SparseSense: Human Activity Recognition from Highly Sparse Sensor Data-streams Using Set-based Neural Networks

Alireza Abedin, S. Hamid Rezatofighi, Qinfeng Shi and **Damith C. Ranasinghe**
School of Computer Science, The University of Adelaide, Australia
{alireza.abedinvaramin, hamid.rezatofighi, javen.shi, damith.ranasinghe}@adelaide.edu.au

Abstract

Batteryless or so called *passive* wearables are providing new and innovative methods for human activity recognition (HAR), especially in healthcare applications for older people. Passive sensors are low cost, lightweight, unobtrusive and desirably disposable; attractive attributes for healthcare applications in hospitals and nursing homes. Despite the compelling propositions for sensing applications, the data streams from these sensors are characterized by *high sparsity*—the time intervals between sensor readings are irregular while the number of readings per unit time are often limited. In this paper, we rigorously explore the problem of learning activity recognition models from temporally sparse data. We describe how to learn directly from sparse data using a deep learning paradigm in an end-to-end manner. We demonstrate significant classification performance improvements on real-world passive sensor datasets from older people over the state-of-the-art deep learning human activity recognition models. Further, we provide insights into the model's behaviour through complementary experiments on a benchmark dataset and visualization of the learned activity feature spaces.

1 Introduction

Understanding human activities using wearables is the basis for an increasing number of healthcare applications such as rehabilitation, gait analysis, falls detection and falls prevention [Bulling *et al.*, 2014]. In particular, older people have expressed a preference for unobtrusive and wearable sensing modalities [Gövercin *et al.*, 2010; Torres *et al.*, 2017]. While traditional wearables employ battery powered devices, new opportunities for human activity recognition applications, especially in healthcare, are being created by batteryless or *passive* wearables operating on harvested energy [Chen *et al.*, 2015; Lemey *et al.*, 2016]. In contrast to using often bulky and obtrusive battery powered wearables, passive sensing modalities provide maintenance-free, often disposable, unobtrusive and lightweight devices highly desirable to both older people and healthcare providers. However, the very nature of these sensors leads to new challenges.

Problem. The process of operating a batteryless sensor and transmitting the data captured is reliant on harvested power. Due to variable times to harvest adequate energy to operate sensors, the data-streams generated are highly sparse with variable inter-sample times. We illustrate the problem in Fig. 1 for a data stream captured by a body-worn passive sensor. We can see two key artefacts: *i*) the variable time intervals between sensor data reporting times; and *ii*) the relatively low average sampling rate. In this paper, we consider the problem of learning human activity recognition (HAR) models from *sparse data-streams* using a deep learning paradigm in an *end-to-end* manner.

Current Approaches. Wearable sensors generate time-series data. Consequently, the dominant human activity recognition pipeline uses fixed duration sliding window partitioning to feed neural networks during both training and inference stages [Wang *et al.*, 2019; Guan and Plötz, 2017; Ordóñez and Roggen, 2016; Hammerla *et al.*, 2016; Yang *et al.*, 2015; Zeng *et al.*, 2014]. When dealing with sparse data partitions, a common remedy is to rely on interpolation techniques as a pre-processing step to synthesize sensor observations to obtain a fixed size representation from time-series partitions as illustrated in Fig. 1 [Wickramasinghe and Ranasinghe, 2015; Gu *et al.*, 2018]. However, we recognize two key issues with an interpolated sparse data-stream:

- Interpolating between sensor readings that are temporarily distant can potentially lead to poor approximations of missing measurements and contextual activity information. Accordingly, adoption of convolutional filters or recurrent layers to extract temporal patterns from the poorly approximated measurements may potentially propagate the estimation errors to the activity recognition model—we substantiate this through extensive experiments in Section 3.3.
- Interpolation is as an intermediate processing step that prevents end-to-end learning of activity recognition models directly from raw data and introduces real-time inference delays in time critical applications—we demonstrate the time overheads imposed in Section 3.3.

