IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 22, 2019, accepted September 8, 2019, date of publication September 19, 2019, date of current version October 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942305

GestureVLAD: Combining Unsupervised Features Representation and Spatio-Temporal Aggregation for Doppler-Radar Gesture Recognition

ABEL DÍAZ BERENGUER[1], MESHIA CÉDRIC OVENEKE[1], HABIB-UR-REHMAN KHALID[1,2], MITCHEL ALIOSCHA-PEREZ[1], ANDRÉ BOURDOUX[2], AND HICHEM SAHLI[1,2]

[1]VUP-NPU Joint Audio-Visual Signal Processing (AVSP) Research Laboratory, Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), 1050 Brussels, Belgium
[2]Interuniversity Microelectronics Centre (IMEC), 3001 Heverlee, Belgium

Corresponding author: Abel Díaz Berenguer (aberengu@etrovub.be)

ABSTRACT In this paper we propose a novel framework to process Doppler-radar signals for hand gesture recognition. Doppler-radar sensors provide many advantages over other emerging sensing modalities, including low development costs and high sensitivity to capture subtle gestures with precision. Furthermore, they have attractive properties for ubiquitous deployment and can be conveniently embedded into different devices. In this scope, current recognition methods still rely in deep CNN-LSTM and 3*D* CNN-LSTM structures that require sufficient labelled data to optimize millions of parameters and significant amount of computational resources for inference; which limits their deployment. Indeed, subtle gestures recognition is a challenging task due to the high variability of gestures among different subjects. To overcome the challenges in the recognition task and the limitations of the current methods, we propose a shallow learning approach for gesture recognition, that is based on unsupervised range-Doppler features representation, along with a learnable pooling aggregation via NetVLAD. The proposed framework can encode extremely valuable information across time, and results in features that are highly discriminative for hand gesture recognition. Experimentation on publicly available Doppler-radar data shows that the proposed framework outperforms state-of-the-art approaches in terms of recognition accuracy and speed for sequence-level hand gesture classification.

INDEX TERMS Convolutional neural networks, Doppler-radar, feature aggregation, hand gesture recognition, unsupervised representation learning.

I. INTRODUCTION

Gesture recognition has become increasingly important in human-computer interaction (HCI) and can support a broad array of emerging applications, such as smart home, virtual reality, driver assistance, and mobile gaming. Prior work in gesture recognition mainly relies on (i) cameras such as RGB cameras, depth cameras or infrared cameras that require line of sight, (ii) dedicated sensors (e.g., Radio Frequency Identification Reader (RFID), gloves, motion sensors) that are worn by the user, and (iii) Radio Frequency (RF) based gesture recognition using either specialized or commodity RF devices. These approaches, however, require significant deployment overhead and incur non-negligible cost [1].

Recently, Doppler-radar sensors showed promising performance in human gesture recognition, attracting significant interests in the microwave community and consumer electronics industry [2]–[11]. Doppler-radar based techniques utilize time-frequency analysis on the Doppler shift introduced by the movement of the hand. The advantage is that it can be implemented in low-cost devices with a simple front-end architecture. Examples, include the Soli sensor developed by Google's Advanced Technology and Projects Group (ATAP) [2]. Soli is a solid-state millimeter-wave radar for micro-interactions in mobile and wearable computing. Soli uses a sensing approach that prioritizes high temporal resolution to detect subtle, non-rigid motion. It utilizes a single broad antenna beam to illuminate the entire hand as

The associate editor coordinating the review of this manuscript and approving it for publication was Yongle Wu.

modulated pulses are transmitted at very high repetition rates (for details see [2]).

Robust gesture recognition is a difficult problem due to the spatio-temporal variations in gesture formation and subjects. It has been addressed using several machine learning approaches, such as Hidden Markov Models [12], Dynamic Time Warping [13], Recurrent Neural Networks (RNNs) [14], Echo State Networks [15] and Convolutional Long Short-Term Memory (CLSTM) [16].

Two categories of literature exist for deep learning-based Doppler-radar gesture recognition algorithms. In the first category, the deep learning framework is trained on the input Doppler-radar data without explicitly modeling the inter-frame temporal information. For such methods, Convolutional Neural Networks (CNNs) are trained on the input Doppler-radar [3], [17], [18]. Regardless of the unquestionable success of deep CNNs trained in a supervised manner for several tasks, to achieve a reasonable local minimum while optimizing the millions of parameters conforming such structures, extremely large labelled datasets are required. Nonetheless, extensive labelled datasets are awkward to build up and there are no guarantees of better results at making such datasets huge. Hence, unsupervised learning approaches, which can learn intermediate features representation directly from the input data distribution have attracted the attention of several researchers. Furthermore, our main motivation towards an unsupervised learning strategy is driven by the high variability across different individuals when performing the same gesture; which has been referred in previous studies as one of the key challenges for hand gesture recognition [19]. In this sense, the experimental evidence [20]–[26] suggests that unsupervised feature learning approaches can enforce representations robust to variations that are irrelevant for the final recognition task.

In the second category of deep learning-based Doppler-radar gesture recognition, researchers model the inter-frame temporal information using CLSTM, in which CNNs and $3D$ CNNs are used to obtain intermediate signal representations, either for frame-based [10], [27], or short spatio-temporal features extraction [28], [29]; and LSTM layers are employed for modeling the inter-frame temporal dynamics. In such approaches, multiple frames, sampled from the radar sequence, are given as input to the network to perform classification. In [10], the authors achieved reasonable accuracy in both frame-based and sequence-based classification. However, one of the key challenges in these approaches is the difficulty of selecting the sampled frames because durations of different gestures can vary significantly.

To alleviate for the above challenges, and motivated by the recently developed powerful pooling techniques to aggregate variable-length inputs into a fixed-length representation when dealing with sequential data [30], [31], in this work, we adopt the NetVLAD approach [32] to explicitly model the inter-frame temporal information.

The main contribution of this work is the combination of an unsupervised representation learning strategy with the NetVLAD approach for Doppler-radar gesture recognition problem. The primary objective of the unsupervised representation learning strategy is to retain short-term temporal structure between frame-level Doppler-radar features and their spatial inter-dependencies in the representation, whereas enforce invariance. Furthermore, the introduction of the NetVLAD approach for gesture recognition drives our proposal to obtain discriminative sequence representations allowing to keep those gestures of different classes far apart and those of the same class near; even when they are potentially variant due to gesture formation and subjects intrinsic differences. We evaluated the proposed framework in terms of representation learning task, gesture recognition task, and speed using Doppler-radar provided by [10]. Our proposal can yield discriminative representations and obtains 98.24% recognition score at sequence-level classification, whereas at inference phase it achieves rather fast recognition rates; which demonstrates its potentials for intelligent control applied to an emerging sensor technology that exhibits unique properties and would enable new input modalities for touchless HCI in a broad array of devices.

II. RELATED WORK

In this section, we briefly review related methods to learn features representation and aggregate frame-based features for sequence classification, particularly those related to our approach developed for gesture recognition.

In general, the first step in sequence classification is to process the frames of the sequence to extract frame-based features using either hand-crafted features [2], [5]–[7] or CNNs [3], [10], [18], to get intermediate layer activations as frame features.

Over the years, several studies based on radar sensors have adopted hand-crafted features representation. Lien *et al.* [2] proposed an entire system to sense and recognize hand gestures using Soli. The authors explored a gesture recognition pipeline based on motion signatures to demonstrate the advantages of Soli. Domain engineering knowledge was required to obtain features representation relevant for recognition. Hence, generic low-level features, such as fine displacement, total measured energy, measured energy from moving scattering centers, scattering center range and velocity centroid were extracted [2]. Moreover, the generic low-level representations were combined with gesture specific features and used to recognize the gestures with a Random Forest classifier; which was selected as a suitable classifier by benchmarking among different classifiers. The machine learning pipeline exposed in [2] mainly relied on gesture specific hand-crafted features since the generic representations were not effective to discriminate across several gestures and not enough to achieve robust recognition. Evermore, the authors pointed out the difficulties in designing features which reflect the gestures and can enable robust recognition. The authors argued that deep CNNs were not suitable for this problem due to the computational burden required for deep structures [2]. Similarly, in [33] Random

Forest was proposed to recognize hand gestures sensed using Soli. They extracted specific features representation to recognize predefined gestures for vehicular infotainment control.

Apart from the Random Forest, k-Nearest Neighbor (k-NN) classifiers, Support Vector Machines (SVM) and similarity measurements for sequences, such as Hausdorff distance and Dynamic Time Warping (DTW), have also been employed along with hand-crafted features for radar-based sequence recognition. Sun *et al.* [6] developed an approach based on hand-crafted features extracted from micro-Doppler signatures to classify gestures using k-NN. In [34] the authors proposed a gesture recognition pipeline for smart home, which made use of k-NN for gesture classification based on application specific features extracted from radar signals. Moreover, the effects of sparsity and time frequency for dynamic micro-Doppler hand gestures recognition have been extensively investigated in a series of studies by Li *et al.* [7], [35], [36]. They studied different hand-crafted features and diverse machine learning approaches such as SVM, Naïve Bayes, and Nearest Neighbor combined with Hausdorff distance to classify dynamic hand gestures. Zhuo *et al.* [5] proposed a whole system to recognize dynamic gestures employing a terahertz radar. The authors made use of high-resolution range profile sequences, whose can be obtained with the terahertz radar, in conjunction with Doppler signatures to extract sequence representations from both signals. For the final classification, they employed an extended DTW algorithm to measure the similarity between multimodal sequences.

