Understanding Deep Neural Networks Performance for Radar-based Human Motion Recognition

Moeness G. Amin and Baris Erol
Center for Advanced Communications, Villanova University, USA (moeness.amin,berol@villanova.edu)

Abstract—**Deep neural networks have recently emerged as a promising tool for radar-based human motion recognition. Their nonlinear structure makes them successful in classifying large-scale datasets. However, due to their complexity, it is difficult to interpret the classification results and identify pixels with the biggest impact on the classification score. In this paper, we investigate recently proposed linear-wise relevance propagation (LRP) method which finds relevant pixels within the image. Based on this method, it is possible to recognize pixels which contain evidence for or against the prediction made by a classifier. Experimental results demonstrate that the LRP method can be successfully applied to detect regions within the radar images responsible for distinguishing human motions.**

Keywords—Deep learning, human motion recognition, radar, time-frequency domain

I. Introduction

Deep learning has attracted widespread interest in various pattern recognition applications due to its superiority compared to traditional methods [1]-[3]. Since image classification still remains as a common issue in various tasks, new applications of deep learning are still emerging. Recently, deep learning methods emerged as an effective tool in human motion recognition (HMR) using both continuous wave (CW) and range-Doppler radar systems [4]-[7].

Depending on the domain data representation, radar images of signal backscattering from moving humans typically depict target radar-cross-section (RCS), range, velocity information and changes in this information over time [8]-[11]. Using radar for classifications of human daily activity finds applications in urban security, health care and assisted living, and medical diagnosis [12]-[15]. The key part in traditional classification of radar images is the feature engineering process, which relies on the knowledge of system operator. In the case of HMR, physical characteristics of motion kinematics are considered important in defining features in the signal joint-variable representations of slow-time, fast-time, and frequency variables. However, this approach suffers from lack of generality and sufficiency. First, it is difficult to define one set of features relevant to all human motions. Second, the rich scatterings of electromagnetic waves from human during daily activities cannot be simply parametrized, accurately modeled, or just described by few predefined features. Deep learning offers automation in feature extraction process and, thus, it is more suitable for the general problem of HMR. In essence, parts of motion signatures which could be considered minor details and readily ignored by manual feature selections can be cast as important and ”relevant” features by deep learning.

Deep learning methods use neural networks with multiple layers, i.e., deep neural networks (DNNs), to automatically learn from the data. Multiple layers combined with nonlinear structure make DNNs successful in capturing intricate data properties. However, due to the absence of theoretical analysis, DNNs are often viewed as a black box. Much study in recent years have been focused on understanding the classification results and visualizing the regions of image which contributed the most and the least to a certain decision. In other words, seek an understanding of which pixels or regions within the image have the biggest and the smallest impacts on the classification score.

There are several methods that attempt to visualize what the network learns [16]-[18]. Most of these methods use some type of backward mapping function to generate visualizations in the form of a heatmap [17], [19]. Heatmap assigns each pixel a relevance score. The linear-wise relevance propagation (LRP) method was recently proposed to address the issue of pixel relevance [19]. This method offers several advantages over other approaches for computing heatmaps, namely,

- it can be used to analyze most of the deep learning architectures,
- it provides direct relationship between the network output (classification score) and the heatmap,
- it provides both the positive and negative evidences.

The last property of LRP allows us to identify pixels with positive evidence, supporting the classification decision, and pixels with negative evidence, working against correct prediction. Although, recent approach of using DNN to classify human motions showed improved classification rates, none has posed the question of what the neural network actually learns from the radar backscattering data when represented in joint-variable domains as images.

In this paper, the LRP method is employed to provide the relative significance of the different time-frequency regions of the spectrograms. These regions manifest different self-and cross-motion articulations as the activity begins, progresses in time, and ends. In addition of spectrograms, we also examine the significance of different regions of the target range maps, which show range translation signature over slow-time. We consider four classes of human motions, namely, walking, sitting, falling, and bending. It is shown that the LRP method can successfully identify regions within the image which have the highest impact on the classification rate. These regions, established by machine, correspond to areas which would, in most cases, be visually recognized by a human operator as relevant.

