Sevgi Zubeyde Gurbuz and Moeness G. Amin

Radar-Based Human-Motion Recognition With Deep Learning

Promising applications for indoor monitoring

©ISTOCKPHOTO.COM/TALAJ

Deep learning (DL) has shown tremendous promise in radar applications that involve target classification and imaging. In the field of indoor monitoring, researchers have shown an interest in DL for classifying daily human activities, detecting falls, and monitoring gait abnormalities. Driving this interest are emerging applications related to smart and secure homes, assisted living, and medical diagnosis. The success of DL in providing an accurate real-time accounting of observed human-motion articulations fundamentally depends on the neural network structure, input data representation, and proper training. This article puts DL in the context of data-driven approaches for motion classification and compares its performance with other approaches employing handcrafted features. We discuss recent proposed enhancements of DL classification performance and report on important challenges and possible future research to realize its full potential.

Introduction

Radar has emerged as an important technology in such areas as commerce, defense, and security. Small, low-cost, solid-state and software-defined radar technologies have enabled new civilian radar applications in medical and automotive fields, advances in human–computer interaction, and the deployment of smart environments. The safety, reliability, portability, and affordability of radar devices have made them prime candidates for use inside office buildings, homes, schools, and hospitals.

Radar possesses unique advantages that complement other sensors such as visual, infrared, acoustic, pressure, and wearable sensors. Radar sensors are noncontact devices that work under any lighting conditions, including darkness, and can even penetrate opaque objects such as tables or walls. Another attribute that has made radar especially attractive for indoor monitoring is that it does not violate the privacy of monitored individuals. Radar backscattering signals can reveal human motion independent of clothing, making it suitable for such environments as hospitals, assisted-living facilities, restrooms, bedrooms, and other locations where people would be uncomfortable in the presence of video cameras. Radar-based remote

Digital Object Identifier 10.1109/MSP.2018.2890128
Date of publication: 26 June 2019

 1053-5888/19©2019IEEE

health-monitoring technologies can directly benefit the elderly, whose worldwide population over age 65 is projected to increase to 1 billion in 2030. The development of in-home motion classifiers using radar is an integral part of the "aging-in-place" paradigm that offers a sense of security and safety for elderly residents and their families. Although wearable devices are available to identify motion, especially falls, they are battery operated and depend either on accelerometers to detect motion or on the individual wearing the device to press a button. Radar offers a nonobstructive passive motion-sensing technology capable of notifying caregivers and first responders of critical events concerning the health and welfare of the observed individual. Moreover, sensing for smart environments and gesture recognition for device control have the potential to transform the way we live and improve quality of life globally by changing the way we interact with our surroundings.

In radar applications, fundamental differences in signal phenomenology have driven the development of unique approaches to DNN design for human-motion classification.

Radar innovations for indoor monitoring have been made possible in part because of advances in machine learning, parallelization, and the speed of graphics processing units. These advances have led to substantial improvements in the performance of deep neural networks (DNNs). In other fields, such as image and natural language processing, DNNs are now able to harvest extremely large amounts of training data to achieve record-breaking performance. This was demonstrated in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), where the winner in 2015 used 1.2 million images to train a 152-layer residual neural network (ResNet) for the classification of 1,000 classes [1]

In radar applications, fundamental differences in signal phenomenology have driven the development of unique approaches to DNN design for human-motion classification. Radar data are not inherently acquired as images. Rather, the received signal of a radar unit is a complex time series, whose amplitude and phase can be related to the electromagnetic scattering and kinematics of the target being observed. Several preprocessing stages, such as filtering and TF analysis, are typically performed to generate 1D, 2D, or 3D data for input. Two-dimensional radar data representations are not typical images. They exhibit different properties from optical imagery, as each pixel value is now not just an intensity but a sample in time and frequency. This has driven the design of novel approaches to DL specific to radar. Recent work on radar-based classification has espoused a knowledge-aided signal processing approach that integrates the fundamentals of physics, phenomenology, and modeling into DNN design and training to develop novel solutions tailored for sensing applications in the radio-frequency (RF) domain.

Radar-based motion characterization relies on observations of the fluctuation in micro-Doppler frequencies induced in the received signal due to vibrations or rotations of parts of the body in addition to translational motion components. The unique kinematics of the bipedal human gait make the respective radar micro-Doppler signature markedly different from those of other living animals that could be present in a home, such as dogs or cats. Moreover, even finer differences in kinematics, corresponding to different daily human activities, are reflected in the intricate patterns of the micro-Doppler signatures. Figure 1 illustrates the radar micro-Doppler signature for daily activities, such as walking and picking up an object. These particular classes possess quite distinct profiles that can also be easily detected by normal visual observation.

Machine learning can enable automatic motion recognition. Conventionally, a set of handcrafted features is first extracted from the micro-Doppler signature, such as bandwidth and stride rate. However, the efficacy of such features depends on operational and situational factors [2], such as the transmit and pulse-repetition frequencies, observation duration (dwell time), signal-to-noise ratio, and aspect angle between observed motion and radar line of sight.

We recognize the promise in radar-based research to provide accurate and nonintrusive tracking of daily human activities. Recent results in DL for the use of radar in indoor monitoring have reported performance superior to that of conventional machine-learning techniques, especially as the number and similarities between different classes increase. For example, in a seven-class problem of visually identifiable signatures, a support vector machine (SVM) classifier yielded roughly the same performance as that of a convolutional neural network (CNN) [3]. But, in a more difficult problem of classifying 12 aided/unaided human gaits [4], the CNN outperformed SVM by approximately 13%.

As part of a broad, in-depth survey of the current state of the art in radar-based motion recognition using DL, DNN performance was compared to conventional machine-learning techniques and

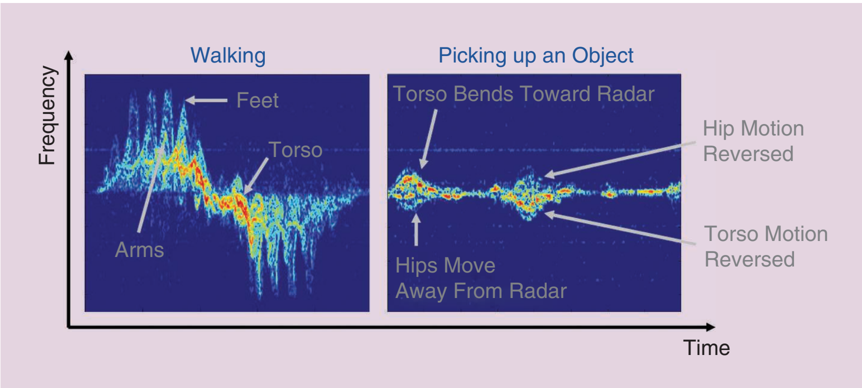

FIGURE 1. Two images showing sample 24-GHz FMCW radar micro-Doppler signatures for daily human activities.

other data-driven approaches. Specifically, the following radar-specific questions were considered:

- *Input data representation*: What is the best way to process the raw, complex data stream for presentation to the input of DNNs?
- *DNN training*: How can knowledge-aided approaches exploiting physics and modeling of human kinematics be developed to mitigate challenges of low training-sample support?
- *Pervasive sequential classification*: How should novel architectures be designed to address optimal classification across multiple input representations to achieve high accuracy in recognizing the dynamic, free-form flow of motion and activity characteristic of daily life?

We conclude the article with a delineation of ongoing challenges and future directions for DNN design in radar-based human-motion recognition.

We recognize the promise in radar-based research to provide accurate and nonintrusive tracking of daily human activities.

Indoor monitoring

Research into the use of radar in civilian applications [5] began almost 20 years ago with work on life-sign detection. Today, applications run the gamut from the monitoring of daily activities in smart environments to home security, assisted living, remote health monitoring, and human–computer interfaces. Figure 2 depicts various applications along with their respective human-motion classes. Radar in assisted-living and remote health applications [6] has been used for the noncontact measurement of heart rate and respiration and for detecting conditions related to heart rate and respiration, such as sleep apnea or sudden infant death syndrome. The detection of falls is also an important application [7] as falling remains a leading cause of mortality and major injuries among the elderly. Rapid response after a fall is critical to minimizing long-term debilitation and maintaining the independence and quality of life of senior citizens [6]. As such, fall detection has become a prime area of research in health monitoring and in the development of sensing technologies for telemedicine and smart homes. The desire to reduce health-care costs, while also expanding medical services to hard-to-reach rural areas, has driven research into the use of radar-based gait monitoring for medical gait analysis. Related applications could include, for example, detecting gait abnormalities [8] for fall-risk prediction, assessing recovery from injuries, measuring progression of neuromuscular disorders and response to treatments, monitoring health after a stroke, evaluating balance, and characterizing pathological gait for physical therapy, disability, and rehabilitation.