Our Approach. Instead of relying on the naturally poor temporal correlations between consecutively received samples in sparse data-streams, we consider incentivizing the activity recognition model to uncover discriminative representa-

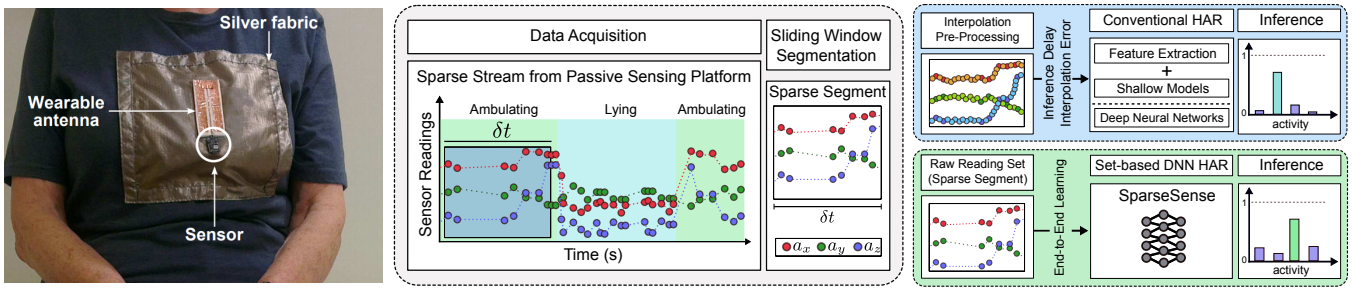


Figure 1: Left: Older volunteer wearing a passive sensor over their clothing in the *clinical rooms* public datasets used in this work (datasets and figure from [Torres *et al.*, 2013]). Right: An overview of the conventional sparse data-stream classification pipeline (blue plane) versus our novel set-based deep learning pipeline (green plane). Initially, data-streams from passive sensors are partitioned. The conventional pipeline then applies interpolation pre-processing on the sparse segments to synthesize fixed temporal context for model training and inference. In contrast, our proposed approach elegantly allows end-to-end learning of activity recognition models directly from sparse segments to deliver highly accurate classification decisions.

tions from the input sensory data partitions of various sizes to distinguish different activity categories. Our intuition is that a few information bearing sensor samples, although not temporally consistent, can capture adequate amount of information. Therefore, we propose learning HAR models directly from sparse data-streams. An illustrative summary of our proposed methodology for sparse data-stream classification in comparison with the conventional treatment is presented in Fig. 1.

In this paper, we describe how human activity recognition with sparse data-streams can be elegantly handled using deep neural networks in an end-to-end learning process. Given that we no longer rely on often poor temporal information, we represent sparse data stream partitions as unordered sets with various cardinalities from which embeddings capable of discriminating activities can be learned. Our approach is inspired by recent research efforts to investigate set-based deep learning paradigms to address a new family of problems where inputs [Qi *et al.*, 2017; Zaheer *et al.*, 2017] of the task are naturally expressed as sets with unknown and unfixed cardinalities. Therefore, our approach here is to develop activity recognition models that can learn and predict from incomplete sets of sensor observations, without requiring any extra interpolation efforts.

Contribution. In particular: *i)* We solve a new problem with a deep neural network formulation—learning from sparse sensor data-streams in an end-to-end manner; *ii)* We show that set learning can tolerate missing information which otherwise would not be possible with conventional DNN; and *iii)* We demonstrate that our novel treatment of the problem yields significantly outperforming recognition models with lower inference delays compared with the state-of-the-art on naturally sparse public datasets—over 4% improvement in the best case. We further compare with a benchmark HAR dataset and provide deeper insights into the performance improvements obtained from our proposed approach.

2 Methodology

We first present a formal description of human activity recognition problem with sparse data-streams and introduce the notations used throughout this paper before elaborating on our proposed activity recognition framework to learn directly from sparse data-streams in an end-to-end manner.

2.1 Problem Formulation

Consider a collected data-stream of raw time-series samples from body-worn sensors of the form $\mathbf{S} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_\mathrm{T})$, where $\mathbf{x}_t \in \mathbb{R}^\mathrm{d}$ is a multi-dimensional vector that contains sample measurements over d distinct sensor channels at time step t and T is the total length of the sequence. Without loss of generality, we assume a hardware-specific sampling rate for the wearable sensors, denoted by f.

