation with mobile phones," in Pervasive computing and communications (PerCom), 2013 IEEE international conference on. IEEE, 2013, pp. 193–200.

[7] M. Wirz, T. Franke, D. Roggen, E. Mitleton-Kelly, P. Lukowicz, and G. Troster, "Probing crowd density through smartphones in city-scale mass gatherings," EPJ Data Science, vol. 2, no. 1, p. 1, 2013.

[8] J. Wilson and N. Patwari, "Radio tomographic imaging with wireless networks," IEEE Trans. Mobile Computing, 2009.

[9] F. Adib and D. Katabi, "See through walls with WiFi!", SIGCOMM, August 12–16, 2013.

[10] F. Adib, Zachary Kabelac and D. Katabi, "Multi-Person Localization via RF Body Reflections", In Usenix NSDI, 2015.

[11] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi and Fredo Durand, "Capturing the Human Figure Through a Wall".

[12] Isaac Cushman, Danda B. Rawat, Abhishek Bhimraj, Malik Fraser,"Experimental approach for seeing through walls using Wi-Fi enabled software defined radio technology", Digital Communications and Network, 2016.

[13] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak. Convolutional, long short-term memory, fully connected deep neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4580–4584, April 2015.

[14] TaraN Sainath,Ron J Weiss, Andrew Senior,Kevin Wilson and Oriol Vinyals. "Learning the speech front-end with raw waveform cldnns".

[15] Timothy J. O'Shea and Johnathan Corgan. Convolutional radio modulation recognition networks. CoRR, abs/1602.04105, 2016

[16] Tim O'Shea, Tamoghna Roy and T. Charles Clancy, "Over the Air Deep Learning Based Radio Signal Classification", DeepSig, 2017.

[17] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S. Jaakkola and Matt T. Bianchi, "Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture", Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[18] A. F. Pérez-Zapata, A. F. Cardona-Escobar, J. A. Jaramillo-Garzón 'and Gloria M. Díaz, "Deep Convolutional Neural Networks and Power Spectral Density Features for Motor Imagery Classification of EEG Signals", Springer International Publishing AG, part of Springer Nature 2018.

[19] Mingmin Zhao, Yonglong Tian, Hang Zhao, Moham- mad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba, "Rf-based 3d skeletons", In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication, pages 267–281. ACM, 2018.

[20] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, Dina Katabi, "Making the Invisible Visible: Action Recognition Through Walls and Occlusions", In Proceedings of the 2019 Conference of the ACM Special Interest Group on Data Communication, pages 267–281. ACM,20 Sep 2019

[21] Saandeep Depatla and Yasamin Mostofi, "Crowd Counting Through Walls Using WiFi", 2018.

[22] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon and Shahrokh Valaee, "A Survey on Behaviour Recognition Using WiFi Channel State Information", IEEE COMMUNICATION MAGAZINE, 2017.

[23] Zhihao Zhang, Chongrong Fang, Yuanchao Shu, Zhiguo Shi† and Jiming Chen, "Demo Abstract: FindIt - Real-time Through-Wall Human Motion Detection Using Narrow Band SDR", SenSys '16 November 14-16, 2016.