Broader applications of human-motion detection and recognition are also being pursued for security and energy-efficient smart-home applications. For example, occupancy (or, conversely, vacancy) sensing can be used for the intelligent operation of systems for the home, such as lighting systems or heating, ventilation, and air conditioning units. Beyond simple motion detectors triggered by fine motion, such as turning the page of a book, radar Doppler sensors can detect and classify larger-scale activity as well as vital signs. The unique patterns in

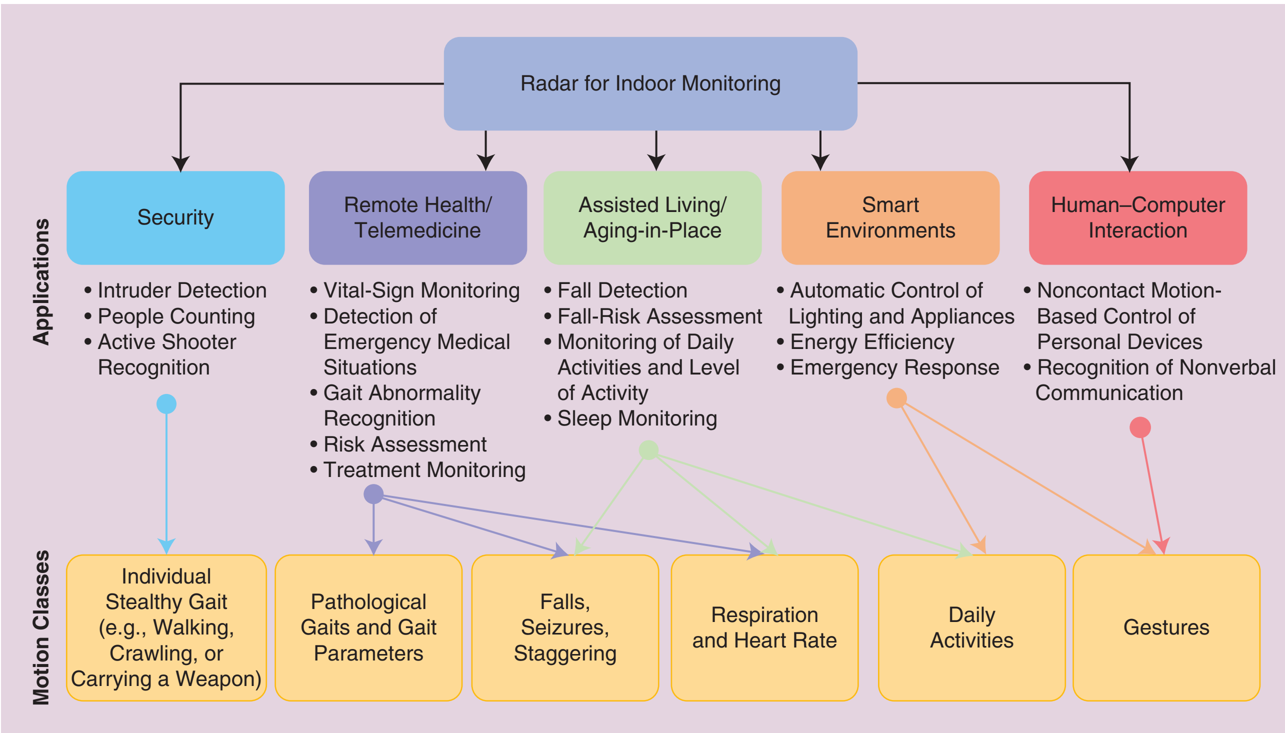

FIGURE 2. An overview of applications of radar-based motion sensing for indoor monitoring.

the micro-Doppler signature of individuals can also be exploited to identify specific people [9]. Within the past two to three years, DNNs have been used to enable radar micro-Doppler signature classification sensors to count the number of people in a room [10], recognize individuals within groups of four to 20 people [11], and identify intruders from among known signatures of household members [12]. In fact, even the way people carry a weapon, such as a rifle, has been shown to reflect unique characteristics in the micro-Doppler signature, thus enabling applications for active-shooter recognition [13] and armed/unarmed-personnel recognition [14].

The increased capability of radar-signal processing and machine-learning algorithms to recognize human motion has spurred parallel advancements in radar hardware. Research in radar-based human-motion classification has explored the use of a wide range of radar waveforms for this purpose at transmit frequencies of 2.4, 5.8, 24, 60, 77, and even 94 GHz, including continuous wave (CW), frequency-modulated CW (FMCW), and pulsed Doppler radar; ultrawideband (UWB), ultrashort pulse, interferometric, and multistatic radar; dual-polarized radar; and even Wi-Fi and through-the-wall radar. Dual-use systems, which can achieve both occupancy sensing and vital-sign monitoring, have also been proposed using UWB impulse radars and hybrid FMCW–interferometric radars. Miniaturization of FMCW radars at 60 GHz, such as the radar technology pioneered by the Google Soli sensor [15], has driven radar-based gesture recognition [16]–[19] for the noncontact control of personal electronics, wearable devices, and vehicle consoles. Examples of motion-based commands include pinching, sliding, and rubbing the fingers; swiping, pushing, pulling, and tilting the hand; and drawing a virtual circle in the air. Automotive applications for driver safety include the multifunction integration of vital-sign monitoring with gross-movement recognition, such as the detection of a cell phone being used or of eye-blinking patterns that indicate drowsiness [20].

The desire to reduce health-care costs, while also expanding medical services to hard-to-reach rural areas, has driven research into the use of radar-based gait monitoring for medical gait analysis.

Radar data domain representations

The signal received by radar is an inherently complex time series that is, in general, a time-delayed, frequency-shifted version of the transmitted signal. The scattering from the entire human body, $x[n]$, may, in turn, be approximated using superposition of the returns from K body parts. Thus, for an FMCW radar,

$$x[n] = \sum_{i=1}^{K} a_i \exp\left\{-j\frac{4\pi f_c}{c} R_{n,i}\right\}, \tag{1}$$

where $R_{n,i}$ is the range to the ith body part at time n, f_c is the transmit frequency, c is the speed of light, and the amplitude a_i is the square root of the power of the received signal as given by

$$a_i = \frac{G\lambda\sqrt{P_t \sigma_i}}{(4\pi)^{3/2} R_i^2 \sqrt{L_s}\sqrt{L_a}}. \tag{2}$$

Here, G is the antenna gain, λ and P_t are the wavelength and power of the transmitted signal, respectively; σ_i is the radar cross section of the ith body part; L_s and L_a are system and atmospheric losses, respectively.

Thus, the kinematic motion and electromagnetic scattering properties of the human body are reflected in both the amplitude and frequency modulations of the received signal. Most work in human-motion classification has relied on the application of a time–frequency (TF) transform to present salient target kinematics to the DNN. Subtle differences in human motion can be observed in the joint-variable signal representation of TF and time–scale. TF analysis methods can be classified into linear transforms and quadratic TF distributions (QTFDs). The former includes short-time Fourier transforms (STFTs), Gabor transforms, fractional Fourier transforms, and wavelet transforms. These transforms capture the signal's local behavior in some ways. QTFD, on the other hand, aims at concentrating the signal power along the instantaneous frequency of each signal component. In the case of radar backscattering from a moving target, such concentration accurately reveals the target velocity, acceleration, and higher-order terms. The Cohen's class of QTFDs of signal $x(t)$ is defined as follows [21]:

$$D(t,\omega) = \mathcal{F}_2[\Phi(\theta,\tau)A(\theta,\tau)] \\ \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi(\theta,\tau)A(\theta,\tau)\exp\{-j(\omega\tau+\theta t)\}\,d\tau d\theta, \tag{3}$$

where $A(\theta,\tau)$ is the ambiguity function, τ is the time lag, θ is the Doppler frequency shift, and $\mathcal{F}_2[\cdot]$ denotes a 2D Fourier transform. The kernel $\Phi(\theta,\tau)$ has typically low-pass filter characteristics to suppress signal cross terms and preserve auto terms, leading to reduced interference distributions [21].