HAR with Uniform Time-series Data. In an ideally controlled laboratory setup, sensor samples are constantly taken at regular intervals of $\frac{1}{\mathrm{f}}$ seconds. In such case, applying the commonly adopted time-series segmentation technique with a sliding window of fixed temporal context δt yields the labeled dataset

$$\mathcal{D}_{\text{uniform}} = \{(\mathbf{X}_1, \mathbf{y}_1), (\mathbf{X}_2, \mathbf{y}_2), ..., (\mathbf{X}_n, \mathbf{y}_n)\}, \quad (1)$$

where $\mathbf{X}_i = [\mathbf{x}_i, ..., \mathbf{x}_{i+\mathrm{m}-1}] \in \mathbb{R}^{\mathrm{d}\times\mathrm{m}}$ is a *fixed size* segment of captured sensor readings, $\mathrm{m} = \mathrm{f}\delta t$ is the constant number of received samples, and $\mathbf{y}_i$ denotes the corresponding one-hot encoded ground-truth from the pre-defined activity space $\mathcal{A} = \{a_1, ..., a_c\}$. The acquired dataset can then be utilized to train activity recognition models using out-of-the-box machine learning techniques.

HAR with Sparse Time-series Data. Unfortunately, sparse time-series data often found in real-world deployment settings, especially with passive sensors have variable inter-sensor observation intervals. In this case, utilizing a fixed time sliding window approach to segment the sparse data-stream results in the labeled dataset:

$$\mathcal{D}_{\text{sparse}} = \{(\mathcal{X}_1^{m_1}, \mathbf{y}_1), (\mathcal{X}_2^{m_2}, \mathbf{y}_2), ..., (\mathcal{X}_n^{m_n}, \mathbf{y}_n)\}, \quad (2)$$

where $\mathcal{X}_i^{m_i} = \{\mathbf{x}_i, ..., \mathbf{x}_{i+m_i-1}\} \in \overbrace{\mathbb{R}^\mathrm{d} \times ... \times \mathbb{R}^\mathrm{d}}^{m_i}$ is a set of sparse sensor observations during a timed window, $m_i \in \mathbb{N}$ is the cardinality of the obtained observation set, and $\mathbf{y}_i$ denotes the corresponding activity class. We emphasize that the number of received sensor readings in the time interval δt is *unfixed* for different sensory segments and upper bounded by

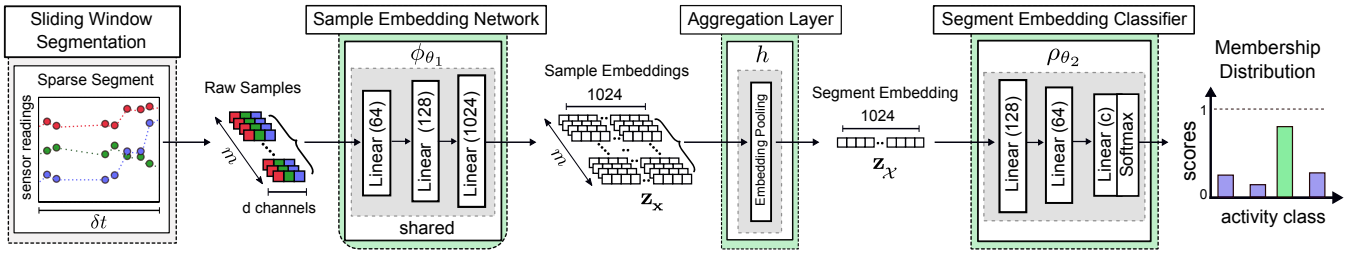


Figure 2: *SparseSense* architecture. The proposed network consumes sets of raw sensor observations with potentially varying cardinalities, uncovers latent projections for individual samples, aggregates sample embeddings into a global segment embedding, and maps the acquired segment embedding to its corresponding activity category. The number of neurons constituting the linear layers are outlined in parenthesis. All layers utilize ReLUs for non-linear transformation of activations except for the last layer which leverages a softmax activation function.

the sensor sampling rate; *i.e.*, for any given sensory segment $\mathcal{X}_i^{m_i}$, we have $m_i \leq \mathrm{f}\delta t$.