In general terms, radar-based hand-crafted features for human action recognition have been extensively investigated. Nonetheless, the main limitation of these solutions is the feature engineering design. The effectiveness of such features depends on the application and the specific human domain knowledge, which leads to features representation that lack generalization capabilities. In consequence, more recent attention has focused on CNNs approaches capable to learn features representation directly from data [1], [28], [37]. Despite the indisputable success of deep CNNs, their reliance on huge amounts of labelled data is a limiting factor when approaching new applications [26], such as Doppler-radar based gesture recognition. We therefore argue that there is a growing need for unsupervised learning strategies capable of training CNNs.

A large body of studies have investigated unsupervised representation learning [38]. Among them, auto-encoders (AEs) [38], denoising auto-encoders (DAEs) [38], [39] and reconstruction contractive auto-encoders (RCAEs) have been extensively employed for unsupervised feature learning. The basic idea behind AEs is to learn a function that minimizes the squared error between the input signal and its reconstruction. Moreover, DAEs yield features robust to noise because can learn to denoise randomly corrupted signals to reconstruct the input data. Furthermore, RCAEs [40] can learn to prioritize relevant factors of the input signal, which leads to useful representations of the data for subsequent learning tasks. Nonetheless, a key challenge for AEs, DAE and RCAEs is that generally these algorithms tend to ignore the structure of two dimensional input signals, which enforces to learn global representations instead of capturing local spatial variations in the data. In [41], the authors proposed the convolutional auto-encoders (CAEs), whose structure is designed based on $2D$ convolutions to overcome the issues related to local spatial representations. They trained the CAEs in an unsupervised manner and subsequently stacked them to conform a CNN. The CNN was initialized with the unsupervised learned hyper-parameters from the CAEs. This approach achieved good performance for visual object recognition tasks. However, the authors relied on a deep CNN architecture.

Dosovitskiy *et al.* [26] proposed an approach to learn generic features using unlabelled data. This proposal is based on a set of image transformations which are used for learning a representation regarding an auxiliary task directly related towards the final classification the authors aimed for. The experimental evidence in [26], suggests that when unsupervised features learning algorithms can exploit unlabelled data effectively, it is possible to extract meaningful feature representation, whose is not related to one specific class instance and robust to image transformations. Nonetheless, the authors required a deep CNN structure to achieve the reported performance.

Unsupervised feature learning has been also studied in radar-based classification of sequences. Seyfioğlu *et al.* [42] presented a deep convolutional auto-encoder to classify indoor human activities. The authors faced the problem of dealing with highly similar radar-based signatures; and proposed a strategy to pretrain, in unsupervised manner, the hyper-parameters of a three-layer CAE, whose are then used as initial values for supervised training. The authors pointed out that such training strategy was useful to learn features representation from two dimensional radar-based signatures.

In this work we are dealing with Doppler-radar hand gesture recognition problem, and ideally, we could directly rely on labelled samples. However, people perform hand gestures in a variety of ways which make the task difficult, because learning a discriminative set of predefined feature related to each class sample is not straightforward due to the intra-class variability. Therefore, in a similar aim as in [22], [26] and [42], we propose to learn unsupervised frame-level representations directly from data. We propose to employ a fast and accurate layer-wise unsupervised learning strategy for CNNs, which trains the convolutional and fully connected (FC) layers using the reconstruction contractive auto-encoding (RCAE) objective as presented in [43], [44]. Essentially, we train the hyper-parameters of the network by transforming the reconstruction contractive auto-encoding objective [40] to a convexified variant in the frequency domain, whose is optimized using the Gauss-Seidel (GS) algorithm [45]. Our class agnostic learning strategy drives the features representation to discover relevant factors in the input frames, whereas enforces invariance to slightly

differences caused by variability among subjects; which leads to a representation that could be directly used for the recognition task or fine-tuned in supervised manner.

Over the recent years, researchers have developed solutions based on deep learning to aggregate frame features extracted by CNNs. Several RNNs solutions have been proposed. Most researchers utilized LSTMs to aggregate frame features extracted by CNNs [10], [28], [46]. Moreover, several solutions have been proposed to improve the performance of RNNs. For video action recognition, Lev *et al.* [47] proposed RNN Fisher Vectors (FV), in which a sequence is represented by using derived gradient from the RNN as features, instead of using a hidden or an output layer of the RNN. The proposed representation is sensitive to the element ordering in the sequence and provides a richer model than the additive ''bag'' model typically used for conventional FV. In [28] is proposed a whole system to recognize unsegmented hand gestures continuously. The authors designed a gesture recognition system for a Frequency Modulated Continuous Wave (FMCW) radar sensor. The recognition is based on $3D$ convolution applied on short sequences of 8 frames, to extract spatio-temporal features which are aggregated with LSTMs aided by a connectionist temporal classification to deal with unsegmented sequence. Likewise, recently Wang *et al.* [29] made use of $3D$ convolution along with LSTM for Doppler-based hand gesture recognition. After preprocessing the FMCW radar signal each gesture was represented as a sequence of 32 range-Doppler frames. The authors used a simplified, however still-deep version, of the $I3D$ network [48] to produce range-time and Doppler-time spatial-temporal representations. Further, the generated range-time and Doppler-time features were used as input to two LSTMs to capture global dynamics. Finally, classification was performed using the Softmax classifier. Note that, the Gated Recurrent Unit (GRU) [49] has also been successfully used in replacement to the LSTM for video-based gesture recognition, while being more computationally efficient. To further enhance the performance of the bidirectional-GRU, Li *et al.* [50] incorporated the Fisher criterion into the Softmax loss function for gesture recognition from accelerometer and gyrometer signals.

Approaches based on deep learning to aggregate intermediate features representation extracted by $2D$ and $3D$ CNNs have achieved an unquestionable success. Admittedly, one major drawback of these methods for hand gesture recognition is that deep network structures increase the computational burden. Thus, in this work we propose a framework which can effectively exploits $2D$ CNNs features learned directly from data, while remaining shallow.

To aggregate frame-based features for sequence-based prediction, early research considered averaging over the sequence for the subsequent regression or classification tasks. Bag of Words (BoW) [51], Vector of Locally Aggregated Descriptors (VLAD) [52], and FV [53] have been used as sequence encoding approaches. BoW and VLAD use k-means to cluster the data, while FV adopts the Gaussian Mixture Model (GMM) approach. BoW and its variants have dominated research in action and gesture recognition for a long time [54], [55]. BoW consists of four main steps: feature extraction, codebook generation, feature encoding and pooling, and normalization. Whereas BoW aggregation keeps a count of the ''visual'' words (a codebook of k centroids), the VLAD [52] stores the sum of the residuals (difference between the descriptor and the mean of its corresponding cluster) for each visual word. Originally developed in the image analysis domain, VLAD has been successfully applied to gesture recognition for sign language recognition [56]. The FV approach [53] transforms an incoming set of descriptors into a fixed-size vector representation, which describes how the sample of the descriptors deviates from a probabilistic visual vocabulary usually modeled by a GMM. FV has also been used for gesture recognition. In [57], the authors combined the generative approach of Hidden Markov Model (HMM) dealing with spatio-temporal motion data with the discriminative approach of SVM for classification. In their approach motion segments are encoded into HMMs, and each segment is converted to FV; the SVM is subsequently trained on the FV.

Lately, based on FV and VLAD, deep learnable pooling techniques have been developed. FisherNet [58] is a differential approach of FV that is incorporated into a CNN network for features aggregation. Inspired by the great advantages of the deep learning model, Arandjelovic *et al.* [32] extended the traditional VLAD coding model to an end-to-end model called NetVLAD. They chose the outputs of the last convolutional layer of a deep CNN to feed a VLAD layer. The entire network trained all the parameters by the back propagation algorithm. NetVLAD introduces a soft assignment of each descriptor to a cluster. To this end, it makes use of Softmax and records the sum of the residuals in the same way as VLAD. The descriptors, once aggregated, are intra-column L2-normalized [59] then squeezed into a final vector that is afterwards L2-normalized. Extending NetVLAD for video-based action classification, Girdhar *et al.* [30] proposed ActionVLAD using NetVLAD for aggregating frame-level features however at different levels of the network.

Gating mechanisms have been actively applied in sequence models through gated RNNs. These mechanisms give RNNs the basis to allow information flows and so gradients do not vanish neither explode. Moreover, the gates, empower RNNs to focus attention towards the most representative components of the input cues. Different variants of gating mechanisms have been explored in RNNs, however most of them exploit the relevant, past and current, information at the present time step. More recently, Dauphin *et al.* [60] incorporated gating mechanisms in feed forward CNN and achieved competitive results on Language Modeling task. In addition, Squeeze-and-Excitation Networks [61] also look for attention mechanisms aggregating information from con-

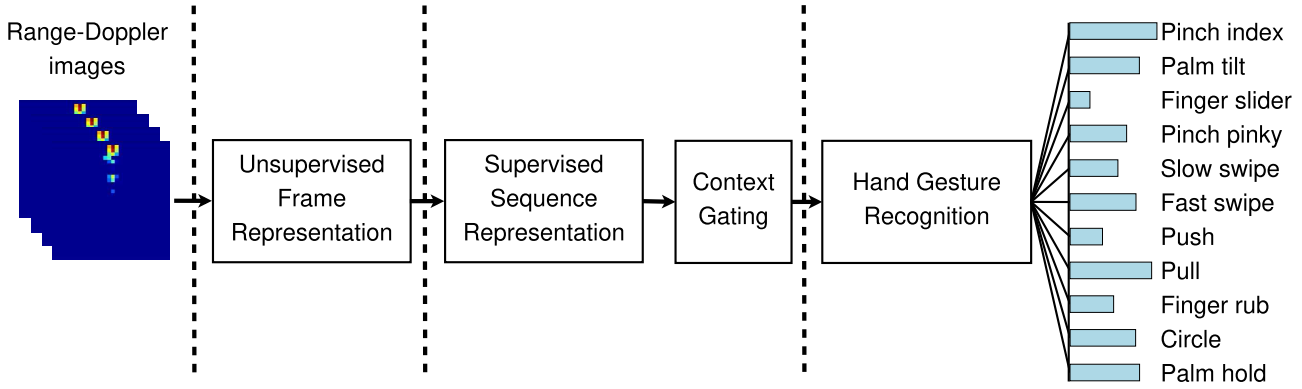


FIGURE 1. Schematic representation of the proposed framework. CNN features are extracted from input gestures composed by range-Doppler images. These features are temporally aggregated and the final representation is used to recognize the gestures.

volutional layers and capturing dependencies among feature channels through a gated mechanism.