V. Appendix

A. Network Architecture

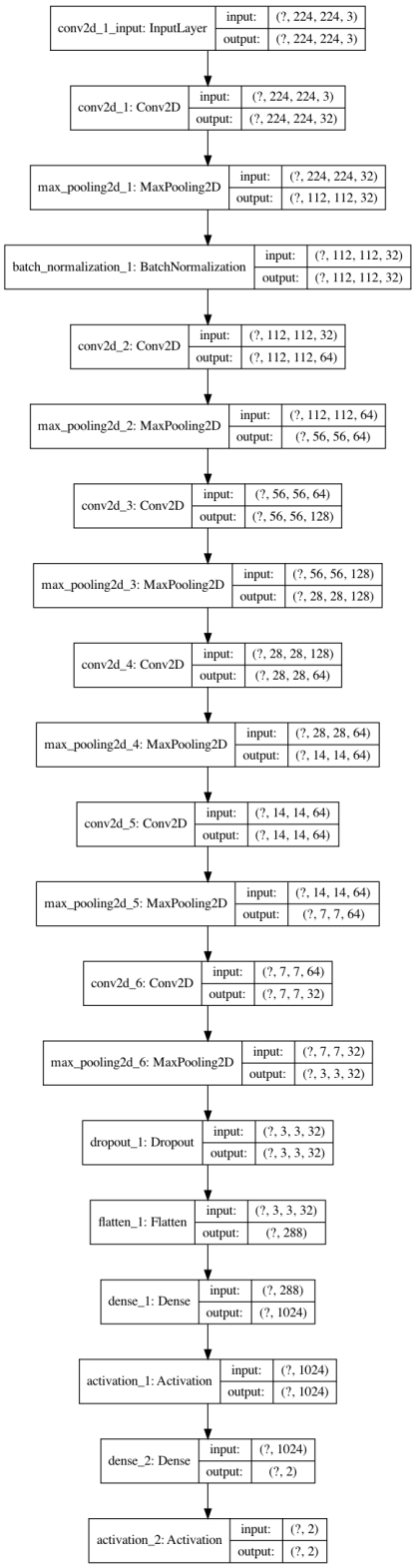

Fig. 16. CNN2 Architecture based on VGGNet (? denotes the iteration of the input fed in)

B. *Feature Maps*

We visualise the feature maps of the intermediate convolutional layers to try to visually interpret what the network is learning about the heatmaps. One can compare feature maps with the input data being fed to the CNN as shown in figure 12. The first layer seems to be retaining the full shape of the spectrogram, although there are a few filters that are not activated and left blank. The activations retain almost all of the information present in the initial picture. As we go deeper in the layers, the activations become increasingly abstract and less visually perceptible. They begin to encode higher level concepts such as the specific frequency bands, their thickness and shade (shade corresponds to power value at that point in time). Higher level features carry less information about the visual content of the image and more information relevant to the specific class of the image. The final layer seems to learn even more complex features about the specific types of edges and blurriness in the shade of the image, although it would be too far fetched to claim an understanding of the activations in such a deep layer. It can be said however that the network is able to learn the features it needs to analyse to successfully distinguish between the presence and absence of a water bottle in the spectrogram.