In this paper, having acquired the training dataset of sparse sensory segments $\mathcal{D}_{\text{sparse}} = \{(\mathcal{X}_i^{m_i}, \mathbf{y}_i)\}_{i=1}^n$, we intend to directly learn a mapping function $\mathcal{F}_{\Theta^*} : 2^{\mathbb{R}^\mathrm{d}} \rightarrow \mathcal{A}$, that operates on input sensory sets with unfixed cardinalities and accurately predicts the underlying activity classes,

$$\mathbf{y}_i = \mathcal{F}_{\Theta^*}(\mathcal{X}_i^{m_i}) = \mathcal{F}_{\Theta^*}(\{\mathbf{x}_i, ..., \mathbf{x}_{i+m_i-1}\}),\ \forall i \in \{1, ..., n\}.$$

2.2 SparseSense Framework

Our work is built upon the insight that incorporating interpolation techniques to recover the missing measurements across large temporal gaps between received sensor observations in sparse data-streams leads to poor estimations and therefore, significant interpolation errors. As we demonstrate in Section 3.3, the adoption of convolutional filters or recurrent layers to extract temporal patterns from the poorly approximated measurements can potentially propagate the estimation errors to the activity recognition model.

Instead of forcing the network to exploit the potentially weak temporal correlations in sparse data-streams, we propose learning global embeddings from sets that encode aggregated information related to an activity. Therefore, we propose formulating sparse segments as unordered sets with unfixed and unknown number of sensor readings. Hence, we design *SparseSense* as a set-based activity recognition framework for the HAR task that directly manipulates sets of received sensor readings with irregular inter-sample observation intervals and outputs the corresponding activity membership distributions. Our approach provides a complete end-to-end learning method that incentivizes the activity recognition model to uncover globally discriminative representations for the input sparse segments with variable number of samples, and distinguish different activity categories accordingly.

Network Architecture

The overall architecture of our proposed *SparseSense* network is illustrated in Fig. 2. Essentially, we approximate the optimal mapping function $\mathcal{F}_{\Theta^*}$ through training of a deep neural network parameterized by Θ. The primary task for integrating set learning into deep neural networks is employing a *shared network* to map each set element independently into a higher dimensional embedding space (to facilitate class separability) and adopting a symmetric operation across the element embeddings to generate a global representation for the entire set that does not rely on the set element orderings. We incorporate this pipeline into the building blocks of our network as elucidated in what follows:

Input. Adopting sliding window segmentation over the sparse data-stream yields sets of sparsely received sensor observations $\mathcal{X}$ in the pre-defined temporal window δt, with potentially varying cardinalities.

Shared Sample Embedding Network. The embedding network $\phi_{\theta_1} : \mathbb{R}^\mathrm{d} \rightarrow \mathbb{R}^\mathrm{z}$ parameterized by θ_1, operates identically and independently on each sample measurement $\mathbf{x}$ within the received observation set $\mathcal{X}$ and learns a corresponding higher dimensional projection $\mathbf{z}_\mathbf{x} \in \mathbb{R}^\mathrm{z}$ to alleviate separability of activity features in the new embedding space; *i.e.*, $\mathbf{z}_\mathbf{x} = \phi_{\theta_1}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$. Technically, ϕ_{θ_1} is a standard multi-layer perceptron (MLP) whose parameters are *shared* between the sensor sample readings; *i.e.*, all samples undergo the same layer operations and are therefore processed identically through a copy of the MLP.

Aggregation Layer. Described by $h : \mathbb{R}^\mathrm{z} \times ... \times \mathbb{R}^\mathrm{z} \rightarrow \mathbb{R}^\mathrm{z}$, the aggregation layer applies a symmetric operation across the latent representations of individual sensor samples and extracts a fixed size global embedding $\mathbf{z}_\mathcal{X} \in \mathbb{R}^\mathrm{z}$ to represent the sensory segment as a whole. Thus, for a given sensory segment $\mathcal{X}_i$, we have

$$\mathbf{z}_{\mathcal{X}_i} = h(\{\mathbf{z}_{\mathbf{x}_i}, ..., \mathbf{z}_{\mathbf{x}_{i+m_i-1}}\}). \tag{3}$$

Notably, the shared sample embedding network coupled with the symmetric aggregation layer allow summarizing sparse segments with effective high-dimensional projections that *i*) do not rely on the weak temporal ordering of the sparse samples, and, *ii*) ensure fixed size tensor representations independent of the number of received readings. Inspired by [Qi *et al.*, 2017], in this paper, we set h to incorporate a feature-wise maximum pooling across sample embeddings which promises robustness against set element perturbations.