Based on FisherNet and NetVLAD, Miech *et al.* [31] introduced a non-linear learnable network unit, named Context Gating (CG). CG aims at better capturing the non-linear inter-dependencies between features as well as among output labels. This approach overcomes the limitations of NetVLAD for creating relationships between descriptors. Moreover, NetVLAD with CG has shown good generalization capability even when trained with small number of samples [31]. In this work, we introduce this approach for Doppler-radar hand gesture recognition. We therefore, make use of NetVLAD combined with CG, to temporally pool frame-level features extracted from Doppler-radar images.

III. PROPOSED FRAMEWORK

A general overview of the proposed framework is given in Figure 1. It is composed by three modules. The first module, denoted as Unsupervised Frame Representation (UFR), is in charge of inferring features from range-Doppler images. The second module, Sequence Representation, aggregates in time the frame-level features to yield a robust representation of the gestures. Finally, a recognition module is used to classify each gesture.

A. UNSUPERVISED FRAME REPRESENTATION

Generally, representation learning aims at disentangling the underlying explanatory factors hidden in the observed data [38]. In the particular case of hand gesture recognition, variabilities such as timing and morphology complicate representation.

Deep CNNs have shown outstanding performance in several recognition tasks. Furthermore, it has been widely documented in the literature that Deep CNNs pretrained on huge datasets can achieve features representations that are intra-class invariant and retain generalization properties [62], [63]. The success of Deep CNNs suggests that their compositional and hierarchical structure induces increasingly invariant data representations by progressively flattening and separating the manifold-shape of the observed data [64]. However, when the CNN structures are not such deep it is required learning strategies to empower the model with capabilities for fine-grained recognition. Indeed, in this work we follow an unsupervised layer-wise learning approach that will ideally drives toward a model that can achieve sufficient representational power, by learning from data potentially subtle differences between very similar classes, whereas remains invariant to nuances within the same class.

The unsupervised representation learning problem can be posed as the optimization of the parameters of a CNN given an unlabelled dataset, subject to discovering the local properties of the data-generating distribution. We choose to discover such properties using an auto-encoder (AE) [38]. Previous experimental evidences show that when AEs are trained using a reconstruction contraction criterion, they learn local properties of the data-generating distribution [40], which is what we aim for. Among the different AEs variants, the reconstruction contractive auto-encoders (RCAEs) scheme has been proven to yield representations that capture the high-density regions of the data-generating distribution [40]. Nonetheless, most previous methods have been widely applied in vision recognition tasks where the input signal are not Doppler-radar images. For this reason, we decided to make use of the approach proposed by [43] which has been proved to be effective dealing with Doppler-radar images in classification tasks [44].

The primary goal of the solution presented in [43] is to transform the RCAE objective [40] into a convex optimization problem via: (i) *random convexification*, i.e. fixing the non-linear encoding filters randomly and only learning the (untied) linear decoding filters, (ii) *spectral minimization*, i.e. learning the decoding filters in the frequency domain. As a direct consequence, the main computational advantages are: (i) very few hyper-parameters to tune, and (ii) fast and guaranteed convergence. For the general case of multi-channel range-Doppler images, we consider a convolutional reconstruction function using 1 convolutional layer with K filters and input space $\mathcal{X} \subset \mathbb{R}^{H \times W \times C}$, i.e. the space of C-channel $H \times W$ images $\mathbf{x} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(C)}\right)$. We then define the

following per-channel convolutional reconstruction function:

$$\mathbf{r}^{(c)}(\mathbf{x}^{(c)};\boldsymbol{\theta}) \triangleq \underbrace{\sum_{k=1}^{K}\mathbf{w}^{(k,c)}}_{\text{linear decoding}} * \underbrace{\mathbf{g}(\mathbf{a}^{(k)} * \mathbf{x} + \mathbf{b}^{(k)})}_{\text{random nonlinear encoding}} = \sum_{k=1}^{K}\mathbf{w}^{(k,c)} * \mathbf{h}^{(k)} \quad (1)$$

with $\mathbf{g}(\cdot)$ denoting the element-wise application of an activation function $g : \mathbb{R} \rightarrow \mathbb{R}$. The model parameters $\boldsymbol{\theta}$ consist of encoding filters $\mathbf{a}^{(k)}$, encoding biases $\mathbf{b}^{(k)}$ and decoding filters $\mathbf{w}^{(k,c)}$. The method in [43] proposes to sample the (non-linear) encoding parameters $\mathbf{a}^{(k)}$ and $\mathbf{b}^{(k)}$ from predetermined density functions $p(\mathbf{a})$ and $p(\mathbf{b})$ respectively, and keep them fixed while learning the (linear) decoding parameters $\mathbf{w}^{(k,c)}$.

Using a complex-valued spectral (re)parametrization, the complex-valued decoding parameters associated to $\mathbf{w}^{(k,c)}$ are defined as $\mathbf{W}^{(k,c)}$, with $\mathbf{W}^{(k,c)} = \mathcal{F}\{\mathbf{w}^{(k,c)}\} \in \mathbb{C}^{H\times W}$, $\mathbf{H}^{(k)} = \mathcal{F}\{\mathbf{h}^{(k)}\}$, and $\mathbf{X} = \mathcal{F}\{\mathbf{x}\}$, where $\mathcal{F}$ is the discrete Fourier transform (DFT).

Then, following the Parseval's theorem along with the convolution theorem [65], the convolution operation is transformed into an element-wise multiplication. As a direct consequence, minimizing the spectral reconstruction error is reduced to solve $\mathbf{H} \odot \mathbf{W}$ independent K-dimensional complex-valued regularized linear least-squares problems, which are solved using the Gauss-Seidel (GS) algorithm [66], [67] (see [43] for a complete derivation of the GS iterations). Once the decoding filters are learned in the frequency domain, they are transformed back to the spatial domain using the inverse DFT: $\hat{\mathbf{w}}^{(k,c)} = \mathcal{F}^{-1}\{\mathbf{W}^{(k,c)}\}$.

Similar to the convolutional layer, the authors in [43] proposed a zero-bias fully connected (FC) auto-encoder. In the case of FC layer, the input space is considered to be $\mathcal{X} \subset \mathbb{R}^d$, i.e. the space of d-dimensional vectors. By fixing the (non-linear) encoding parameters $\{\mathbf{A}\}$ randomly, the (linear) decoding parameter $\boldsymbol{\theta} = \mathbf{W}$ are fitted optimally using a *convex minimization* strategy resulting in a linear least-squares minimization problem with Tikhonov regularization, which has a closed-from solution.

Finally, the feature representation for a given range-Doppler image $\mathbf{x}$ is obtained by:

$$\mathbf{f} = UFR(\mathbf{x};\boldsymbol{\theta}) \quad (2)$$

where $\mathbf{f}$ stands for the frame-level descriptor, given the learned hyper-parameters $\boldsymbol{\theta}$ of the convolutional and FC layers of the UFR module.

B. SUPERVISED SEQUENCE REPRESENTATION

In general, when dealing with sequences the intermediate representations obtained at frame-level are subsequently aggregated to exploit the temporal information from the input signal, which leads to a sequence representation. As stated above, most methods have relied on RNNs to model the dynamics of gestures. In this work, we cope with sequences

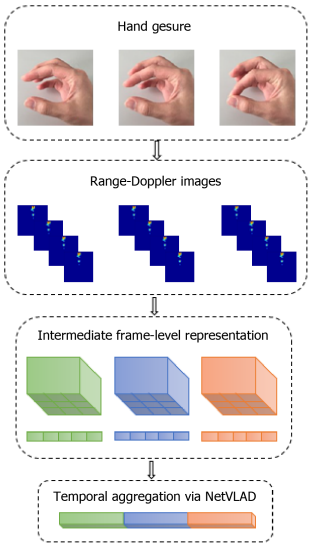


FIGURE 2. Frame-level descriptors are temporally pooled to attain the synthesized sequence representation.

of gestures that have different lengths, therefore it could drive our framework to model the dynamics using LSTMs as in previous studies [10], [19], [27]. However, loosely speaking there are two reasons why we do not select LSTMs in our framework. Firstly, the gestures used in this study are not such long sequences, therefore they have high short-term correlation across frames, and secondly LSTMs increase the computational demand, which is not cost-effective for ubiquitous deployment towards "real-time" recognition. Indeed, all that we require is modeling the dynamics of gestures and capture "relevant" cues from the intermediate frame-level representations to yield a synthesized sequence representation, whose will ideally allow to recognize gestures at high rates during the inference phase. Therefore, in this work we adopt NetVLAD [32] to capture the dynamics of Doppler-radar hand gestures across time.