The paper is organized as follows. Section II describes the deep learning approach for HMR. Section III focuses on the LRP approach and on generating heatmaps based on the radar images of human motions. Experimental results which demonstrate the impact of specific image regions on classifications, are shown in Section IV. The conclusion is given in Section V.

II. Deep learning for HMR

In this section, we first describe the radar images of human motions which are used to train the DNN. Next, the DNN used for HMR is presented.

A. Radar images used in HMR

A wide-band radar, such as Frequency Modulated Continuous Wave (FMCW), provides the means to analyze target returns in different joint-variable domains. Different domains profess different suitabilities for different motions, and not a single domain has been proven to discriminate best among all motions [20]. Each 2D joint-variable domain representation can be converted to a gray-scale image. In this paper, we observe the images obtained from the time-frequency (TF) domain as well as the range map.

The TF domain has been employed to depict the changes in velocity of the human body parts over time. Typically, the spectrogram is used as the TF signal representation,

$$SPEC(n,k) = \left| \sum_{m=0}^{N-1} \mathrm{h}(m)\mathrm{s}(n-m)e^{-j2\pi km/N} \right|^2, \quad (1)$$

where $s(n), n = 0,..,N-1$ is the signal and $h(m)$ is a window function. We deal with the radar signal as non-stationary but deterministic, in lieu of a random process [21], [22]. There are quadratic time-frequency distributions (QTFDs), other than spectrogram, which can offer higher resolution and power concentration in time and frequency. Since high resolution associate with high fidelity and is likely to depict more details of the human motion signature in the TF domain, it is expected that DNN would benefit from applying high resolution kernels underlying QTFD [23], [24]. In this work, however, we confine learning to the commonly used spectrograms. Within the spectrogram paradigm, we also recognize that the window length, shape and shifts can readily affect the motion signature appearance and representation, and thus possibly alter the respective heatmap.

B. DNN for motion recognition

In this paper, we use a deep learning architecture for motion classification proposed in [4] as depicted in Figure 1. In the general scheme, this method follows three steps, namely, pre-processing, feature extraction and classification. The preprocessed representation is used as input to stacked auto-encoders that perform feature extraction.

Sparse autoencoder is a neural network that attempts to obtain the sparse representation of the input data. The learning is achieved via a single hidden layer that typically has fewer neurons than the input and output layers. Connections between layers are established by the weights and biases. Each hidden neuron applies a sigmoid function $\sigma\{\bullet\}$ to the weighted and biased input data units x_m, i.e., the output of the hidden layer nth neuron is

$$a_n = \sigma\bigg(\sum_m x_m w_{m,n} + b_n \bigg), \quad (2)$$

where $w_{m,n}$ and b_n denote the weight and the bias term, respectively. The values of output layer units are obtained in a similar way, by applying the nonlinear function $\sigma\{\bullet\}$ to weighed and biased hidden layer units a_n. The weights and biases of neurons are learned in such manner which minimizes the reconstruction error and promotes sparsity. Once the features are extracted using stacked autoencoders, the classification is performed using softmax regression classifier. More details about the architecture can be found in [4].

III. Relevant pixels according to the DNN

In this section, we describe the LRP approach. This approach is based on the layer-wise conservation principle which states that the relevance R should be preserved when back-propagating from one layer to the next, i.e.,

$$R^{(1)} = ... = R^{(l)} = R^{(l+1)} = ... = f(x). \quad (3)$$

The relevance of the last layer is equal to the classification function $f(x)$ of the input image x. Relevance in each layer is defined as the sum of relevances of all N neurons,

$$R^{(l)} = \sum_{n=1}^{N} r_n^{(l)}. \quad (4)$$

In order to obtain the relevance in the first layer, i.e., the heatmap $h_n = r_n^{(1)}$, the output signal is propagated backward by the following rule:

$$r_n^{(l)} = \sum_m \left(\alpha \frac{z_{n,m}^+}{\sum_{n'} z_{n',m}^+} + \beta \frac{z_{n,m}^-}{\sum_{n'} z_{n',m}^-} \right) r_m^{(l+1)}. \quad (5)$$

We use the values of the parameters α and β as set in [19], i.e., $\alpha = 2$ and $\beta = -1$. $z_{n,m}^+$ and $z_{n,m}^-$ are the positive and the negative part of $z_{n,m}$, respectively. $z_{n,m}$ is the contribution of nth neuron at layer l to the activation of the mth neuron in the next layer $l+1$,

$$z_{n,m} = a_n^{(l)} w_{n,m}^{(l,l+1)}, \quad (6)$$

and a_n is the activation of neuron n defined by eq. 2.