Segment Embedding Classifier. Described by $\rho_{\theta_2} : \mathbb{R}^\mathrm{z} \rightarrow \mathcal{A}$ parameterized by θ_2 is trained to exploit the segment embeddings $\mathbf{z}_\mathcal{X}$ through multiple layers of non-linearity and predict the corresponding activity class probability distributions $\hat{\mathbf{y}}$; *i.e.*, $\hat{\mathbf{y}} = \rho_{\theta_2}(\mathbf{z}_\mathcal{X})$. Here, a softmax activation function governs the output of our network to yield posterior probability distributions over the activity space $\mathcal{A}$.

Dataset (clinical room)	HAR Model	Interpolant (acceleration)	Input	Window Size (δt)	Precision$_m$ (mean±std)	Recall$_m$ (mean±std)	F-score$_m$ (mean±std)
Roomset1	SVM^{lin*}	Cubic	Hand-crafted features	4 seconds	87.87±2.55	83.44±1.72	84.96±1.23
	SVM^{rbf*}	None	Hand-crafted features	8 seconds	90.39±2.70	87.42±1.42	88.45±1.68
	CRF*	Linear	Hand-crafted features	2 seconds	85.97±2.43	82.35±3.08	83.73±2.40
	Bi-LSTM	Linear	Raw sensor readings	2 seconds	89.97±0.78	85.11±0.99	86.96±1.06
	DeepCNN	Quadratic	Raw sensor readings	4 seconds	92.43±1.21	87.93±1.74	89.73±1.55
	DeepConvLSTM	Linear	Raw sensor readings	4 seconds	91.87±1.43	88.88±1.79	90.42±1.54
	(Ours) SparseSense	None	Raw sensor readings	2 seconds	**95.0±0.75**	**94.08±0.78**	**94.51±0.62**
Roomset2	SVM^{lin*}	Cubic	Hand-crafted features	2 seconds	87.06±4.10	84.00±2.90	84.97±3.74
	SVM^{rbf*}	None	Hand-crafted features	8 seconds	90.97±4.11	83.88±2.04	85.53±2.86
	CRF*	None	Hand-crafted features	16 seconds	83.68±6.50	78.29±3.58	79.99±4.76
	Bi-LSTM	Previous	Raw sensor readings	2 seconds	92.38±0.91	91.4±0.62	91.78±0.58
	DeepCNN	Linear	Raw sensor readings	4 seconds	93.11±0.94	91.7±1.18	92.36±0.99
	DeepConvLSTM	Previous	Raw sensor readings	4 seconds	94.16±0.52	93.05±0.78	93.77±0.63
	(Ours) SparseSense	None	Raw sensor readings	2 seconds	**97.07±0.52**	**96.88±0.34**	**96.97±0.37**

Table 1: Performance comparison for the naturally sparse clinical room datasets. Reported results and design choices for the baselines with asterisks are quoted directly from [Wickramasinghe and Ranasinghe, 2015]. The remaining baselines are replicated following their original paper descriptions. To ensure a fair comparison, the reported results are the highest achieving combination of window sizes (explored in the range of $\{2, 4, 8, 16\}$ seconds) and interpolants (*linear*, *cubic*, *quadratic* and *previous*) selected for all the competing approaches.

Summary. Now, we can express the mathematical operations constituting the forward pass of our proposed activity recognition model for a given sparse sensory segment $\mathcal{X}_i$ as:

$$\mathcal{F}_\Theta(\mathcal{X}_i^{m_i}) = \rho_{\theta_2}\left(h(\{\phi_{\theta_1}(\mathbf{x}_i), ..., \phi_{\theta_1}(\mathbf{x}_{i+m_i-1})\})\right), \quad (4)$$

where Θ denotes the collection of all network parameters; *i.e.*, $\Theta = (\theta_1, \theta_2)$.