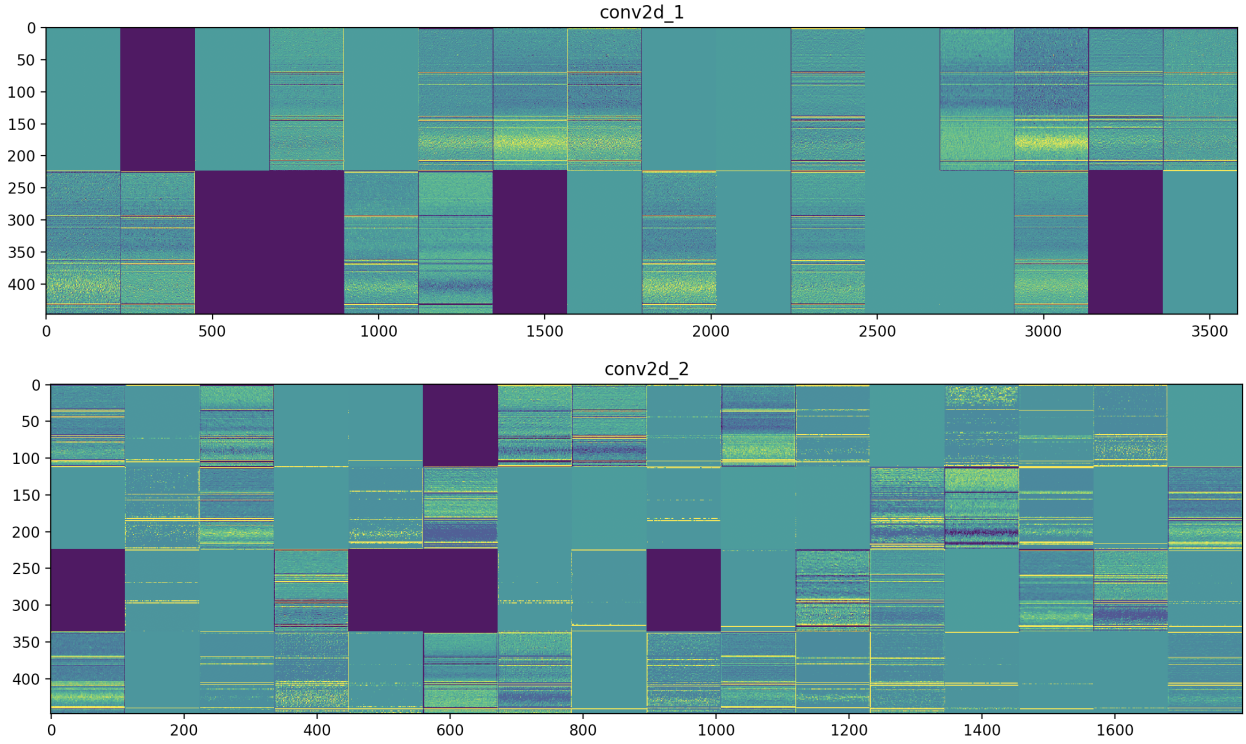

Fig. 17. Feature maps of each convolutional layer

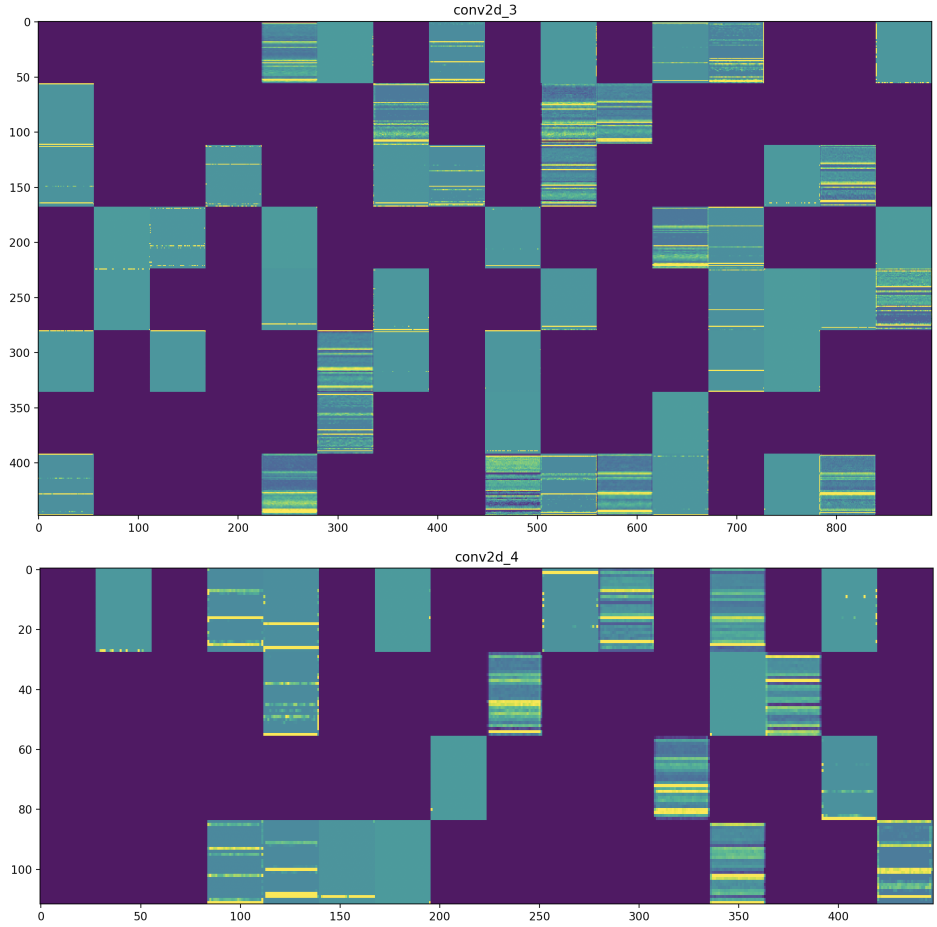