IV. Discussion and analysis of the results

The FMCW radar experiments were conducted in the Radar Imaging Lab, at the Center for Advanced Communications, Villanova University. The radar system used in the experiments, named SDRKIT 2500B, is developed by Ancortek, Inc. The center frequency is 25 GHz, whereas the bandwidth is 2 GHz which provides 0.075 m range resolution.

The dataset contained four human motions: falling, sitting, bending and walking. Each motion was observed during a time

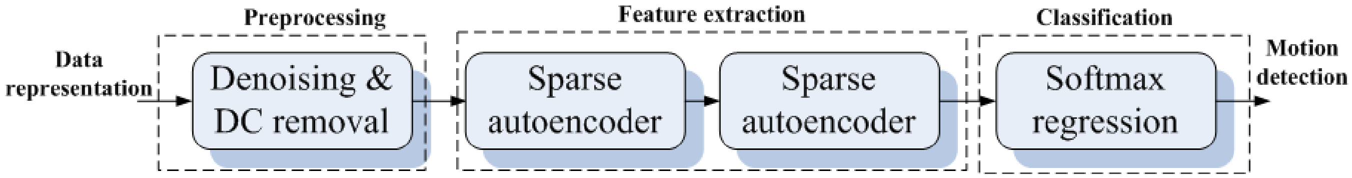


Fig. 1. Deep learning based architecture for motion classification.

span of 4 seconds. Resulting spectrograms and range maps were then converted to gray-scale images with a grid size of 64x64, and used as inputs to the DNN. The dataset contains 408 signals: 117 fall, 111 sit, 115 bend and 65 walk signals. The dataset is divided into two sets in order to train and test the DNN. Training set consisted of 200 samples, where each motion class was represented by 50 samples. The rest of the dataset was used for testing. The number of units in the hidden layer for the first auto-encoder was set to 300, meaning that the network would attempt to compress 4096 coefficients into 300. The 300 outputs were further compressed using only 100 units in the second hidden layer.

Figure 2-(a) depicts spectrograms of the four observed motions. Based on visual information of these images, we can discern some prominent features for each class. For example, fall signature depicts sudden drop in frequency in the shape of a negative ”hump”. Sitting signature also exhibits sudden drop in frequency, but the drop is not as pronounced as in the case of a fall. Picking up an object and standing up, i.e., bending, exhibits both positive and negative frequency components, while the main characteristic of the walking signature are the periodic components.

The heatmaps corresponding to these spectrograms are shown in Figure 2-(b). The heatmap values are normalized to be in the range $[-1, 1]$ and each value is the relevance that can be positive or negative. Namely, red and yellow shades denote positive evidence (evidence that support prediction), while blue color denotes negative evidence (evidence that goes against prediction). Areas with positive evidence are extracted from heatmaps and depicted in Figure 2-(c). For example, it can be noticed that relevant pixels for fall signature capture mostly the area where maximum frequency and negative hump shape are present. Similarly, in the case of walk, periodic components are denoted as relevant. We also show range maps with corresponding heatmaps in Figure 3. These results demonstrate that DNN makes conclusions that are consistent with those made by humans when deciding which image regions favor certain decision.