Network Training and Activity Inference

During the training process, the goal is to learn the network parameters Θ such that the disagreement between the network outputs and the corresponding ground-truth activities is minimized for the training dataset. We can precisely express this discrepancy minimization by adopting an end-to-end optimization of the negative log-likelihood loss function $\mathcal{L}_{\text{NLL}}$ on the training dataset $\mathcal{D}_{\text{sparse}}$; *i.e.*,

$$\Theta^* = \arg\min_{\Theta} \sum_{i=1}^{n} \mathcal{L}_{\text{NLL}}\big(\mathcal{F}_\Theta(\mathcal{X}_i^{m_i}), \mathbf{y}_i\big). \quad (5)$$

As the training process progresses and the corresponding objective function is minimized, the SparseSense network uncovers highly discriminative embeddings for sparse segments that allow effective separation of classes in the activity space.

Once the training procedure converges and the optimal network parameters Θ^* are learned from the training dataset, we adopt a maximum a posteriori (MAP) inference to promote the most probable activity category for any given set of sparse sensor readings; *i.e.*, the highest scoring class in the softmax output of the network is chosen to be the final prediction.

3 Experiments and Results

3.1 Datasets

To ground our study, we evaluate our proposed framework on two naturally sparse public datasets collected in clinical rooms with older people using a body-worn batteryless sensor intended for ambulatory monitoring in hospital settings. For further insights, we also present extensive empirical analysis of our approach on a HAR benchmark dataset with synthesized sparsification and provide comparisons against the state-of-the-art deep learning based HAR models.

Clinical Room Datasets [Torres *et al.*, 2013]. The dataset is collected from fourteen older volunteers, with a mean age of 78 years, performing a set of broadly scripted activities while wearing a W^{2}ISP over their attire at the sternum level (see Fig. 1). The W^{2}ISP is a passive sensor-enabled RFID (Radio Frequency Identification) device that operates on harvested electromagnetic energy emitted from nearby RFID antennas to send data with an upper-bound sampling rate of 40 Hz. Data collection was carried out in two clinical rooms with two different antenna deployment configurations to power the sensor and capture data; resulting in *Roomset1* and *Roomset2* datasets. Each sensor observation in the obtained datasets records triaxial acceleration measurements as well as contextual information from the RFID platform indicating the antenna identifier and the strength of the received signal from the sensor. These recordings were manually annotated with *lying on bed*, *sitting on bed*, *ambulating* and *sitting on chair* to closely simulate hospitalized patients' actions. Consecutive samples in the sparse datastreams from *Roomset1* and *Roomset2* exhibit high mean time differences of 0.37 s and 0.72 s respectively.

WISDM Benchmark Dataset [Kwapisz *et al.*, 2011]. This dataset contains acceleration measurements from 36 volunteers collected through controlled, laboratory conditions while performing a specific set of activities. The sensing device used for data acquisition is an Android mobile phone with a constant sampling rate of 20 Hz and placed in the subjects' front pant's pocket. The sensor samples carry annotations from *walking*, *jogging*, *climbing up stairs*, *climbing down stairs*, *sitting* and *standing*. The collected dataset delivers high quality data and has frequently been used in HAR studies for benchmarking purposes. Accordingly, we find this dataset a suitable choice for thorough investigation of our SparseSense network under different levels of synthesized data sparsification.

3.2 Experiment Setup

In this study, we initially perform per-feature normalization to scale real-valued observation attributes to the $[0, 1]$ inter-