The most common TF representation used in micro-Doppler analysis is the spectrogram, denoted by $S(t,\omega)$, which is the square modulus of the STFT, and a special case of (3). It can be expressed in terms of the employed window function, $w(t)$, as

$$S(t,\omega) = \left|\int_{-\infty}^{\infty} w(t-u)x(u)\,du\right|^2. \tag{4}$$

The target behavior in $S(t,\omega)$, referred to as the *micro-Doppler signature*, depicts how target Doppler frequencies vary with time and reflects unique patterns caused by the target motion.

Supervised classification

Handcrafted features

The recognition of human motion through the classification of micro-Doppler signatures typically involves first extracting a set of handcrafted features before inputting into a classifier. The process of computing the spectrogram (micro-Doppler signature) converts the complex time series of radar measurements into a 2D image. However, the radar data usually contain reflections from stationary objects in the room, known as *ground clutter*, which can obscure low-frequency components of the motion. A number of preprocessing steps can be implemented to isolate the micro-Doppler frequency components corresponding to only the human motion of interest. Prior to the computation of the spectrogram, high-pass filtering or moving-target indication (MTI) filtering can be implemented for clutter suppression, and range gating may be used to mitigate noise and clutter by isolating those data samples corresponding to the person's location. Figure 3 gives an example of the effect of MTI filtering and range gating on the micro-Doppler signature. It has been reported [10], however, that filtering to remove clutter weakens DNN performance because DNNs are able to extract information from signal components not masked by clutter. Once an image has been formed, thresholding and other image processing methods may also be used, if necessary, to further reduce unwanted artifacts and noise in the data. In handcrafted feature extraction, these types of preprocessing operations typically increase classification accuracy. A wide variety of features have been proposed over the years [22]:

- *Physical features*: These are related to the physical characteristics of target motion, such as Doppler bandwidth, stride rate, and the maximum/minimum of the Doppler signature envelope.
- *Transform-based features*: These use the coefficients of transforms, such as the fast Fourier transform and discrete cosine transform (DCT).
- *Speech-inspired features*: These features, originally designed for use in speech processing, include mel-frequency cepstral coefficients and linear predictive coding (LPC) coefficients.

Features can be extracted not only from the micro-Doppler signature but also from representations derived from the spectrogram, such as the cadence velocity diagram (CVD), which is obtained by performing a Fourier transform along each frequency bin of the spectrogram. The CVD measures how often the different velocities repeat (i.e., cadence frequencies). Hence, it may be used to extract physical features, such as velocity and stride rate, or transform-based features, such as pseudo-Zernike moments.

Due to the curse of dimensionality, a subset of features is often selected using wrapper methods that iteratively search for the optimal subset or filter methods that use a metric, such as mutual information or distance. The selected subset of features is extracted from each sample within a labeled set of training data and fed to a machine-learning algorithm, which attempts to find an optimal mapping of a feature space to class. The efficacy of the resulting model is evaluated through the application of an independent test data set to evaluate metrics, such as classification accuracy, specificity, and precision.

Data-driven feature learning

In contrast to handcrafted feature extraction, data-driven feature-learning methods can modify or adapt the process of extracting features by exploiting the knowledge gained from the analysis of the training data set. Figure 4 shows a flowchart comparing the processing involved with conventional classification of micro-Doppler signatures using handcrafted features and data-driven feature learning.

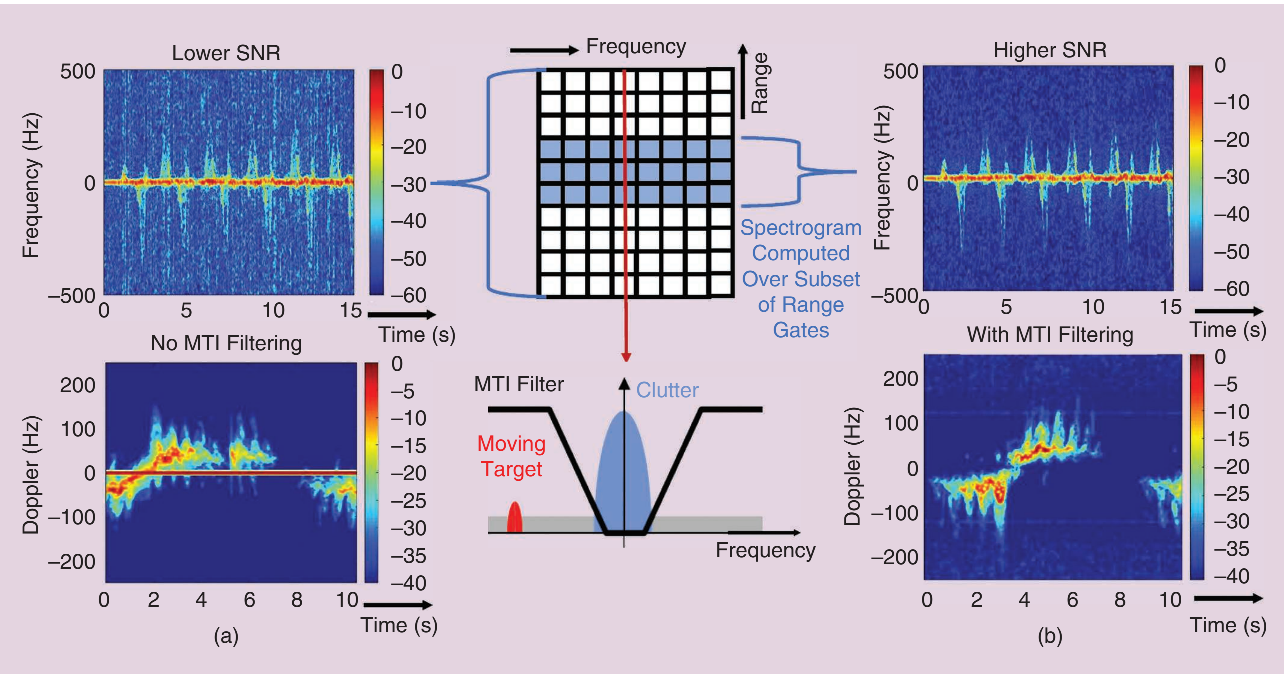

FIGURE 3. Four spectrograms and two illustrations showing the effect of clutter-suppression and noise-reduction techniques on spectrograms of radar data. Clutter-suppression and noise-reduction techniques were not applied to spectrograms in (a). Clutter-suppression and noise-reduction techniques were applied to spectrograms in (b). SNR: signal-to-noise ratio.

Principal component analysis

One unsupervised approach to data-driven feature learning is principle component analysis (PCA), a linear transformation of the data that finds the direction with the most variance. By also requiring that the basis vectors be mutually orthogonal, PCA ensures the process results in a representation with lower dimensionality. In conventional PCA, typically referred to as a *PCA* or *1D PCA*, the spectrogram depicting each micro-Doppler motion signature is vectorized and M corresponding vectors, x_i, of the same class are used to form an $N \times M$ input data matrix $X = (x_1 | x_2 | \ldots x_M)$. PCA computes the representation $z = W^{\top} x$, where the matrix W is found from the eigendecomposition of the unbiased data covariance matrix: $X^{\top} X = W \Lambda W^{\top}$. Extensions that preserve the 2D structure of the micro-Doppler signature are 2D PCA and generalized 2D PCA [23]. Whereas the former performs PCA on all row vectors of the spectrograms, the latter considers the correlation in both rows and columns and removes redundancies across both variables. Both 1D and 2D PCA are effective for feature extraction and are especially suited for embedded implementations that require computational efficiency [24].