Considering a sequence of T range-Doppler frames, and their corresponding N-dimensional frame-level descriptors $\mathbf{f}_t$ provided by the UFR module, we aggregate the descriptors using the NetVLAD [32] approach with M clusters (a parameter we can adjust as a trade-off between computation cost and performance). To this end, each frame-level descriptor is firstly encoded to be a feature matrix of $N \times M$ dimension using the following equation:

$$\mathbf{V}_t(j,m) = \alpha_m(\mathbf{f}_t)(\mathbf{f}_t(j) - \mathbf{c}_m(j)) \quad (3)$$

with $t \in \{1,\cdots,T\}$, $j \in \{1,\cdots,N\}$, $m \in \{1,\cdots,M\}$, $\mathbf{c}_m$ is the N-dimensional anchor point of cluster m and $\alpha_m(\mathbf{f}_t)$ is a soft assignment of $\mathbf{f}_t$ to cluster m (which measures the proximity of f_t to cluster m). The proximity function is modeled using a single FC layer with a Softmax activation:

$$\alpha_m(\mathbf{f}_t) = \frac{e^{\mathbf{W}_m^T\mathbf{f}_t+\mathbf{b}_m}}{\sum_{s=1}^{M} e^{\mathbf{W}_s^T\mathbf{f}_t+\mathbf{b}_s}} \quad (4)$$

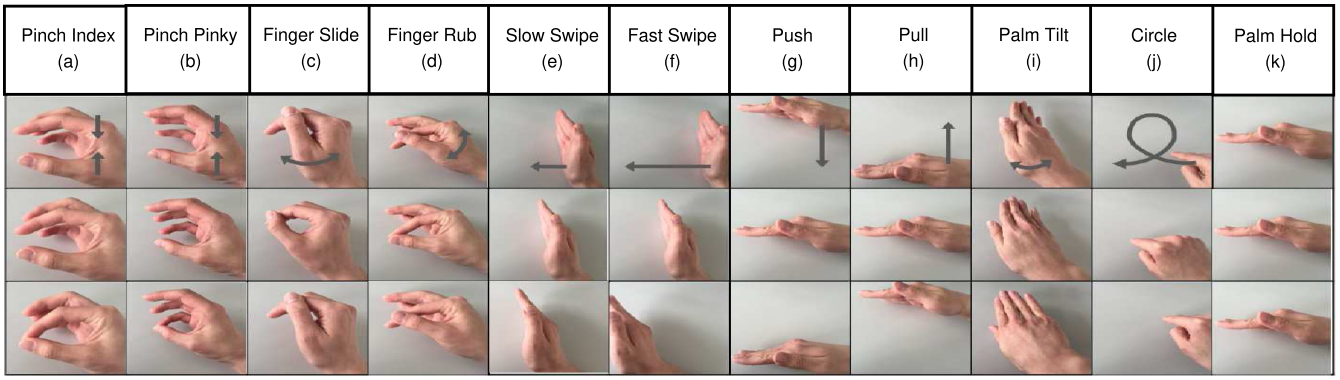


FIGURE 3. Portrayal of the set of gestures used in the experiments. Illustration taken from [10].

Secondly, a sequence-level descriptor $\mathbf{z}$ is obtained by column-wise intra-normalization of the feature matrix $\mathbf{V}_t$, followed by column-wise vectorization and L2-normalization.

In figure 2 we schematically illustrate the steps to yield the sequence representation.

C. CONTEXT GATING AND GESTURE RECOGNITION

We assume that the components of the vector $\mathbf{z}$ contain the representation of relevant information from each range-Doppler image belonging to hand gestures. Moreover, we aim to model inter-dependencies among these feature representations.

The Context Gating (CG), to recalibrate the relevance of each dimension in the sequence-level descriptor $\mathbf{z}$, obtained via NetVLAD, allows to capture inter-feature relationships from range-Doppler features as follows [31]:

$$\hat{\mathbf{z}} = g(\mathbf{W} \odot \mathbf{z} + \mathbf{b}) \tag{5}$$

here g represents a non linear activation function, $\mathbf{z}$ is the input vector, $\hat{\mathbf{z}}$ the transformed output vector, $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters.

Finally, the constructed sequence-level descriptor $\hat{\mathbf{z}}$ is passed to a Softmax layer to yield the gesture recognition probability score $\mathbf{y}$, formally referred to us as:

$$\mathbf{y} = Softmax(\hat{\mathbf{z}}) \tag{6}$$

IV. EXPERIMENTS

We performed qualitative and quantitative experiments to evaluate the effectiveness of the proposed framework in terms of representation learning, gesture recognition accuracy and speed. Furthermore, to compare with previous works [10] and validate the effectiveness of our core pipeline, we conducted experiments to explore alternative approaches using LSTMs in the Sequence Representation module to recognize gestures at frame and sequence-level.

A. DATASET

We make use of a publicly available[1] dataset, that was collected in the study of Wang *et al.* [10], the reader is referred to [10] for the data acquisition details. The dataset contains range-Doppler images of size 32×32 with 4 channels. These range-Doppler images correspond to 11 gestures recorded from 10 different subjects in 25 times, accounting for a total of 2750 gesture sequences. The considered gestures are illustrated in Figure 3. In addition, the authors provided a set of 2750 sequences from a single subject. In summary, the dataset has three benchmarks:

1) 50% − 50%. In this benchmark the 2750 sequences from 10 subjects were equally split into training and testing sets. We used the provided annotations to create the training and testing splits.
2) Leave one out cross subject evaluation on the 2750 sequences from 10 subjects. We used this benchmark to train our models on the data from 9 subjects and testing them on the subject out of the training set. This was done 10 times and the reported results measure the final average recognition accuracy over the 10 subjects.
3) Leave one out cross session evaluation on the 2750 sequences recorded from a single subject. This subset of hand gestures has 6 sessions recorded from a single user. It is used to evaluate the classification performances as a personalized classifier. Thus, we trained our models using the data from 5 sessions and tested on the remaining session.

B. IMPLEMENTATION DETAILS

We implemented the UFR module using MATLAB with MatConvNet [68]. The Supervised Sequence Representation module was implemented on top of TensorFlow [69]. Additionally, we made use of the Tensorflow Toolbox: Learnable mOdUle for Pooling fEatures (LOUPE) [31]. A graphics process unit, NVIDIA TITAN Xp, with 12 GB of RAM has been used for training and evaluating (unless indicated otherwise) our framework.

1) ARCHITECTURES

For all our experiments, we used the same CNN shallow architecture in the UFR module: consisting of one convolutional layer of 64 3×3 filters with the Rectified Linear Unit (ReLU) as activation function, followed by a FC layer

[1] https://github.com/simonwsw/deep-soli (accessed 2 January 2019)

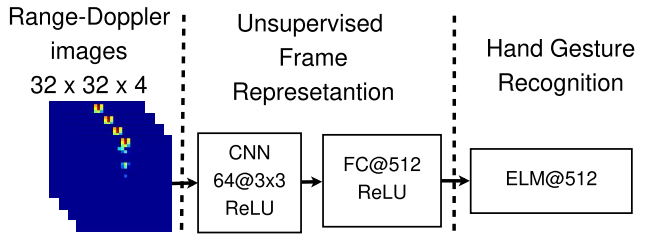


FIGURE 4. Illustration of the UFR-ELM pipeline implemented to validate the UFR module.

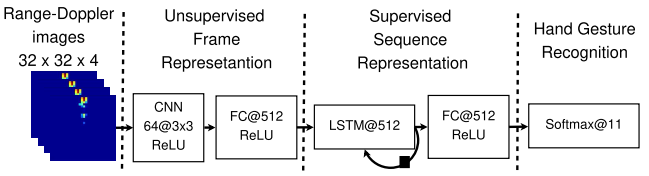


FIGURE 5. UFR-LSTM pipeline to asses the recognition rate and speed using recurrent neural networks in the supervised sequence representation module.

with 512 neurons and ReLU as activation function. Following the UFR methodology explained in section III, the encoding filters were randomly fixed by sampling their entries independently from a zero-mean normal distribution with standard deviation $1e-4$.

To validate the resulting frame-level descriptors $\mathbf{f}_t$, we considered a frame-based classification scheme using an Extreme Learning Machine (ELM) [70] with 512 neurons. This processing pipeline, denoted as UFR-ELM, is illustrated in Figure 4.

To compare to state-of-art CNN-LSTM approaches, in terms of frame-based and sequence-based accuracies, we considered a LSTM architecture on top of the UFR module, as illustrated in Figure 5. This processing pipeline is denoted as UFR-LSTM. The used LSTM has 512 recurrent neurons, followed by a FC layer with 512 neurons and as activation function the ReLU. In this pipeline Softmax was selected as classifier after the FC layer.

In order to asses the performance using short spatio-temporal features followed by global sequence representations, we explored the 3*D* CNN to obtain intermediate short spatio-temporal features along with LSTM as in [29] for the final sequence representation. Such pipeline is referred to as 3D-CNN-LSTM. It should be noted, however, that in this alternative approach the model is completely trained in a supervised manner since the UFR module has not been employed.