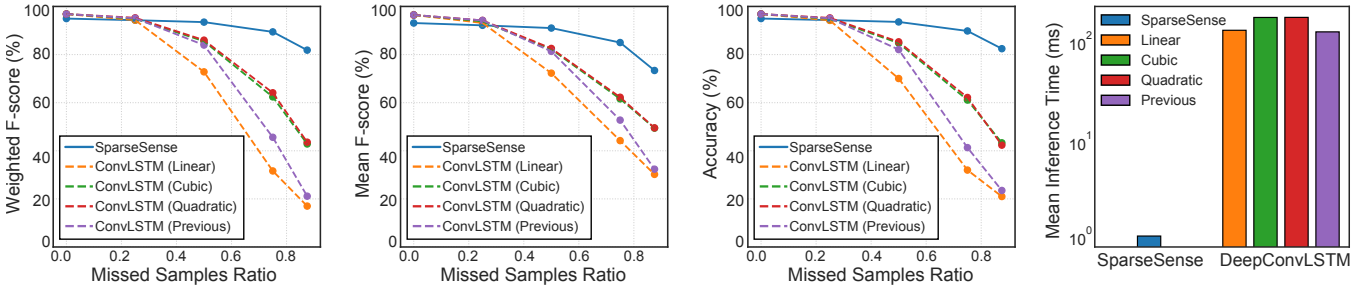


Figure 3: Activity recognition performance and computational complexity of our proposed *SparseSense* framework for sparse data-stream classification against the state-of-the-art DeepConvLSTM HAR model. DeepConvLSTM is tailored for sensory segments of fixed size temporal context and thus, requires sparse segments be re-sampled through adoption of interpolation methodologies prior to inputting them.

val. We consider a fixed temporal context δt and obtain sensory partitions by sliding a window over the recorded data-streams. The acquired segments are assumed to reflect adequate information related to a wearer's current activity and are thus, assigned a categorical activity label based on the most observed sample annotation in the time-span of the sliding window.

We implement the experiments in Pytorch [Paszke *et al.*, 2017] deep learning framework on a machine with a NVIDIA GeForce GTX 1060 GPU. The SparseSense deep human activity recognition model is trained in a fully-supervised fashion by back-propagating the gradients of the loss function in mini-batches of size 128; *i.e.*, the network parameters are iteratively adjusted according to the RMSProp [Tieleman and Hinton, 2012] update rule in order to minimize the negative log-likelihood loss using mini-batch gradient descent. The optimizer learning rate is initialized with 10^{-4}, reduced by a factor of 0.1 after 100 epochs, and the optimization is ceased after 150 epochs. Further, a weight decay of 10^{-4} is imposed as L_2 penalty for regularization. Following previous studies, we employ 7-fold stratified cross-validation on the datasets and preserve activity class distributions across all folds. Each constructed fold is in turn utilized once for validation while the remaining six folds constitute the training data.

3.3 Baselines and Results

Clinical Room Experiments

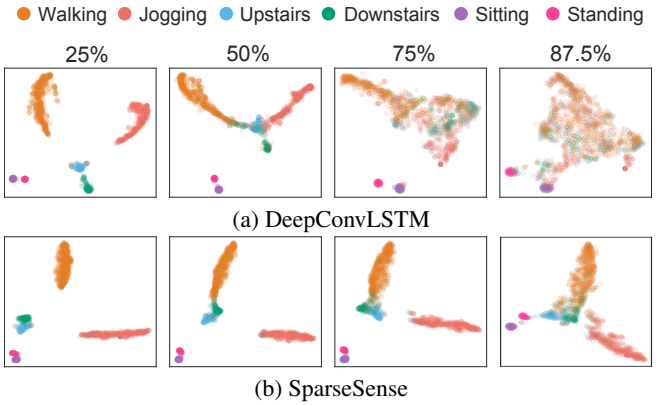
In Table 1, we report the mean F-measure (F-score$_m$) as the widely adopted evaluation metric and compare SparseSense with activity recognition models previously studied for the naturally sparse clinical room datasets as well as the state-of-the-art deep learning based HAR models. [Wickramasinghe and Ranasinghe, 2015] has explored shallow models including support vector machines (*SVM*^{lin} and *SVM*^{rbf}) and conditional random fields (*CRF*) trained using hand-crafted features extracted from either raw or interpolated sparse segments. In addition, we investigate the effectiveness of *Bi-LSTM* [Hammerla *et al.*, 2016], *DeepCNN* and *DeepConvLSTM* [Ordóñez and Roggen, 2016] as solid deep learning baselines representing the state-of-the-art for HAR applications.