Autoencoders

DNNs [25] can be interpreted as a nonlinear extension to PCA, which could be implemented using a one-layer artificial neural network with linear hidden and activation units. One special case of neural networks is the autoencoder (AE), which uses unsupervised learning to reconstruct the input at the output. In other words, for a given input vector x, the AE aims to approximate the identity operation $h_w(x) \approx x$. This is accomplished by using a symmetric encoder–decoder structure in the AE, as illustrated in Figure 5. First, the encoder computes a nonlinear mapping of the inputs as $e_i = \sigma(Wx_i + b)$, where σ denotes a nonlinear activation function, W denotes weights, and b denotes the biases of the encoder. The encoded features are then decoded to reconstruct the input vector x using $z_i = \sigma(\tilde{W} x_i + \tilde{b})$, where $\tilde{W}$ and $\tilde{b}$ are the weights and biases of the decoder, respectively. In unsupervised pretraining, the AE tries to minimize the reconstruction error between output and input:

$$\operatorname{argmin}_{\theta} J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (x_i - z_i)^2 + \beta \sum_{j=1}^{h} KL(p \| p_j). \qquad (5)$$

Here, θ is a parameter vector that includes the weights and biases of both the encoder and decoder: $\theta = [W, b, \tilde{W}, \tilde{b}]$; h denotes the number of neurons in the hidden layer; β denotes sparsity proportion; and $\sum_{j=1}^{h} KL(p \| p_j)$ denotes the Kullback–Leibler (KL) divergence between the Bernoulli random variables with mean p and p_j, respectively. The addition of the second term involving the sparsity parameter β functions as a regularizer and prevents the network from learning the identity function.

If the decoder is removed from the network, the remaining encoder components can be fine-tuned in a supervised manner by adding a softmax classifier after the encoder. An important advantage of AEs is that the unsupervised pretraining step

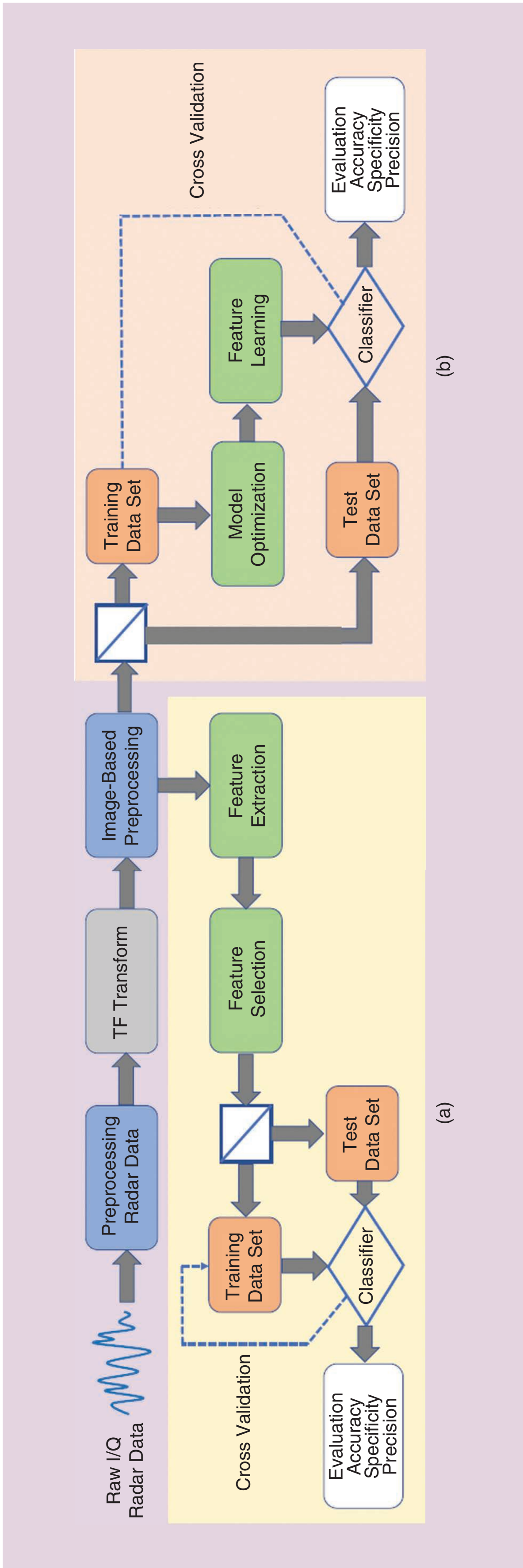

FIGURE 4. A flowchart showing (a) radar-signal classification with extraction of handcrafted features and (b) data-driven feature learning I/Q: in phase/quadrature.

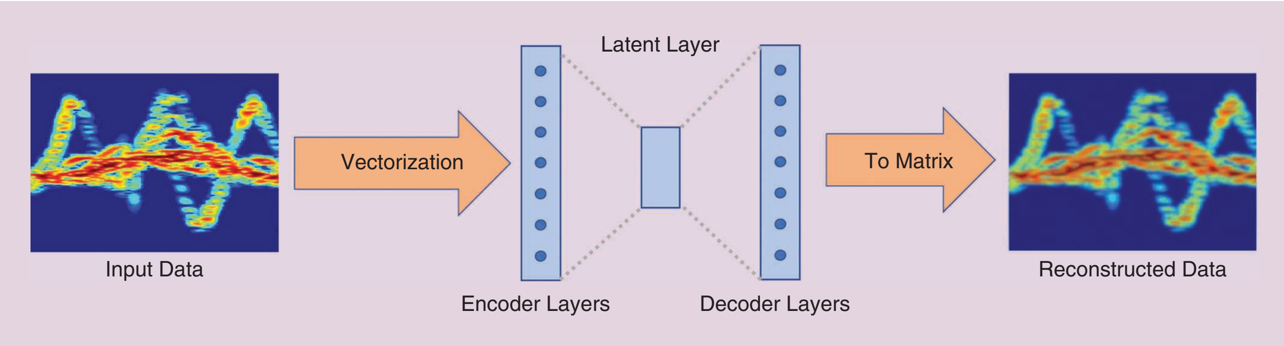

FIGURE 5. An illustration that shows how an AE works.

minimizes requirements for labeled training data sets. This means the AE is still effective even when only a small number of data is available. This advantage is particularly relevant for radar applications, where collecting a large number of measurements is time-consuming and costly, and where it is often not feasible to conduct experiments that span all expected scenarios and target profiles. For example, whereas the ILSVRC ImageNet database includes 1.5 million images, most work on classification of radar micro-Doppler signatures involve just 1,000–2,000 measured data samples.

CNNs

A highly popular alternative architecture is the CNN, which uses spatially localized convolutional filtering to capture the local features of input images. Basic features, such as lines, edges, and corners, are learned in the initial layers, while more abstract features are learned in deeper layers. CNNs typically comprise three types of layers: convolutional, pooling, and fully connected (see Figure 6). For a given matrix P, the mth neuron in the CNN calculates

$$M[i, j] = \sigma\left(\sum_{x=-2k-1}^{2k+1} \sum_{y=-2k-1}^{2k+1} f_m[x, y]P[i-x, j-y] + b\right). \quad (6)$$

The size of each side of the image is $2k+1$, M is the activation map of the given input P, f_m is the mth convolution filter, and σ is the nonlinear activation function. Generally, a max pooling layer follows each convolutional layer, in which local maxima are used to reduce computational complexity in the forward layers, and adds translation invariancy to the network. Finally, fully connected layers are used to learn nonlinear combinations of extracted features from previous layers.

Unlike AEs, CNNs rely entirely on a large amount of labeled training data for supervised optimization of network weights. The objective function of CNNs is highly nonconvex, with the result that the parameter space of the model contains many local minima. Conventionally, network weights are randomly initialized at the beginning of training. However, there is no guarantee that the local minimum to which optimization algorithms converge will also be optimum in the global sense. The use of a large training database increases the likelihood of convergence to a good, if not optimum, solution. In RF applications, where a large number of data may not be available, alternative approaches suitable to training under low sample support are required [e.g., convolutional AEs (CAEs) and transfer learning].