Fig. 18. Feature maps of each convolutional layer

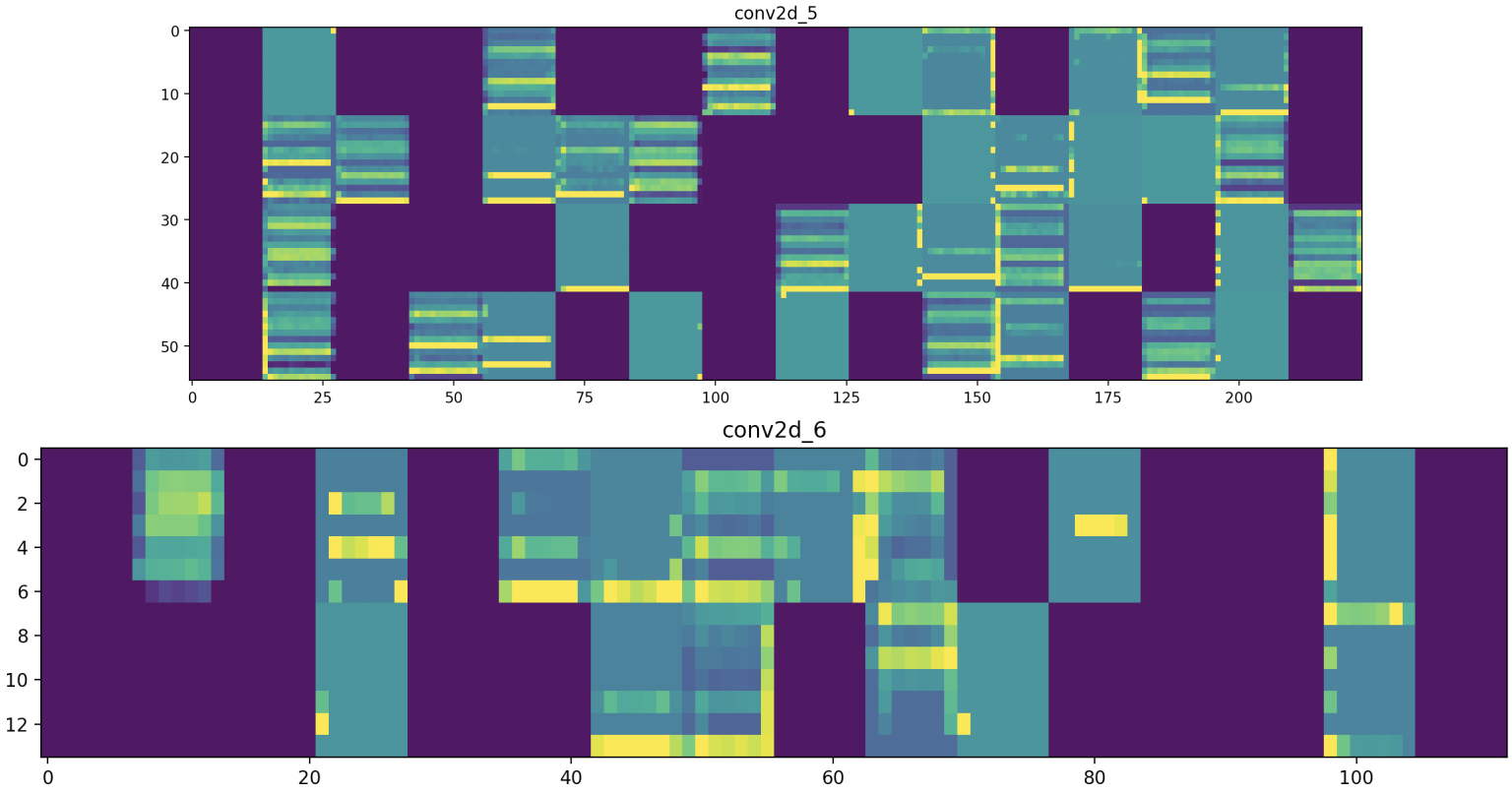

Fig. 19. Feature maps of each convolutional layer