Figure 6 portrays the schematic representation of the 3D-CNN-LSTM. 3*D* tensors consiting of 3 frames (defined empirically) are input to the 3*D* convolutional layer of 64 $3 \times 3 \times 3$ filters with the Rectified Linear Unit (ReLU) as activation function, followed by a FC layer with 512 neurons and ReLU as activation function. The employed LSTM has 512 recurrent neurons, followed by a FC layer with 512 neurons and as activation function the ReLU. Finally, Softmax was used to classify the sequences after the FC layer.

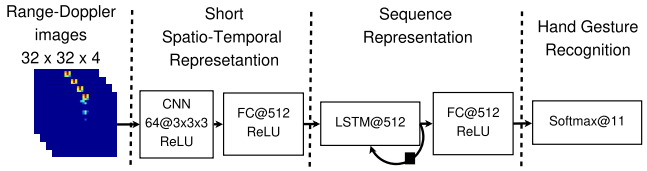


FIGURE 6. Illustration of the 3D-CNN-LSTM pipeline to asses the recognition rate and speed employing 3*D* convolutions and recurrent neural networks.

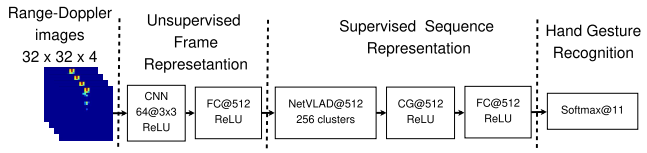


FIGURE 7. Details of the UFR-NetVLAD implementation to evaluate the recognition rate and speed using our proposal.

The proposed framework for Doppler-radar based gesture recognition, denoted as UFR-NetVLAD, is depicted in Figure 7. UFR has the same configuration as explained above and NetVLAD is employed with 256 clusters and 512 dimensional output descriptor $\mathbf{z}$, followed by Context Gating to obtain $\hat{\mathbf{z}}$. It must be noted that, different from prior works using CNN-LSTM [10], we did not sub-sample any sequence of the dataset, neither applied any preprocessing to the input frames.

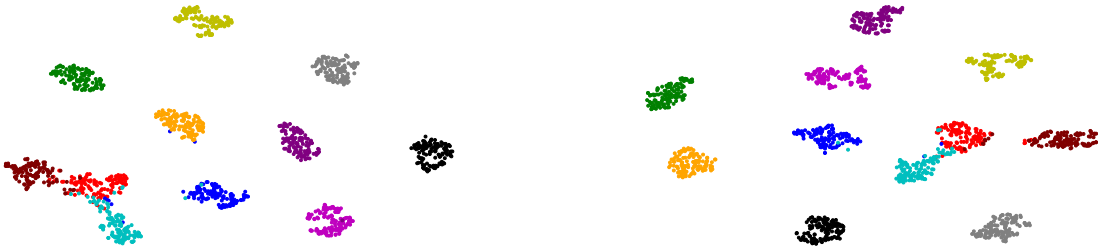
2) TRAINING

The different implemented architectures have a wide range of options for training. Thus, we explored different configurations aiming at validating our approach.

First, for all the experimental settings where the UFR module was used, it has been trained with the full training data as *unlabelled data*. To classify gestures at frame-level in UFR-ELM, the learned UFR hyper-parameters are used to infer the range-Doppler image features, which are then fed into the ELM model trained in a supervised fashion. For the proposed framework, UFR-NetVLAD, and its alternative design, UFR-LSTM, once the UFR was trained, two supervised training strategies were followed:

1) Freezing the hyper-parameters in lower layers to infer the features at UFR. Thus, training only the Supervised Sequence Representation module. This strategy is used to train the UFR-LSTM and UFR-NetVLAD models.
2) Fine-tuning the hyper-parameters in the lower layers, by pairing supervised learning with the unsupervised pretraining of UFR. We initialize the weights in lower layers with the unsupervised learned hyper-parameters from UFR and optimize all the learnable parameters in the CNN-LSTM and CNN-NetVLAD frameworks by training them in an end-to-end fashion. We denote the fine-tuned models as UFR-CNN-LSTM and UFR-CNN-NetVLAD receptively.

Second, for the alternative architectures, we have trained the network structures entirely in an end-to-end manner; to compare the performance by learning the networks



(a) UFR-NetVLAD (b) UFR-CNN-NetVLAD

FIGURE 8. t-SNE visualization for 50% evaluation split in the 50% – 50% split benchmark. (a) UFR-NetVLAD and (b) UFR-CNN-NetVLAD. Pinch index (red), Palm tilt (green), Finger slider (blue), Pinch pinky (cyan), Slow swipe (magenta), Fast swipe (yellow), Push (black), Pull (grey), Finger rub (orange), Circle (purple) and Palm hold (maroon).

hyper-parameters completely supervised, instead of training unsupervised the UFR module or fine-tuning it as exposed above. The models that have been completely trained end-to-end in a supervised manner are the CNN-LSTM and 3D-CNN-LSTM.

During training we required 1 epoch to optimize the UFR, we used 200 cycles for Gauss-Seidel optimization and the regularization parameter was set to $1e-4$. The ELM model also required just 1 epoch. For training the UFR-LSTM and UFR-CNN-LSTM, we used 200 epochs with Adam and a learning rate of $1e-3$. For UFR-NetVLAD and UFR-CNN-NetVLAD, we just changed the number of epochs to 50. Similarly, for the end-to-end models, namely the CNN-LSTM and 3D-CNN-LSTM, we used Adam with a learning rate of $1e-3$ and 200 epochs.

C. QUALITATIVE RESULTS

To qualitatively assess the learned representations of the considered target gestures, we use a t-SNE plot. Figure 8 depicts the two-dimensional t-SNE embedding from the FC layer before Softmax classifier in UFR-NetVLAD and UFR-CNN-NetVLAD, respectively. As can be seen the features belonging to different gestures are well separated; the natural clusters formed during t-SNE embedding are an indication of the obtained discriminative representations. For most of gestures, their instances are grouped, and the formed clusters are clearly isolated, which suggests invariant and discriminative representation properties.

In the case of UFR-NetVLAD, Pinch index (red), Pinch pinky (cyan) and Palm hold (maroon) have instances closely distributed and sometimes mixed. Pairing supervised learning with the unsupervised pretraining, using UFR-CNN-NetVLAD, allowed further separating the Palm hold cluster from Pinch index and Pinch pinky gestures. The latter gestures stay close to each other. Indeed, from Figure 3 one can notice that it is difficult to discriminate between them if we do not track the fingers.

D. QUANTITATIVE RESULTS

In the following, the outcomes of our quantitative experimentation are compared to the accuracy scores reported in [10].

The results of our tests on the 50%-50% split benchmark are summarized in Table 1 for both, frame-based and sequence-based recognition. Our proposed framework using NetVLAD is very competitive compared to CNN-LSTM based state-of-art approaches. This benchmark nearly resembles real world conditions when a model is deployed for its exploitation, as it should recognize gestures from different potential users.

Considering the frame-based classification, the UFR-ELM approach outperforms CNN-based approaches reported in [10]. Moreover, one can notice that, adding a LSTM layer on top of CNN (CNN-LSTM, UFR-LSTM and UFR-CNN-LSTM) improves the recognition accuracy of frame-level classification. This demonstrates the importance in capturing dependencies among frame representations to get more attributes and synthetic descriptions of the gestures. These findings are in accordance with the works presented in [10] and [19].

With respect to sequence-based classification, our proposed framework, combining the unsupervised features learning with NetVLAD and CG outperforms state-of-art Convolutional Long Short-Term Memory Networks. Thus, consistent with our assumptions about exploiting the progressive and time-sequential nature of the gestures by capturing dependencies among features at frame-level. We also notice that combining $3D$ CNN with LSTM improves the recognition accuracy compared to CNN-LSTM, this is due to the fact that $3D$ CNN captures spatio-temporal features before feeding LSTM. Regarding, UFR-NetVLAD and UFR-CNN-NetVLAD, better results are observable with UFR-CNN-NetVLAD due to the fine-tuning of the convolutional and FC layers. Moreover, the CG after NetVLAD increases more than 2% and 1% the recognition scores for UFR-NetVLAD and UFR-CNN-NetVLAD respectively. In particular, UFR-NetVLAD clearly boosts the performance to recognize Pinch index when using CG.

Analyzing the recognition scores of the considered gestures, one can notice that UFR-NetVLAD and UFR-CNN-NetVLAD recognize well gestures with very similar characteristics. The gestures Slow swipe and Fast swipe, which differ only in the velocity of motions, as well as Push and

TABLE 1. Accuracy on 50% – 50% split of data for training and evaluation.