CAEs

CAEs combine the benefits of convolutional filtering in CNNs with the unsupervised pretraining of AEs. In CAEs, however, the encoder also contains convolutional layers, while the

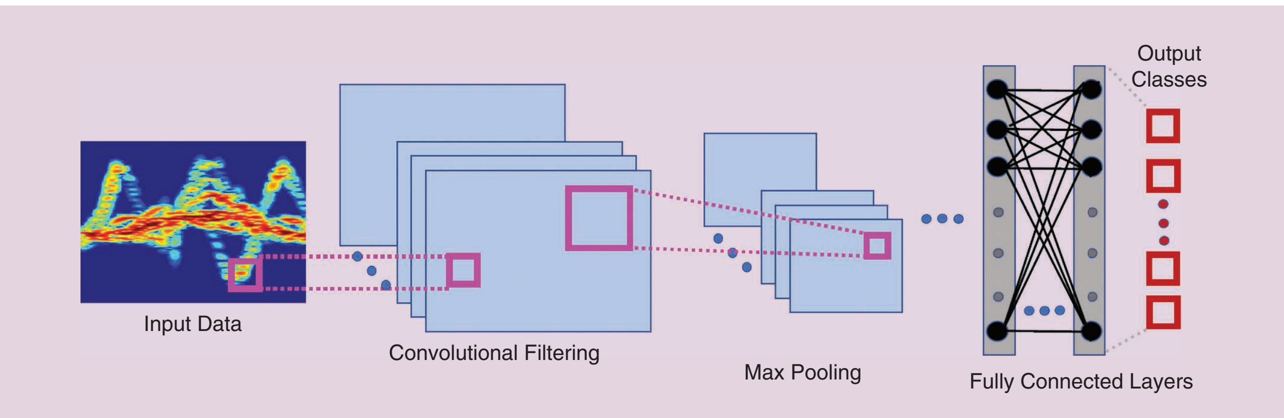

FIGURE 6. An illustration that shows how a CNN works.

decoder contains deconvolutional layers. The deconvolutional filters may be defined as transposed versions of the convolutional filters or they may be learned from scratch. Moreover, each deconvolutional layer must be followed by an unpooling layer, performed by storing the locations of the maximum values during pooling, thus preserving the values of these locations during unpooling and zeroing the rest. CAEs implement a two-stage training process, where unsupervised pretraining is used first to initialize network parameters, and then supervised fine-tuning with labeled measured data is used to optimize final values. Because the unsupervised pretraining starts the optimizer at values closer to the global optimum, better performance is typically achieved on small training data sets.

Transfer learning

Alternatively, knowledge gained from a different domain can be exploited to initialize the weights of a DNN for radar classification. As with CAEs, starting the supervised training process from a better set of initial weights reduces radar data requirements while also improving the classification accuracy. Pretrained models for VGGnet, AlexNet, GoogleNet, and ResNet have all been exploited for micro-Doppler classification. However, transfer learning from the optical domain has only outperformed CAEs when there are few real radar training data (e.g., fewer than 550 samples [26]).

DNN performance for daily human activities

Let us now consider a case study that used a common data set to evaluate these various approaches. A software-defined radio platform was programmed to transmit a CW signal at 4 GHz through two horn antennas of 48° beamwidth mounted side by side 1 m off the ground. Each of the 11 participants conducted 12 daily activities, moving along the radar line of sight. The activities observed were walking, jogging, limping, sitting, walking with a cane, walking with a walker, walking with crutches, crawling on hands and knees, creeping while dragging the abdomen on the floor, using a wheelchair, falling after tripping, and falling off a chair. A total of 1,007 measurements were collected. Each classification approach was tested using 10-fold cross validation unless otherwise noted. Given the limited measured data available, each method was first evaluated to determine the features or network structure that offered the highest classification accuracy:

1) *Supervised classification with handcrafted features*: A total of 127 micro-Doppler features were extracted: 10 DCT coefficients, three cepstral coefficients, 101 LPC coefficients, and 13 physical features—bandwidth and mean of the torso response; minimum, maximum, and mean of the upper and lower envelopes; total Doppler bandwidth; difference between the means of the upper and lower envelopes; and fundamental, second, and third harmonics of the CVD. The best performance was achieved using 50 features selected by sequential backward elimination: the bandwidth of the torso response, mean torso frequency, mean of the upper envelope, mean of the lower envelope, first two CVD features, first two cepstral coefficients, 37 LPC coefficients, and five DCT features. A multiclass SVM classifier with a linear kernel outperformed polynomial and radial basis function kernels as well as random forest and xgboost classifiers.
2) *2D PCA*: Generalized 2D PCA [23] was implemented where the projection was performed bilaterally on the rows and columns of the spectrogram, viewed as a matrix of dimension 128×128. The right (rows) and left (columns) projection matrices were of dimension 128×15 and 15×128, respectively, describing 15 principal components. A 3k-nearest-neighbor classifier was implemented with the size of training, validation, and testing sets chosen, respectively, to be 80, 10, and 10% of the data.
3) *AE*: A three-layer AE with layers of 200, 100, and 50 neurons, respectively, was found to yield the best results considering depth and width of the network. An adaptive moment estimation algorithm was used for optimization with a learning rate of 0.0001. The KL divergence term for regularization and sparsity parameter β were chosen as 2 and 0.1, respectively.
4) *CNN*: Two different convolutional filter sizes of 3×3 and 9×9 were applied and concatenated in each convolutional layer. Three convolutional layers using 2×2 max-pooling followed by two fully connected layers of 150 neurons each yielded the highest accuracy. To mitigate the potential for overfitting, 50% dropout was implemented.
5) *CAE*: The convolutional and deconvolutional layers of a three-convolutional layer CAE were populated with 30 3×3 and 9×9 convolutional filters. After unsupervised pretraining, two fully connected layers with 150 neurons each were used with a softmax classifier.
6) *Transfer learning from VGGnet pretrained on the ImageNet database*: Although there are many possible pretrained models that could be exploited to illustrate transfer learning, VGGnet outperforms GoogleNet and performs similar to ResNet-50 on radar micro-Doppler data [27]. Significantly, this contrasts with the results of these same networks on ImageNet and underscores the fact that results or conclusions from other domains do not directly translate to the radar domain.

Figure 7 compares the performance of these techniques and two emergent methods, GA-FWCC and DivNet-15 (discussed in the section "Emergent Techniques for DNN Performance Improvement in Radar"). The benefits of DL can be clearly seen as SVM offers just 76.9% accuracy in sharp contrast to the CAE at 94.2%. It is interesting to note that transfer learning from optical imagery yields only a slight improvement in performance. Yet, VGGnet is a 16-layer CNN while the CAE, a significantly less complex network, has only three convolutional layers. These results underscore the importance of using source training data that closely match those of the target domain.