It is also possible to view heatmaps as filters which pass only the relevant pixels and suppress noise and artifacts that can cause misclassification. In order to verify the filtering aspect of heatmaps, we applied masks (shown in Figure 2-(c)) to the testing samples. The initial classification rate using DNN approach depicted in Figure 1 was 86.7% (Table I). Once the filtering is employed, the success rate is increased to 89.2% with bend and walk professing 100% (Table II). Even though this denoising approach requires that the heatmaps to be employed on the samples with known class labels, the results show promise in the use of heatmaps for improving success rates. We observed that when denoising is applied to training data only where test data remained unmasked, the improvement in classification rates is slightly lower than that when both training and test data are masked according the corresponding class heatmaps.

TABLE I. Confusion matrix for the deep learning based approach for motion classification using spectrograms. Success rate is 86.7%.

Classified/Actual Class	Fall	Sit	Bend	Walk
Fall	74%	3%	-	3%
Sit	13%	84%	3%	-
Bend	3%	10%	97%	3%
Walk	10%	3%	-	94%

TABLE II. Confusion matrix for the deep learning based approach for motion classification using spectrograms and the heatmaps. Success rate is 89.2%.

Classified/Actual Class	Fall	Sit	Bend	Walk
Fall	70%	-	-	-
Sit	-	87%	-	-
Bend	-	13%	100%	-
Walk	30%	-	-	100%

V. Conclusion

This paper represented the first study of characterizing what a Neural Network (NN) learns from the human motion micro-Doppler signatures to render desirable classification rates. We applied heatmap based method to determine what is relevant in motion signatures according to the deep neural networks (DNN). Results demonstrate that this method can successfully determine pixels which have the highest impact on the classification score. This impact can be either positive or negative, i.e., some pixels help in classification while the others are responsible for misclassification. It is shown that DNN tends to make similar assessments as a human operator when determining which image regions are relevant for the classification.

Acknowledgment

The authors would like to thank Ms. Branka Jokanovic for her input and help in MATLAB programs.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