Frame-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
CNN-Shallow [10]	41.13%	20.00%	34.40%	1.20%	46.80%	37.20%	97.60%	29.60%	22.00%	54.00%	26.4%	83.20%
CNN-Deep [10]	48.18%	47.60%	34.40%	6.80%	39.60%	43.20%	98.80%	56.00%	31.20%	52.80%	48.40%	71.20%
UFR-ELM	64.10%	49.92%	75.39%	49.47%	42.66%	60.22%	80.64%	58.08%	74.57%	65.24%	68.16%	80.67%
CNN-LSTM	86.41%	69.90%	94.04%	83.33%	56.05%	95.65%	98.85%	99.01%	91.89%	93.78%	98.41%	69.60%
CNN-LSTM [10]	87.17%	67.72%	71.09%	77.78%	94.48%	84.84%	98.45%	98.63%	88.89%	94.85%	89.56%	92.63%
UFR-LSTM	90.23%	89.54%	93.70%	79.83%	68.62%	96.91%	98.70%	98.14%	95.03%	92.52%	97.55%	82.09%
UFR-CNN-LSTM	**91.06%**	82.75%	94.78%	81.21%	79.11%	96.34%	99.82%	98.24%	94.41%	93.09%	99.15%	82.80%
Sequence-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
UFR-NetVLAD without CG	95.45%	67.74%	99.19%	97.58%	94.35%	100.00%	100.00%	100.00%	100.00%	100.00%	98.38%	92.74%
UFR-CNN-NetVLAD without CG	96.77%	97.58%	100.00%	98.38%	83.06%	100.00%	100.00%	99.19%	99.19%	96.77%	98.38%	91.93%
CNN-LSTM	97.70%	81.12%	100.00%	100.00%	99.32%	99.19%	100.00%	99.19%	100.00%	100.00%	99.19%	96.77%
3D-CNN-LSTM	97.80%	94.35%	100.00%	100.00%	87.09%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	94.35%
UFR-NetVLAD	97.87%	87.09%	100.00%	96.77%	96.77%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	95.96%
UFR-LSTM	98.02%	97.58%	100.00%	97.58%	92.74%	100.00%	100.00%	100.00%	100.00%	99.19%	97.58%	93.54%
UFR-CNN-LSTM	98.16%	95.16%	100.00%	99.19%	91.93%	99.19%	99.19%	100.00%	100.00%	100.00%	100.00%	95.16%
UFR-CNN-NetVLAD	**98.24%**	91.12%	99.19%	99.19%	95.96%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	95.16%

TABLE 2. Accuracy on leave one subject out, cross subject evaluation on 10 participants.

Frame-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
CNN-LSTM	67.21%	48.22%	92.94%	46.99%	40.63%	68.49%	51.89%	91.88%	90.90%	85.77%	80.35%	41.23%
UFR-LSTM	77.18%	61.80%	93.59%	60.23%	62.41%	74.67%	62.42%	94.77%	90.08%	89.89%	82.10%	76.99%
UFR-CNN-LSTM	78.00%	65.00%	93.37%	62.50%	62.78%	76.95%	60.18%	95.98%	90.46%	90.51%	85.90%	74.41%
CNN-LSTM [10]	**79.06%**	58.71%	67.62%	64.80%	91.82%	72.31%	72.91%	93.40%	89.99%	95.16%	82.80%	80.24%
Sequence-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
UFR-LSTM	87.52%	77.60 %	99.20%	77.60%	76.40%	79.60%	87.60%	99.20%	100.00%	87.60%	88.00%	90.00%
CNN-LSTM	87.69%	74.80%	85.00%	82.40%	78.00%	81.60%	93.20%	95.60%	100.00%	88.40%	94.40%	91.20%
CNN-LSTM [10]	88.27%	70.80%	76.80%	83.20%	97.20%	80.40%	83.60%	94.80%	100.00%	100.00%	97.20%	86.82%
UFR-CNN-LSTM	90.32%	80.00%	98.80%	83.60%	79.20%	79.20%	96.39%	99.20%	100.00%	91.20%	93.20%	92.80%
3D-CNN-LSTM	90.94%	74.40%	93.20%	91.20%	74.80%	98.00%	92.39%	100.00%	100.00%	88.40%	100.00%	88.00%
UFR-NetVLAD	90.94%	79.60%	98.40%	82.00%	76.80%	93.60%	91.99%	98.00%	99.20%	88.80%	97.60%	94.40%
UFR-CNN-NetVLAD	**91.38%**	84.80%	98.40%	88.00%	78.40%	87.60%	99.20%	90.00%	99.20%	96.40%	93.99%	89.20%

Pull, which just differ in their directions are recognized accurately. On the other hand, Pinch index, Pinch pinky and Palm hold are recognized with lower accuracy, however still distinguishable.

Table 2 summarizes the obtained results for the leave one out cross subject evaluation (using the 2750 sequences from 10 subjects). Comparing the reported accuracy scores to the ones of Table 1, we notice a decrease, likewise reported in [10]. Here also, the combination UFR, NetVLAD and CG provides competitive results compared to state-of-art. Regarding the accuracy scores per gestures, Pinch index and Pinch pinky continue being difficult to classify.

Considering the single subject scenario, the recognition results in the leave-one-out evaluation are summarized in Table 3. The overall sequence-based accuracy is drastically improved compared to Table 2, clearly supporting our approach of combining an unsupervised representation learning step before fine-tuning and NetVLAD. On the other hand, similar to the previous indicative results, it can be noted that Pinch gestures have lower recognition scores.

All the results reported in Table 1 to Table 3 have been obtained by training UFR in an unsupervised fashion using all the training data. To further evaluate the effect of the unsupervised learning, we conducted experiments using a reduced set of training data to perform representation learning using UFR. For these experiments we used the 50% – 50% benchmark, from the 50% training data we randomly selected 25% for training UFR, and the remaining 25% for the Supervised Sequence Representation models (UFR-LSTM and UFR-NetVLAD), as well as fine-tuning the CNN for the UFR-CNN-LSTM and UFR-CNN-NetVLAD models. The 50% testing data has been employed to assess the learned models. Table 4 provides the obtained accuracies. One can notice that our results are rather accurate, even with reduced training data for both UFR and the Supervised Sequence Representation models. Indeed, compared to [10]

TABLE 3. Accuracy on leave one session out, cross session evaluation on single subject.

Frame-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
CNN-LSTM	81.23%	60.82%	96.34%	65.91%	54.52%	87.25%	91.08%	98.15%	96.23%	94.62%	92.85%	55.75%
UFR-CNN-LSTM	84.01%	59.58%	96.63%	73.51%	63.52%	86.74%	84.55%	99.06%	96.12%	93.66%	91.00%	79.84%
UFR-LSTM	83.97%	58.31%	96.13%	75.00%	66.13%	85.71%	83.96%	98.64%	95.58%	93.94%	91.51%	78.79%
UFR-CNN-LSTM	84.01%	59.58%	96.63%	73.51%	63.52%	86.74%	84.55%	99.06%	96.12%	93.66%	91.00%	79.84%
CNN-LSTM [10]	**85.75%**	56.69%	61.98%	76.43%	96.83%	92.73%	81.38%	98.42%	97.79%	95.33%	96.92%	89.10%
Sequence-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
CNN-LSTM [10]	94.15%	79.20%	74.40%	95.60%	100.00%	97.60%	94.80%	100.00%	100.00%	100.00%	100.00%	94.09%
CNN-LSTM	95.38%	89.20%	100.00%	95.20%	85.20%	96.8%	98.40%	100.00%	99.20%	96.40%	99.60%	89.20%
UFR-LSTM	96.75%	96.00%	100.00%	96.33%	81.00%	97.66%	98.66%	100.00%	100.00%	97.33%	100.00%	97.33%
3D-CNN-LSTM	96.78%	89.00%	100.00%	95.66%	85.99%	100.00%	100.00%	100.00%	100.00%	99.66%	100.00%	94.33%
UFR-NetVLAD	97.15%	95.00%	99.00%	96.66%	89.33%	97.66%	100.00%	100.00%	99.33%	100.00%	98.66%	93.00%
UFR-CNN-LSTM	97.30%	95.66%	100.00%	97.66%	85.99%	98.00%	99.66%	100.00%	99.66%	98.33%	100.00%	95.33%
UFR-CNN-NetVLAD	**97.75%**	94.33%	99.33%	97.66%	90.33%	97.66%	100.00%	100.00%	99.66%	100.00%	99.66%	96.66%

TABLE 4. Accuracy on the 50% – 50% benchmark, using 25% training data for UFR, 25% for training the supervised sequence representation and 50% for testing.

Frame-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
CNN-LSTM	83.31%	51.93%	93.59%	80.88%	54.40%	93.98%	98.20%	94.85%	90.68%	90.69%	95.50%	71.70%
UFR-CNN-LSTM	86.79%	75.64%	94.28%	76.83%	64.69%	94.44%	98.88%	96.55%	90.13%	93.26%	95.64%	74.42%
UFR-LSTM	**87.69%**	78.72%	93.99%	76.54%	68.53%	94.27%	98.92%	98.31%	91.42%	91.78%	95.10%	77.06%
Sequence-level recognition	Ave. Acc.	Pinch index (a)	Palm tild (b)	Finger slider (d)	Pinch pinky (d)	Slow swipe (e)	Fast swipe (f)	Push (g)	Pull (h)	Finger rub (i)	Circle (j)	Palm hold (k)
CNN-LSTM	95.60%	79.03%	100.00%	100.00%	85.48%	100.00%	99.19%	99.19%	100.00%	100.00%	98.38%	90.32%
UFR-LSTM	96.04%	92.74%	100.00%	95.96%	83.87%	100.00%	100.00%	100.00%	100.00%	100.00%	99.19%	84.67%
UFR-CNN-LSTM	96.04%	87.09%	100.00%	95.16%	88.70%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	85.48%
3D-CNN-LSTM	96.48%	81.45%	100.00%	100.00%	89.51%	100.00%	100.00%	100.00%	100.00%	100.00%	99.19%	91.12%
UFR-CNN-NetVLAD	96.55%	86.29%	100.00%	91.93%	91.12%	100.00%	100.00%	99.19%	100.00%	100.00%	100.00%	93.54%
UFR-NetVLAD	**96.92%**	88.70%	100.00%	95.16%	91.93%	100.00%	100.00%	99.19%	100.00%	100.00%	99.19%	91.93%

for the 50% – 50% scenario of Table 1, it is noticeable that, despite reduced training data, the shallow proposals UFR-LSTM and UFR-CNN-LSTM still can achieve competitive recognition scores for frame-level recognition. Moreover, for sequence-level recognition we also observe similar trends. This is due to the proposed unsupervised learning strategy that learns representations which are robust to variations. Furthermore, for sequence-based recognition, the UFR-CNN-NetVLAD and UFR-NetVLAD yield the best average accuracy values, slightly outperforming 3D-CNN-LSTM.