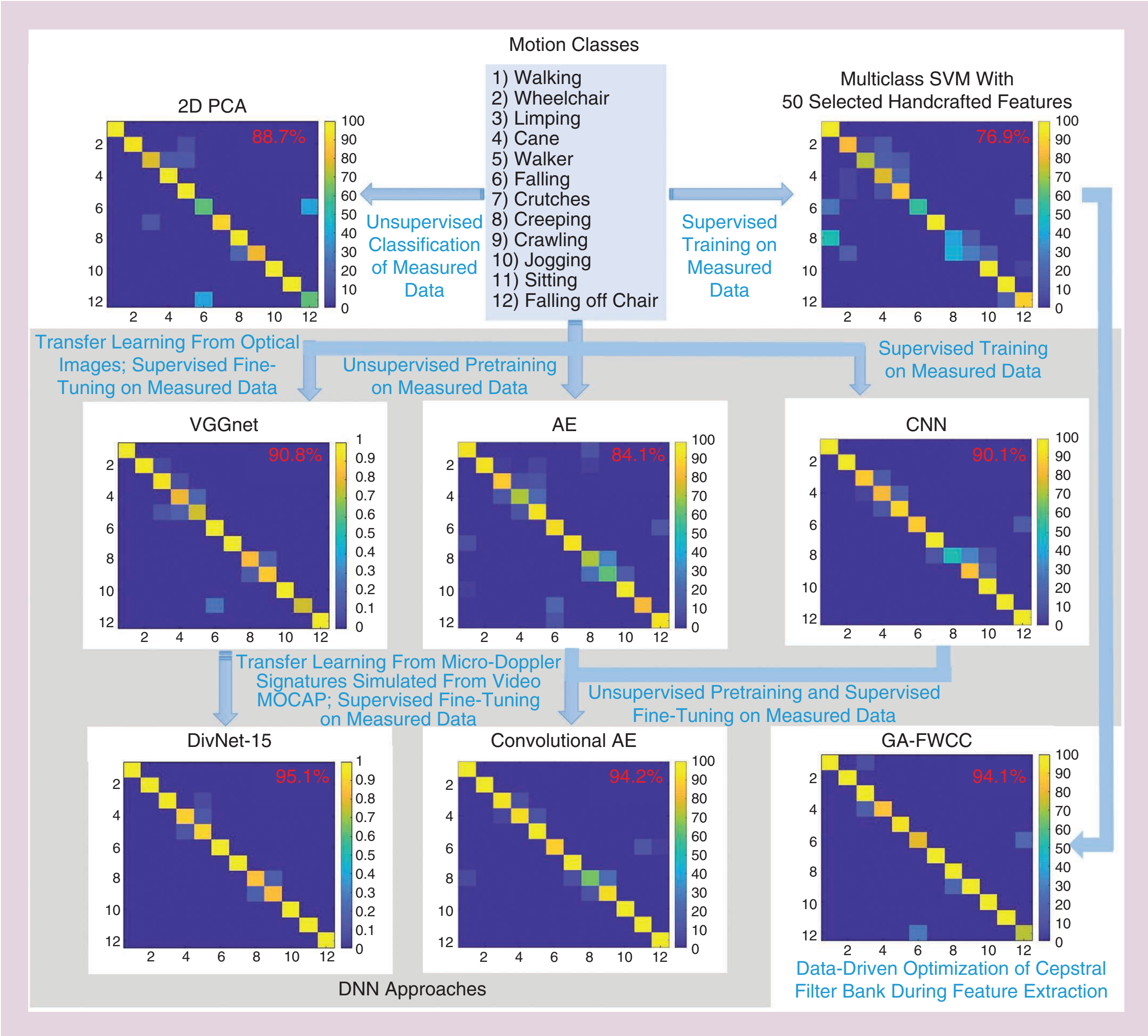

FIGURE 7. A diagram comparing conventional, data-driven, and DL classification approaches.

Emergent techniques for DNN performance improvement in radar

Transfer learning from simulated data sets

As highlighted by the case study, the depth and accuracy of DNNs are extremely limited by the characteristics and amount of training data available. Because the collection of radar data can be costly, only small measured data sets are typically acquired for training. Thus, the generation of simulated micro-Doppler signatures for use in training has been a focus of recent research. One recently proposed technique [27] exploits the motion capture (MOCAP) data acquired from RGB-D (red, green, blue, depth) cameras to perform skeleton tracking and record the time-varying distance to various points on the human body. While high-accuracy MOCAP systems typically track reflective markers attached to the body, lower-cost markerless skeleton tracking is also possible using such devices as the Kinect sensor. Thus, MOCAP data can provide estimates of the distance to K points on the human body, which can be used in lieu of $R_{n,i}$ measured by radar. The model of the electromagnetic backscatter provided by (1) can then be used to simulate the expected signal received by radar.

Because the skeleton corresponding to the motion being recorded by the Kinect sensor is accessible, transformations of the skeleton can be applied to generate new instances of the micro-Doppler signatures that correspond to different target profiles. This is essentially a form of data augmentation, a technique used in computer vision for creating minor alterations of the image (e.g., flipping, rotating, scaling) in an attempt to generate new samples and add to the training data set. However, in the case of radar, such transformations, when directly applied to the target 2D joint TF representation, would yield noninterpretable and kinematically impossible samples that would have an adverse effect on training. Instead, by applying transformations on the source Kinect data and underlying skeleton tracking, alterations can be induced that are consistent with the electromagnetic interaction of the RF waves with the target and its kinematics. Two possible transformations are extension or compression of the time axis to speed up or slow down the measured motion. Also, the physical dimensions of the skeleton can be scaled to account for smaller or larger people. Another possible transformation seeks to represent variations due to individual gait styles. This transformation can be obtained by perturbing the coefficients of the Fourier series that best fits the trajectory of each tracked point on the body versus time. Applying combinations of these transformations permits a controlled simulation that spans the range of expected human motion for any radar system. Not only does this approach help with target generalization, but it can be used to build up a large number of simulated data from a small number of MOCAP measurements, thereby enabling deeper DNN training. Once the model for the received data is generated with (1), any desired data representation can be generated as training data. After network initialization using the simulated training data, only a small number of real radar data is required to fine-tune the training of the weights and test performance.

The efficacy of this training approach can be seen by applying the common data set in the case study described in the section "DNN Performance for Daily Human Activities." A Kinect sensor was used to collect 55 MOCAP measurements from five participants, none of whom were among the 11 participants from whom the radar data were collected. The diversification methodology, detailed in [27], was applied to expand the data set to 32,000 statistically independent samples, after which the spectrogram was computed. The spectrograms of both real and simulated domain data were computed to form a 2D input representation to the DNN. Because of the significantly larger amount of diversified MOCAP data available for training, the depth of the network could now be increased to 15 layers without overfitting. Modified residual units were used to mitigate training errors. The resulting performance of the 15-convolutional-layer residual neural network, labeled DivNet-15 and shown in Figure 7, yields the highest accuracy of the methods surveyed in this article: 95.2%.

In computer vision, generative adversarial networks (GANs) have been proposed to generate highly realistic simulated images. These networks have been successfully used to generate simulated data for the classification of synthetic aperture radar imagery. However, this approach has yet to be applied to the radar data generation of human micro-Doppler signatures. One problem is that variations in the synthetic data created by GANs may be unrelated to the underlying target kinematics. This could potentially result in the generation of misleading training data. Nevertheless, broad studies on the generation of synthetic radar micro-Doppler data with GANs deserve detailed investigation, with special emphasis placed on comparing the properties of that data with those of the diversified MOCAP data previously described.

Input data representations

Classification is typically not directly performed on the raw 1D radar measurements, as this would drastically increase the computational cost and complexity of DNNs now forced to learn extremely high dimensional mapping between the raw data and motion classes. Instead, radar engineers have espoused a knowledge-aided approach using a number of tools, such as range processing, clutter mitigation, and TF analysis, to present a preprocessed form of the data that reveals salient target properties. The micro-Doppler signature, as computed with a spectrogram, has been the most commonly used input data representation for motion recognition. Spectrograms highlight the variation in Doppler shift induced by human motion as a function of time, and thus provide the DNN with a valuable representation for feature learning. However, radar measures not just Doppler frequency or velocity but also target range. Researchers have shown increasing interest in joint-domain representations involving alternative ways of presenting the time-varying range and Doppler information to the DNN. Figure 8 shows that three types of 2D inputs can

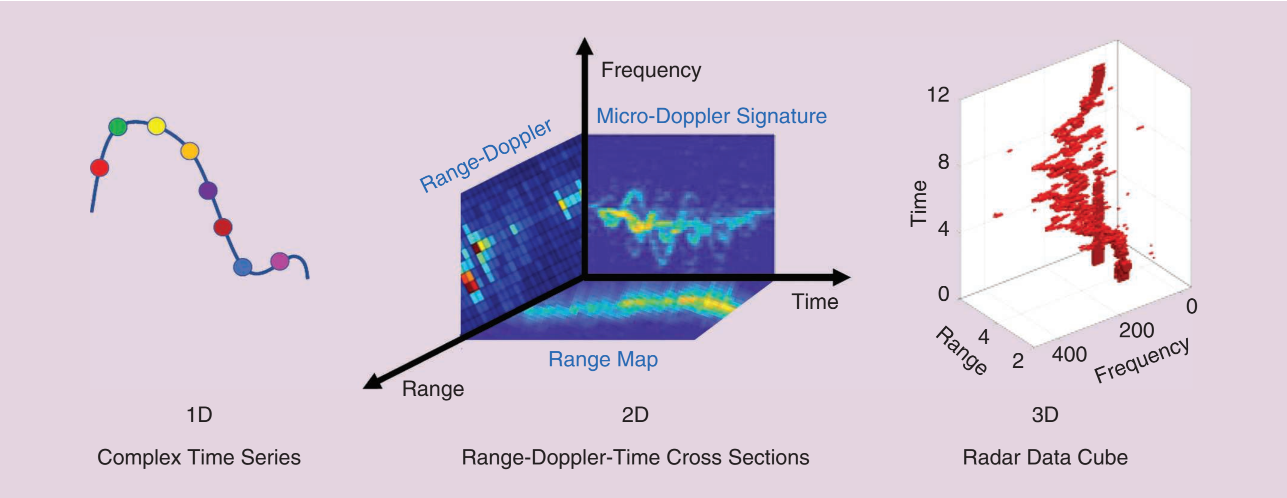

FIGURE 8. Three input representations of radar data.

be generated from the raw radar data: micro-Doppler signatures, range maps, and range-Doppler images. The range map shows the variation of target position with slow-time, while range-Doppler images show both the target position and velocity. Analysis with multiple joint-domain representations has been shown to be more effective than when individual representations are used separately [28]. These 2D representations are cross sections that can be expanded to a 3D representation, the radar data cube, which permits the use of the correlation and interdependency among the three variables.