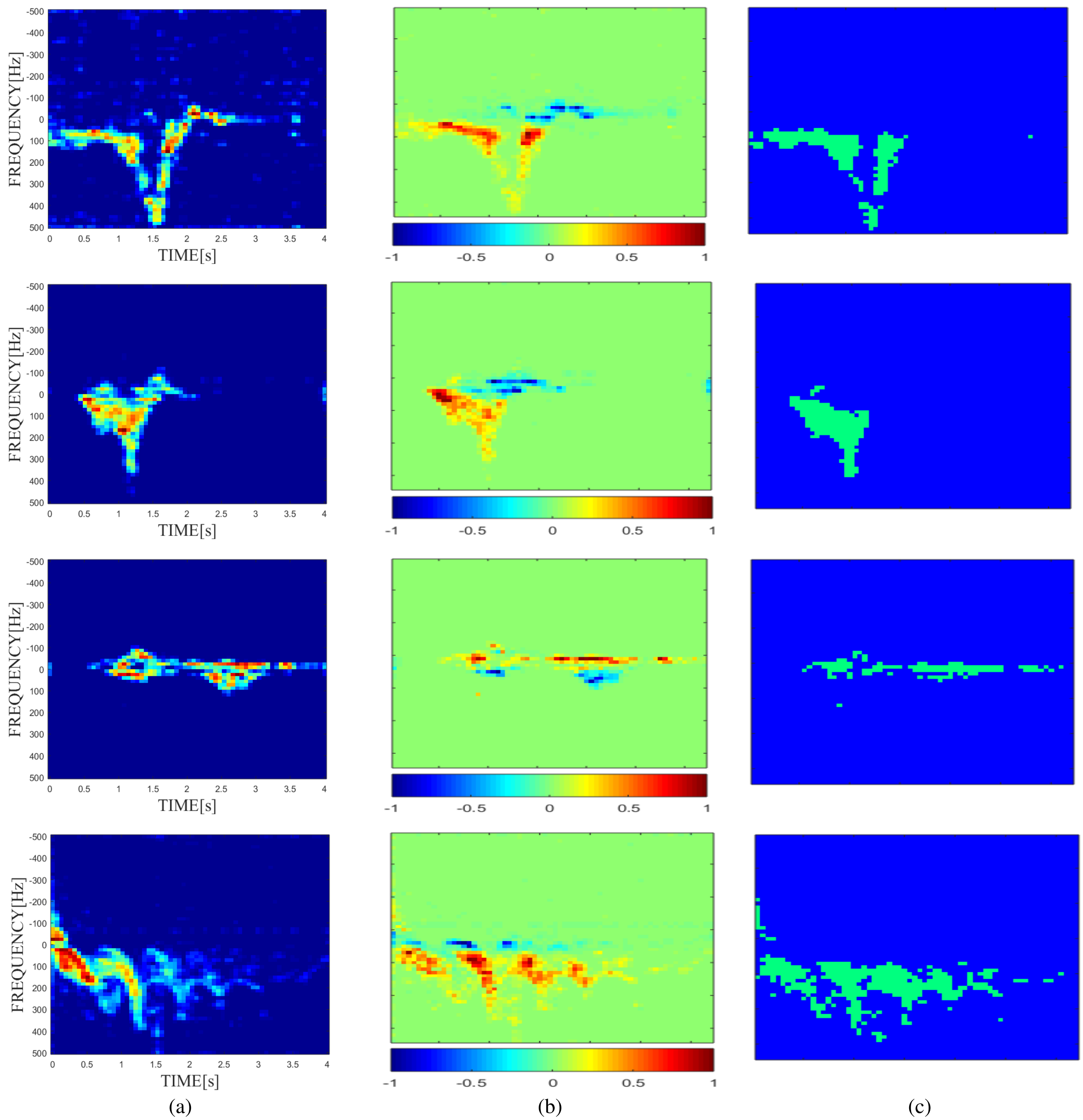
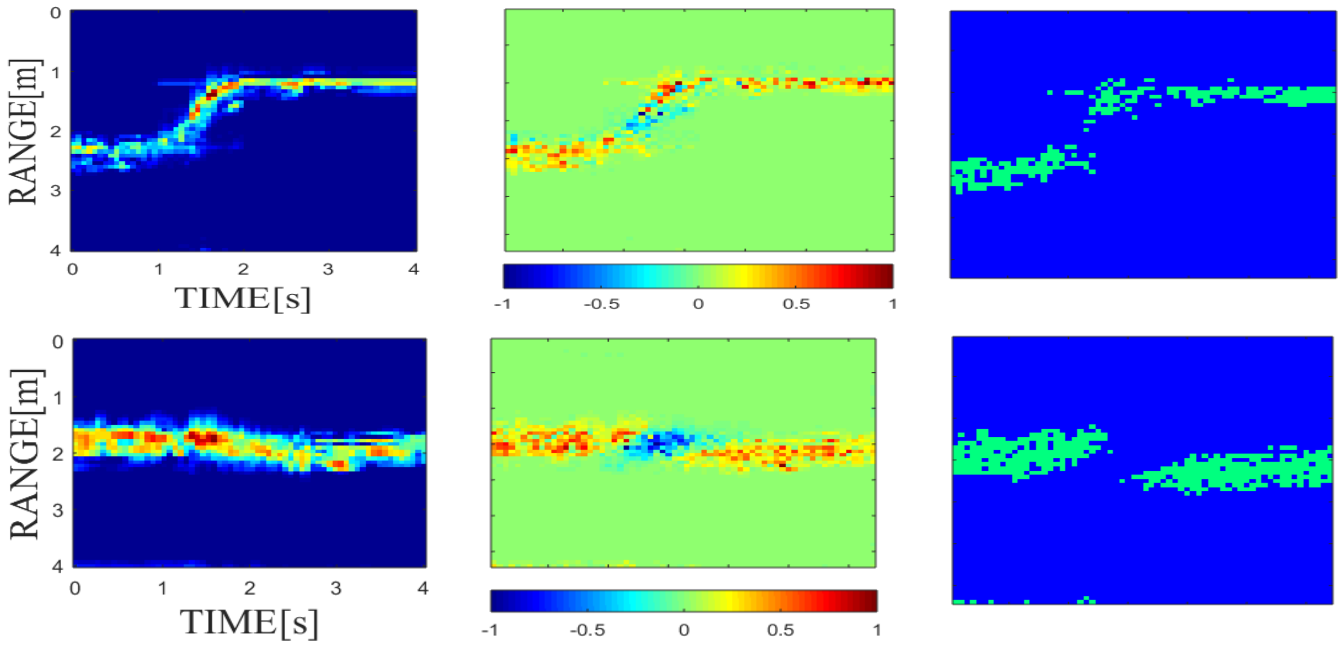


Fig. 2. (a) Spectrogram. (b) Corresponding heatmap. (c) Mask which contains pixels with positive evidence. First row: fall, second row: sit, third row: bend, forth row: walk.