Regarding the recognition scores per gesture, we can notice that Pinch index and Pinch picky gestures are the most difficult to recognize, as in previous experimental scenarios. However, for the other gestures used in the experiments, the proposed framework can achieve good sequence-based recognition scores. These outputs confirm that the UFR enforce invariance, even when it is trained with small amount of data. Moreover, these results corroborate that combining UFR with the NetVLAD approach, to aggregate frame-level features, leads to quite accurate recognition of tiny hand gestures from Doppler-radar images. Furthermore, the outcomes shown in Table 4 illustrate the fact that the proposed shallow architecture can generalize well with reduced number of samples.

Since the final goal of gesture recognition is to recognize, in pseudo real time, gestures and translate them into system commands to perform specific tasks. We therefore assessed our framework in terms of recognition speed, measured as the time required, during inference, to recognize a gesture giving a full sequence with 62 frames in average. Table 5 reports the measured inference time. One can notice that approaches based on CNN-LSTM are slower. On the other hand, the UFR-NetVLAD and UFR-CNN-NetVLAD architectures required 4.8 *ms*, and 4.9 *ms* receptively to recognize a sequence. These results represent a gesture recognition rate of 208 *Hz* and 204 *Hz* for UFR-NetVLAD and UFR-CNN-NetVLAD. Thus, very competitive compared to the 150*Hz* frame rate recognition reported in [10].

Finally, we evaluated our framework in commodity CPU hardware, an Intel $i7 - 4800$ at 2.70*Ghz*. In such hardware,

TABLE 5. Time and recognition rate required for the different implemented architectures to recognize gestures.

Solution	Time (in ms, lower is better)	Recognition rate (in Hz, higher is better)
CNN-LSTM	5.7	175
3D-CNN-LSTM	5.1	196
UFR-LSTM	5.8	172
UFR-CNN-LSTM	5.6	178
UFR-CNN-NetVLAD	4.9	204
UFR-NetVLAD	**4.8**	**208**

our proposal achieved gesture recognition at a rate of 28 *Hz* and 26 *Hz* for UFR-NetVLAD and UFR-CNN-NetVLAD respectively; which corroborate its potential for ubiquitous deployment.

V. CONCLUSION AND DISCUSSION

In this paper, we presented a framework to recognize hand gestures attributes combining Unsupervised Frame Representation and Supervised Sequence Representation. Our proposed framework is based on (i) an unsupervised representation learning strategy for greedy layer-wise unsupervised training of CNNs to extract range-Doppler features at frame-level, followed by (ii) the NetVLAD learnable pooling aggregation technique, that captures compact invariant synthesis to recognize patterns in tiniest motion gestures at sequence-level. Comprehensive experiments on publicly available data proved the effectiveness of our approach in terms of representation task, recognition accuracy and speed. Furthermore, the spectrum of design alternatives illustrated in the experimental section, shows the capability of the framework to be expanded to different sequence aggregation methods.

The proposed shallow CNN architecture (UFR) could be directly used to generate feature representations from images or fine-tuning the hyper-parameters in the lower layers of a CNN architecture, by pairing supervised learning with the unsupervised pretraining of UFR. UFR not only yields features robust to noise but also discovers relevant factors in the input frames, whereas it enforces invariance to slightly differences caused by variability among subjects; which leads to a representation that could be directly used for the recognition task or fine-tuned in supervised manner. UFR combined with NetVLAD, as well as with LSTM, provide good recognition results compared to the deeper CNN architecture used in [10]. Moreover, compared to 3*D* CNN-LSTM architectures, our framework delivers better recognition accuracy. Additionally, in contrast to 3*D* CNNs our UFR model requires less parameters to train, as in our model we rely on NetVLAD to learn the interframe temporal information. Indeed, 3*D* CNNs compute feature maps from both spatial and temporal dimensions and have exhibited promising spatio-temporal feature learning ability, however, the number of model parameters (depth and width of the network) and therefore the number of required examples increase with the temporal depth of input sequences, limiting their use to short-term sequences [28], [29].

In the future, we plan to move towards an adaptive framework able to personalize in an online manner the models, considering different subjects as well as unseen gestures. Moreover, we will investigate how to incorporate recent learnable aggregation approaches which attempt increasing attention [78] and dealing with unseen data [71], [72].

REFERENCES

[1] M. J. Cheok, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 1, pp. 131–153, Jan. 2019.

[2] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev,"Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Trans. Graph.*, vol. 35, no. 4, p. 142, 2016.

[3] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks," *IEEE Sensors J.*, vol. 19, no. 8, pp. 3041–3048, Apr. 2019.

[4] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–8.

[5] Z. Zhou, Z. Cao, and Y. Pi, "Dynamic gesture recognition with a terahertz radar based on range profile sequences and Doppler signatures," *Sensors*, vol. 18, no. 1, p. 10, Dec. 2017.

[6] Y. Sun, T. Fei, F. Schliep, and N. Pohl, "Gesture classification with handcrafted micro-Doppler features using a FMCW radar," in *Proc. IEEE MTT-S Int. Conf. Microw. Intell. Mobility (ICMIM)*, Apr. 2018, pp. 1–4.

[7] G. Li, R. Zhang, M. Ritchie, and H. Griffiths, "Sparsity-based dynamic hand gesture recognition using micro-Doppler signatures," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2017, pp. 928–931.

[8] F. Khan, S. K. Leem, and S. H. Cho, "Hand-based gesture recognition for vehicular applications using IR-UWB radar," *Sensors*, vol. 17, no. 4, p. 833, Apr. 2017.

[9] T. Sakamoto, X. Gao, E. Yavari, A. Rahman, O. Boric-Lubecke, and V. M. Lubecke, "Radar-based hand gesture recognition using I-Q echo plot and convolutional neural network," in *Proc. IEEE Conf. Antenna Meas. Appl. (CAMA)*, Dec. 2017, pp. 393–395.

[10] 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 Softw. Technol.*, 2016, pp. 851–860.

[11] F. Bernardo, N. Arner, and P. Batchelor, "O soli mio: Exploring millimeter wave radar for musical interaction," in *Proc. 17th Int. Conf. New Interfaces Musical Expression (NIME)*. Aalborg, Denmark: Aalborg Univ., May 2017, pp. 283–286. [Online]. Available: http://www.nime.org/proceedings/2017/nime2017_paper0054.pdf

[12] G. Rigoll, A. Kosmala, and S. Eickeler, "High performance real-time gesture recognition using hidden Markov models," in *Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and M. Fröhlich, Eds. Berlin, Germany: Springer, 1998, pp. 69–80.

[13] T. Darrell and A. Pentland, "Space-time gestures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 1993, pp. 335–340.

[14] N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. W. Taylor, and F. Nebout, "A multi-scale approach to gesture detection and recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Jun. 2013, pp. 484–491.

[15] D. Jirak, P. Barros, and S. Wermter, "Dynamic gesture recognition using echo state networks," in *Proc. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn. (ESANN)*, vol. 475. Louvain-la-Neuve, Belgium: Presses Univ. Louvain, 2015, chs. 591–596.

[16] E. Tsironi, P. Barros, and S. Wermter, "Gesture recognition with a convolutional long short-term memory recurrent neural network," in *Proc. ESANN*, 2016, pp. 1–6.

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

[18] J. Zhang, J. Tao, and Z. Shi, "Doppler-radar based hand gesture recognition system using convolutional neural networks," in *Proc. Int. Conf. Commun., Signal Process., Syst.*, 2017, pp. 1096–1113.

[19] A. Oulasvirta, X. Bi, and A. Howes, *Computational Interaction*. London, U.K.: Oxford Univ. Press, 2018.

[20] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[21] K. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun, "Learning convolutional feature hierarchies for visual recognition," in *Advances in Neural Information Processing Systems*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Red Hook, NY, USA: Curran Associates, 2010, pp. 1090–1098. [Online]. Available: http://papers.nips.cc/paper/4133-learning-convolutional-feature-hierarchies-for-visual-recognition.pdf

[22] S. Rifai, P. Vincent, X. Müller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. 28th Int. Conf. Mach. Learn.* Madison, WI, USA: Omnipress, 2011, pp. 833–840.

[23] S. Rifai, G. Mesnil, P. Vincent, X. Müller, Y. Bengio, Y. Dauphin, and X. Glorot, "Higher order contractive auto-encoder," in *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer, 2011, pp. 645–660.

[24] K. Sohn and H. Lee, "Learning invariant representations with local transformations," in *Proc. 29th Int. Conf. Mach. Learn. (ICML)*. Madison, WI, USA: Omnipress, 2012, pp. 1339–1346. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042573.3042745

[25] F. Anselmi, L. Rosasco, and T. Poggio, "On invariance and selectivity in representation learning," *Inf. Inference, J. IMA*, vol. 5, no. 2, pp. 134–158, 2016.

[26] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, Sep. 2016.

[27] J.-W. Choi, S.-J. Ryu, and J.-H. Kim, "Short-range radar based real-time hand gesture recognition using LSTM encoder," *IEEE Access*, vol. 7, pp. 33610–33618, Mar. 2019.