The micro-Doppler signature, as computed with a spectrogram, has been the most commonly used input data representation for motion recognition.

The 3D range-Doppler-time data cube has been used to improve fall-detection rates to more than 96% [29], while recent research on eight-class and 11-class hand-gesture recognition reports more than 85% accuracy in identifying motions for device control, such as a turning a virtual knob or moving a virtual slider. In [19], 3D input was generated by stacking time-shifted versions of a fixed-duration micro-Doppler signature, while in [15], the range-Doppler image was stacked as a function of time.

An important motivation for seeking new ways to present radar data to the DNN is to better understand what the neural network learns. Activation maps and heatmaps are two ways to visualize network learning. Examination of activation maps has revealed, for instance, that various DNNs respond differently to spurious signals at the input [26], whereas heatmaps assign each pixel a relevance score that identifies the parts of the image that have the most impact on classification performance.

Since spectrograms are not typical 2D images, frequency-based metrics, in addition to spatial metrics, have yielded great insights. Consider a frequency-warped cepstral analysis, which has recently been proposed for data-driven classification. The approach is essentially a modification of mel-frequency cepstral analysis, first proposed for speech processing and poorly matched to the spectral properties of radar micro-Doppler. Genetic-algorithm optimized frequency-warped cepstral coefficients (GA-FWCCs) are tailored to the spectral properties of radar micro-Doppler. During training, the GA-FWCC of the center frequency and spectral width of each cepstral filter bank are optimized for maximum classification accuracy. The GA-FWCC's efficacy can be assessed using the 12-class case study described in the section "DNN Performance for Daily Human Activities." When 500 GA-FWCC features, generated from 10 optimized filters, are supplied to a single-layer, 50-node artificial neural network, a classification accuracy of 94.1% is achieved. This is on par with the accuracy of the CAE and just 1% less than that of DivNet-15, as shown in Figure 7. The high accuracy achieved by GA-FWCCs highlights the frequency bands important for classification, as revealed by the GA-optimized filters. Thus, researchers looking into how to best present radar data to DNNs should work on designing attention-driven neural networks that exploit relevancy analysis in both spatial and frequency domains.

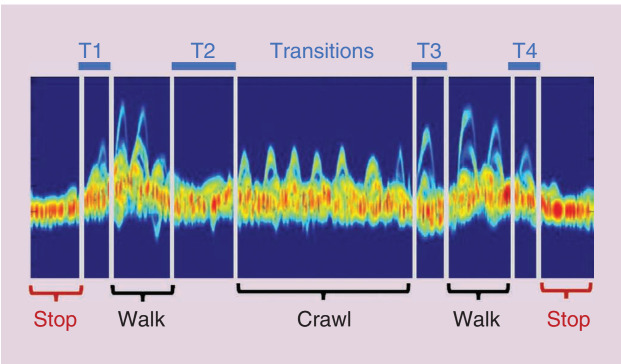

FIGURE 9. A micro-Doppler dynamic motion sequence recording walking, then crawling, then walking, with transitions (T1, T2, T3, T4) in between.

Recurrent neural networks

The methods surveyed so far have relied exclusively on capturing a snapshot of the motion over a finite window of time. However, in daily life, human motion is a continuous stream of many different motions of varying durations. Moreover, natural human-motion sequences include not just short periods over which a well-defined action occurs but also transition periods during which the body repositions itself appropriately to perform the subsequent action. Consider, for example, the sequence of "walk–crawl–walk" depicted in Figure 9. The variation in micro-Doppler frequencies over the transition period matches neither the characteristic of walking nor that of crawling. Accurate classification of daily human motion thus requires not only the ability to identify and parse long-duration signatures according to motion class, but also the ability to identify transition periods of varying duration.

While the sequential classification of dynamic motion remains an open problem, recent work exploiting recurrent neural networks (RNNs) has made important progress toward addressing this issue. RNNs produce an output at each time step and include connections between nodes that form a directed graph along a sequence. This structure enables RNNs to effectively model temporal dynamic behavior. Long short-term memory (LSTM) RNNs model longer-term behavior through the inclusion of a memory block that consists of a cell, input, output, and forget gates. In gesture recognition, a combination of 3D inputs has been used with LSTM RNNs for classification [15], [19], but results were only presented for data samples containing a single motion class. Another RNN variant, which has been proposed for micro-Doppler analysis, is the gated recurrent unit (GRU). This approach has a simple structure and shows better performance on small data sets. In [30], a stacked GRU network was proposed for the sequential classification of micro-Doppler, where the micro-Doppler signatures of different activities were concatenated to generate a sequence of different motion classes.

This network was shown to significantly outperform a CNN in classifying the entire sequence.

Challenges and future directions

The improved performance offered by DL has opened the door for the use of radar in a wide range of applications, which in turn drives more stringent requirements for pervasive and robust motion recognition. The main challenges facing indoor human-motion classifications using DL are the same as those for any other classification technique: specifically, developing devices that can achieve both high specificity and sensitivity while also being computationally efficient and inexpensive. Sensitivity is the ability to correctly classify a given motion, whereas specificity is the ability to correctly classify its nonoccurrence. For fall detection, these measures translate, respectively, into probability of detection and false alarm. However, for daily-activity and gesture recognition, a multiple-hypothesis testing problem emerges with the goal of minimizing the probability of error. This goal must be achieved in a variety of indoor settings with many sources of clutter and interference. Specific challenges include:

1) *Motion generalization*: The radar micro-Doppler signature depends not just on the motion being performed but also on the physique, speed, and walking style of the individual being observed. The ability to generalize human motion to include unknown or unobserved people is critical to the credibility of the radar classifier. In essence, desired performance must be maintained for a "generic" person. Innovations in human-motion modeling and simulations for DNN training have been introduced to address in-class variations. Another aspect of motion generalization relates to the fact that there is an infinite number of possible human activities. New approaches must, therefore, be designed to address the open-set problem, including cognition-inspired algorithms that can aid in generating and revising multimission classification problems.
2) *Dynamic motion characterization*: Human motion is an inherently dynamic time stream of actions, where an observation can include multiple types of motions as well as transitions in between. A key idea driving innovation in DNNs for radar applications is to integrate physics-based models with the DNN architecture, resulting in a knowledge-aided approach to classification. The change in micro-Doppler during transitions, or even potential motion sequences, is not completely unknown or random but, instead, is constrained by the human body and its kinematics. The progression of micro-Doppler can be modeled a priori to understand what transition regions should resemble. Moreover, the observation of daily habits can be used to generate the likelihood for certain motion sequences—e.g., standing is more likely to follow sitting, after which walking is expected to follow. By determining the likelihood, some motion classes can be discarded as too unlikely to occur. Such memory-based models can aid in motion sequencing as well as serve to limit the openness of the classification problem.
3) *Personalized classification*: Remote health-monitoring applications require the human gait to be characterized with great sensitivity to detect gait abnormalities. Such problems as imbalance play a role in fall-risk assessment as well as in post-stroke rehabilitation, while the long-term estimation of gait parameters is critical to the treatment of neuromuscular disorders. In such applications, the ailments involved cause such small differences that they can only be detected by a highly sensitive, "personalized" radar specifically attuned to the motion patterns of the individual being monitored.
4) *Motion decomposition*: Another important feature with applications to health and gesture recognition is the ability to separate and disentangle the micro-Doppler signatures from people near each other, as well as to isolate those motion components resulting from gestures versus gross body movements. The latter requires a high resolution in range and angle. Thus, while MOCAP has been helpful for synthesizing total micro-Doppler signatures from measurements of body parts, the inverse problem of decomposing the micro-Doppler signature remains unsolved.
5) *Pervasive recognition*: In typical indoor environments, the radar line of sight may not always be aligned with the person's direction of motion. As the aspect angle increases, the bandwidth of the micro-Doppler signature decreases because radar is sensitive to radial velocity not absolute velocity. Moreover, it is likely that furniture or other objects may obstruct the line of sight and block the radar from observing part of the human motion. Through-the-wall radar and model-based approaches that take into account blocked signal components during training can mitigate some of these challenges. Multistatic radar networks have also been proposed to overcome some physical limitations. Researchers are looking into fusion, which are adaptive approaches driven by information sharing, and novel DNNs [9] that can facilitate multisensor exploitation, including passive Wi-Fi sensing.