(a) (b) (c)

Fig. 3. (a) Range map. (b) Corresponding heatmap. (c) Mask which contains pixels with positive evidence. First row: fall, second row: sit.

[2] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2012, pp. 3642–3649.

[3] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna- and rna-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[4] B. Jokanovic, M. Amin, and F. Ahmad, "Radar fall motion detection using deep learning," in *2016 IEEE Radar Conference (RadarConf)*, Philadelphia, PA, 2016, pp. 1–6.

[5] M. S. Seyfioğlu, S. Z. Gürbüz, A. M. Özbayoğlu, and M. Yüksel, "Deep learning of micro-doppler features for aided and unaided gait recognition," in *2017 IEEE Radar Conference (RadarConf)*, Seattle, WA, 2017, pp. 1125–1130.

[6] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, 2016.

[7] R. Trommel, R. Harmanny, L. Cifola, and J. Driessen, "Multi-target human gait classification using deep convolutional neural networks on micro-doppler spectrograms," in *2016 European Radar Conference (EuRAD)*, London, 2016, pp. 81–84.

[8] P. van Dorp and F. Groen, "Human walking estimation with radar," *IET Radar, Sonar and Navigation*, vol. 150, no. 5, pp. 356–365, 2003.

[9] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using a support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1328–1337, 2009.

[10] S. Z. Gurbuz, C. Clemente, A. Balleri, and J. J. Soraghan, "Micro-Doppler-based in-home aided and unaided walking recognition with multiple radar and sonar systems," *IET Radar, Sonar & Navigation*, vol. 11, no. 1, pp. 107–115, 2016.

[11] C. Li, J. Cummings, J. Lam, E. Graves, and W. Wu, "Radar remote monitoring of vital signs," *IEEE Microwave Magazine*, vol. 10, no. 1, pp. 47–56, 2009.

[12] M. G. Amin, Ed., *Radar for Indoor Monitoring*. CRC Press, 2017.

[13] A. K. Seifert, A. M. Zoubir, and M. G. Amin, "Radar-based human gait recognition in cane-assisted walks," in *2017 IEEE Radar Conference (RadarConf)*, Seattle, WA, 2017, pp. 1–6.

[14] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. C. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 71–80, 2016.

[15] B. Y. Su, K. C. Ho, M. J. Rantz, and M. Skubic, "Doppler radar fall activity detection using the wavelet transform," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 865–875, 2015.

[16] D. Erhan, A. Courville, and Y. Bengio, "Understanding representations learned in deep architectures," *Department dInformatique et Recherche Operationnelle, University of Montreal, QC, Canada, Tech. Rep*, 2010.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. ICLR*, 2014.

[19] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 11, pp. 2660–2673, 2017.

[20] B. Jokanovic and M. G. Amin, "Suitability of data representation domains in expressing human motion radar signals," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2370–2374, 2017.

[21] S. Hearon and M. G. Amin, "Minimum variance time-frequency distribution kernels," *IEEE Trans. on Signal Processing*, vol. 43, no. 1, pp. 1258–1262, 1995.

[22] W. Martin and P. Flandrin, "Wigner-ville spectral analysis of nonstationary processes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 6, 1985.

[23] M. G. Amin and W. Williams, "High spectral resolution time-frequency distribution kernels," *IEEE Transactions on Signal Processing*, vol. 46, pp. 2796–2804, 1998.

[24] B. Barkat and B. Boashash, "High resolution quadratic time-frequency distribution for multicomponent signals analysis," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2232–2239, 2001.