[28] 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, Feb. 2018.

[29] Y. Wang, S. Wang, M. Zhou, Q. Jiang, and Z. Tian, "TS-I3D based hand gesture recognition method with radar sensor," *IEEE Access*, vol. 7, pp. 22902–22913, 2019.

[30] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "ActionVLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3165–3174.

[31] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," 2017, *arXiv:1706.06905*. [Online]. Available: https://arxiv.org/abs/1706.06905

[32] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 5297–5307.

[33] K. A. Smith, C. Csech, D. Murdoch, and G. Shaker, "Gesture recognition using mm-Wave sensor for human-car interface," *IEEE Sens. Lett.*, vol. 2, no. 2, Jun. 2018, Art. no. 3500904.

[34] Q. Wan, Y. Li, C. Li, and R. Pal, "Gesture recognition for smart home applications using portable radar sensors," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 6414–6417.

[35] 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.

[36] G. Li, S. Zhang, F. Fioranelli, and H. Griffiths, "Effect of sparsity-aware time–frequency analysis on dynamic hand gesture classification with radar micro-Doppler signatures," *IET Radar, Sonar Navigat.*, vol. 12, no. 8, pp. 815–820, 2018.

[37] F. Luo, S. Poslad, and E. Bodanese, "Human activity detection and coarse localization outdoors using micro-Doppler signatures," *IEEE Sensors J.*, vol. 19, no. 18, pp. 8079–8094, Sep. 2019.

[38] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[39] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 899–907.

[40] G. Alain and Y. Bengio, "What regularized auto-encoders learn from the data-generating distribution," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3563–3593, 2014.

[41] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning—ICANN*. Berlin, Germany: Springer, 2011, pp. 52–59.

[42] M. S. Seyfioğlu, A. M. Özbayoglu, and S. Z. Gürbüz, "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.

[43] M. C. Oveneke, M. Aliosha-Perez, Y. Zhao, D. Jiang, and H. Sahli, "Efficient convolutional auto-encoding via random convexification and frequency-domain minimization," in *Proc. NIPS Int. Workshop Efficient Methods Deep Neural Netw. (EMDNN)*, 2016, pp. 1–6.

[44] K. N. Parashar, M. C. Oveneke, M. Rykunov, H. Sahli, and A. Bourdoux, "Micro-Doppler feature extraction using convolutional auto-encoders for low latency target classification," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2017, pp. 1739–1744.

[45] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 265–272.

[46] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*. [Online]. Available: https://arxiv.org/abs/1609.08675

[47] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "RNN Fisher vectors for action recognition and image annotation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 833–850.

[48] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6299–6308.

[49] M. Maghoumi and J. J. LaViola, Jr., "DeepGRU: Deep gesture recognition utility," 2018, *arXiv:1810.12514*. [Online]. Available: https://arxiv.org/abs/1810.12514

[50] C. Li, C. Xie, B. Zhang, C. Chen, and J. Han, "Deep Fisher discriminant learning for mobile hand gesture recognition," *Pattern Recognit.*, vol. 77, pp. 276–288, May 2018.

[51] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 2, Oct. 2003, pp. 1470–1477.

[52] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.

[53] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[54] X. Peng, L. Wang, Z. Cai, Y. Qiao, and Q. Peng, "Hybrid super vector with improved dense trajectories for action recognition," in *Proc. ICCV Workshops*, vol. 13, 2013, pp. 109–125.

[55] M. Benmoussa and A. Mahmoudi, "Machine learning for hand gesture recognition using bag-of-words," in *Proc. Int. Conf. Intell. Syst. Comput. Vis. (ISCV)*, Apr. 2018, pp. 1–7.

[56] J. Rodríguez and F. Martínez, "A kinematic gesture representation based on shape difference VLAD for sign language recognition," in *Proc. Int. Conf. Comput. Vis. Graph.* Cham, Switzerland: Springer, 2018, pp. 438–449.

[57] Y. Goutsu, W. Takano, and Y. Nakamura, "Gesture recognition using hybrid generative-discriminative approach with Fisher vector," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2015, pp. 3024–3031.

[58] P. Tang, X. Wang, B. Shi, X. Bai, W. Liu, and Z. Tu, "Deep FisherNet for object classification," 2016, *arXiv:1608.00182*. [Online]. Available: https://arxiv.org/abs/1608.00182

[59] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2013, pp. 1578–1585.

[60] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 933–941. [Online]. Available: http://dl.acm.org/citation.cfm?id=3305381.3305478

[61] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7132–7141.

[62] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 647–655.

[63] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.
[64] P. P. Brahma, D. Wu, and Y. She, "Why deep learning works: A manifold disentanglement perspective," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 1997–2008, Oct. 2016.
[65] D. W. Kammler, *A First Course in Fourier Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[66] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: JHU Press, 2012, vol. 3.
[67] A. Hefny, D. Needell, and A. Ramdas, "Rows vs. columns: Randomized Kaczmarz or gauss-seidel for ridge regression," 2015, *arXiv:1507.05844*. [Online]. Available: https://arxiv.org/abs/1507.05844
[68] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[69] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf
[70] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.
[71] Y. Zhong, R. Arandjelović, and A. Zisserman, "GhostVLAD for set-based face recognition," 2018, *arXiv:1810.09951*. [Online]. Available: https://arxiv.org/abs/1810.09951
[72] S. Kmiec, J. Bae, and R. An, "Learnable pooling methods for video classification," 2018, *arXiv:1810.00530*. [Online]. Available: https://arxiv.org/abs/1810.00530

ABEL DÍAZ BERENGUER received the B.Sc.Eng. degree in informatics sciences and the M.Sc.Eng. degree in applied informatics from the University of Informatics Sciences (UCI), Havana, Cuba, in 2009 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Joint VUB-NPU Audio-Visual Signal Processing (AVSP) Laboratory, Electronics and Informatics (ETRO) Department, Vrije Universiteit Brussel (VUB), under the supervision of Prof. H. Sahli. His current research interests include automatic human social behavior analysis, pedestrian detection, visual tracking, smart video surveillance, audio–visual signal processing, and machine learning.

MESHIA CÉDRIC OVENEKE received the B.Sc.Eng. degree in electronics and information technology and the M.Sc.Eng. degree (Hons.) in computer science (artificial intelligence) from the Vrije Universiteit Brussel (VUB), Brussels, Belgium, in 2011 and 2013, respectively, and the Ph.D. degree in engineering sciences from the Joint VUB-NPU Audio-Visual Signal Processing (AVSP) Laboratory, Electronics and Informatics (ETRO) Department, VUB, in 2018. He is currently a Teaching Assistant in computer vision. His current research interests include developing machine learning-based approaches to audio–visual signal processing, applied to affective computing and behavior analysis, image, video, and speech processing, large-scale optimization, machine learning, and the Bayesian estimation.

HABIB-UR-REHMAN KHALID received the B.Sc. degree in electrical engineering from The University of Lahore (UoL), Pakistan, in 2013, and the master's degree in electrical and electronics engineering from the Katholieke Universiteit Leuven (KUL), Belgium, in 2018. He is currently pursuing the Ph.D. degree in engineering sciences with the Vrije Universiteit Brussel (VUB) in collaboration with the Interuniversitair Micro-Elektronica Centrum (IMEC) under the supervision of Prof. H. Sahli. His main research interests include digital signal processing, real-time embedded and control systems, multi-sensor data fusion, energy-efficient machine learning, and heterogeneous sensor-based human activity recognition.

MITCHEL ALIOSCHA-PEREZ received the B.Sc. degree *(summa cum laude)* in computer science from the Central University of Las Villas (UCLV), Santa Clara, Cuba, the M.Sc.Eng. degree in systems and signals with a focus on digital signals and images processing from UCLV, in 2011, and the Ph.D. degree in engineering sciences from the Joint VUB-NPU Audio-Visual Signal Processing (AVSP) Laboratory, Electronics and Informatics (ETRO) Department, Vrije Universiteit Brussel, Brussels, Belgium, in 2018. His current work involves large-scale computer vision and machine learning problems, with applications to computational biology and audio–visual signals processing. His current research interests include image and video analysis, large-scale machine learning, variational models, and non-convex optimization.

ANDRÉ BOURDOUX received the M.Sc. degree in electrical engineering from the Université Catholique de Louvain-la-Neuve, Belgium, in 1982. In 1998, he joined IMEC, where he is currently a Principal Member of Technical Staff with the Internet-of-Things Research Group. He is a system-level and signal processing expert for both the mm-wave wireless communications and radar teams. He has more than 15 years of research experience in radar systems and 15 years of research experience in broadband wireless communications. He holds several patents in these fields. He has authored or coauthored over 160 publications in books and peer-reviewed journals and conferences. His research interests include advanced signal processing, and machine learning for wireless physical layer and high-resolution 3D/4D radars.

HICHEM SAHLI is currently a Professor of computer vision and machine learning with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Brussels, Belgium, and a Group Coordinator with the Interuniversitair Micro-Elektronica Centrum vzw (IMEC), Leuven, Belgium. He also coordinates the Joint VUB-NPU Audio-Visual Signal Processing (AVSP) Laboratory. AVSP deals with applied and theoretical problems related to computer vision, machine learning, and signal, audio, and image processing, for applications linked to affective computing and multi-modal interaction. His work deals with the development of algorithms, analysis, and novel principles for learning. His current research interests include computer vision and machine learning, especially in the areas of object detection and tracking, recognition, shape reconstruction, and image segmentation.