Conclusions

This overview of DL for human-motion classification has covered a broad range of applications. The tremendous performance gains offered by DL have been illustrated with a 12-activity class case study in which the accuracy of a variety of DNNs was compared with that of conventional classification using handcrafted features and other data-driven approaches. A particular challenge to applying DL to radar has been the small amount of real data available for training. Unsupervised pretraining and transfer learning have been discussed as methods to overcome the limitation in depth and accuracy incurred with a small training sample. The value of integrating physics-based models into the DNN training process has been demonstrated to yield the highest classification accuracy of the compared methods. This is related as well to the way radar data are presented to the DNN using 1D, 2D, or 3D input representations that best capture the salient spatial and frequency-based signal characteristics relevant to classification. The development of novel approaches that leverage advancements in DNN design and optimization with knowledge-aided radar-signal processing techniques will be key to addressing current challenges related to generalizing motion, characterizing the transient and sequential flow of motion, developing high-sensitivity personalized devices, and designing robust, pervasive recognition systems.

Authors

Sevgi Zubeyde Gurbuz (szgurbuz@ua.edu) received her B.S. and M.Eng. degrees from the Massachusetts Institute of Technology, Cambridge, in 1998 and 2000, respectively, and her Ph.D. degree from the Georgia Institute of Technology, Atlanta, in 2009, all in electrical engineering. She served as a radar-signal processing research engineer in the U.S. Air Force Research Laboratory, Rome, New York, from 2000 to 2004. Formerly with TOBB University and the Scientific and Technological Research Council of Turkey Space Technologies Research Institute in Ankara, she is currently an assistant professor of electrical and computer engineering at the University of Alabama, Tuscaloosa. Her research interests include radar-signal processing, machine learning, and remote sensing. She is a Senior Member of the IEEE.

Moeness G. Amin (Moeness.amin@villanova.edu) received his B.Sc. degree from Cairo University, Egypt, in 1976, his M.Sc. degree from the University of Petroleum and Minerals, Dhahran, Saudi Arabia, in 1980, and his Ph.D. degree from the University of Colorado, Boulder in 1984, all in electrical engineering. He is director of the Center for Advanced Communications, Villanova University, Pennsylvania. He is the recipient of the 2017 Fulbright Distinguished Chair, the 2016 Alexander von Humboldt Research Award, the 2016 Institution of Engineering and Technology Achievement Medal, the 2014 IEEE Signal Processing Society Technical Achievement Award, the 2009 European Association for Signal Processing Technical Achievement Award, the 2015 IEEE Aerospace and Electronic Systems Society Warren White Award for Excellence in Radar Engineering, and the IEEE Third Millennium Medal. He is a Fellow of the IEEE, the International Society of Optical Engineering, the Institution of Engineering and Technology, and the European Association for Signal Processing.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 770–778.

[2] S. Gurbuz, B. Erol, B. Cagliyan, and B. Tekeli, "Operational assessment and adaptive selection of micro-Doppler features," *IET Radar, Sonar Navigation*, vol. 9, no. 9, pp. 1196–1204, 2015.

[3] 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, Jan. 2016.

[4] M. Seyfioglu, A. Ozbayoglu, and S. Gurbuz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 4, pp. 1709–1723, Aug. 2018.

[5] M. Amin, *Radar for Indoor Monitoring: Detection, Classification, and Assessment.* Boca Raton, FL: CRC Press, 2017.

[6] C. Li, V. M. Lubecke, O. Boric-Lubecke, and J. Lin, "A review on recent advances in Doppler radar sensors for noncontact healthcare monitoring," *IEEE Trans. Microw. Theory Techn.*, vol. 61, no. 5, pp. 2046–2060, May 2013.

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

[8] A. Seifert, A. M. Zoubir, and M. G. Amin, "Radar classification of human gait abnormality based on sum-of-harmonics analysis," in *Proc. 2018 IEEE Radar Conf. (RadarConf)*, pp. 940–945.

[9] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Personnel recognition and gait classification based on multistatic micro-Doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 669–673, 2018.

[10] X. Yang, W. Yin, and L. Zhang, "People counting based on CNN using IR-UWB radar," in *Proc. 2017 IEEE/CIC Int. Conf. Communications in China (ICCC)*, pp. 1–5.

[11] P. Cao, W. Xia, M. Ye, J. Zhang, and J. Zhou, "Radar-ID: Human identification based on radar micro-Doppler signatures using deep convolutional neural networks," *IET Radar, Sonar Navigation*, vol. 12, no. 7, pp. 729–734, 2018.

[12] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. D. Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3941–3952, July 2018.

[13] Y. Li, Z. Peng, and C. Li, "Potential active shooter detection using a portable radar sensor with micro-Doppler and range-Doppler analysis," in *Proc. 2017 Int. Applied Computational Electromagnetics Society Symp. (ACES)*. doi: 10.1109/JSEN.2018.2879223.

[14] F. Fioranelli, M. Ritchie, and H. Griffiths, "Classification of unarmed/armed personnel using the netrad multistatic radar for micro-Doppler and singular value decomposition features," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1933–1937, 2015.

[15] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with Soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proc. 29th Annu. Symp. User Interface Software and Tec*hnology, 2016, pp. 851–860.

[16] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[17] T. Sakamoto, X. Gao, E. Yavari, A. Rahman, O. Boric-Lubecke, and V. M. Lubecke, "Hand gesture recognition using a radar echo IQ plot and a convolutional neural network," *IEEE Sensors Lett.*, vol. 2, no. 3, pp. 1–4, Sept. 2018.

[18] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-driven micro-Doppler feature extraction for dynamic hand gesture recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 2, pp. 655–665, Apr. 2018.

[19] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3278–3289, Apr. 2018.

[20] Y. Kim, "Detection of eye blinking using Doppler sensor with principal component analysis," *IEEE Antennas Wireless Propag. Lett.*, vol. 14, pp. 123–126, 2015.

[21] L. Cohen, "Time–frequency distributions—a review," *Proc. IEEE*, vol. 77, no. 7, pp. 941–981, July 1989.

[22] S. 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, 2017.

[23] H. Kong, L. Wang, E. Teoh, X. Li, J. Wang, and R. Venkateswarlu, "Generalized 2D principal component analysis for face image representation and recognition," *Neural Netw.*, vol. 18, pp. 585–594, 2005.

[24] J. Zabalza, C. Clemente, G. D. Caterina, J. Ren, J. J. Soraghan, and S. Marshall, "Robust PCA micro-Doppler classification using SVM on embedded systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 3, pp. 2304–2310, July 2014.

[25] I. Goodfellow, Y. Bengio, A. Courville, and F. Bach, *Deep Learning.* Cambridge, MA: MIT Press, 2016.

[26] M. Seyfioglu and S. Gurbuz, "Deep neural network initialization methods for micro-Doppler classification with low training sample support," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2462–2466, Dec. 2017.

[27] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "DNN transfer learning from diversified micro-Doppler for motion classification," *IEEE Trans. Aerosp. Electron. Syst.*, Dec. 2018. doi: 10.1109/TAES.2018.2883847.

[28] B. Jokanovic and M. Amin, "Fall detection using deep learning in range-doppler radars," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 1, pp. 180–189, Feb. 2018.

[29] B. Erol and M. Amin, "Radar data cube analysis for fall detection," in *Proc. 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2446–2450.

[30] M. Wang, G. Cui, X. Yang, and L. Kong, "Human body and limb motion recognition via stacked gated recurrent units network," *IET Radar, Sonar Navigation*, vol. 12, no. 9, pp. 1046–1051, 2018.

SP