Bi-LSTM leverages bidirectional LSTM recurrent layers to directly learn the temporal dependencies of samples within the sensory segments. Both DeepCNN and DeepConvLSTM adopt four layers of 1D convolutional filters along the temporal dimension of the fixed size segmented data to automatically extract feature representations. However, DeepCNN is then followed by two fully connected layers to aggregate the feature representations while DeepConvLSTM utilizes a two layered LSTM to model the temporal dynamics of feature activations prior to the final softmax layer. We refer interested readers to the original papers introducing the HAR models for further details and network specifications. Following [Wickramasinghe and Ranasinghe, 2015], for each baseline we explore progressively increasing window durations, *i.e.* $\delta t \in \{2, 4, 8, 16\}$, adopt per-channel interpolation schemes (*linear*, *cubic*, *quadratic* and *previous*) to compensate for the missing acceleration data and report the highest achieving configurations in Table 1 for all competing approaches. In this regard, *cubic* and *quadratic* interpolation schemes respectively refer to a spline interpolation of second and third order, and the *previous* scheme fills missed values with the previously received sensor readings.

From the outlined results, we observe that the SparseSense network outperforms all the baseline models with a large margin in the task of sparse data-stream classification. Notably, the baselines are: *i*) well-engineered shallow models that require a large pool of domain expert hand-crafted features; and *ii*) state-of-the-art deep learning HAR models that demand interpolation techniques to synthesize regular sensor sampling rates. In contrast, SparseSense seamlessly operates on sparse sets of sensory observations without requiring any extra interpolation efforts or manually designed features, and automatically extracts highly discriminative embeddings for the classification task in an end-to-end framework.

WISDM Benchmark Experiments

To provide additional insights onto the model's behavior, we conduct experiments on WISDM benchmark dataset and analyze the network's classification performance under different levels of synthesized data sparsification. Taking into account the superior performance of DeepConvLSTM among the baselines in Table 1, here we only present comparisons with this model. Following [Kwapisz *et al.*, 2011; Alsheikh *et al.*, 2016], we partition the data-streams into fixed size sensory segments using a sliding window of 10 seconds duration (corresponding to 200 sensor readings) and train the HAR models on the acquired segmented data. Subsequently



(a) DeepConvLSTM

(b) SparseSense

Figure 4: A 2D visualization of the learned feature spaces for (a) DeepConvLSTM and (b) SparseSense under different data sparsity levels indicated by the percentage of artificially imposed missed readings. SparseSense learns robust embeddings that maintain cluster separation even under significant missing sample settings.

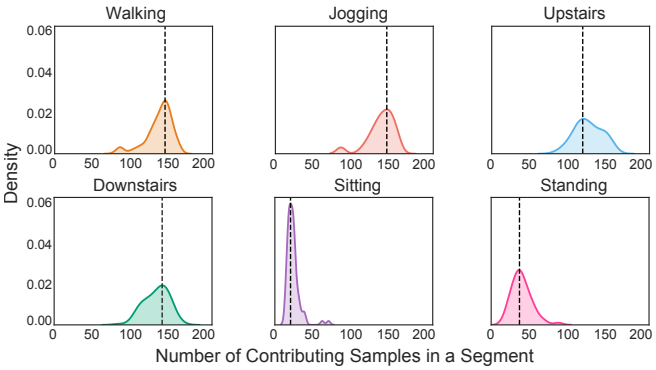


Figure 5: The density plots for the number of contributing samples that constitute the aggregated segment embeddings for each activity category of WISDM dataset.

at test time, we drop sensor readings at random time-steps in order to generate synthetic sparse segments.

Tolerance to Data Sparsity and Delays

In Fig.3, the obtained evaluation measures are plotted for both HAR models under different sparsification settings. When data segments are received in full, DeepConvLSTM performs better than SparseSense due to its ability in capturing temporal dependencies between consecutive sensor readings. However, as the data sparsity increases and the temporal correlation weakens, we observe a significant drop in classification performance of DeepConvLSTM. Notably, with large temporal gaps between sensor observations, interpolation techniques cannot produce good estimations of the missing samples and fail to recover the original acceleration measurements which in turn impacts the classification decisions of DeepConvLSTM. In contrast, not only does SparseSense achieve comparable classification results for completely received sensory segments, but it also displays great robustness to data sparsity by making accurate decisions for incomplete segments of sensor data. In addition, we show in the bar plot the mean processing time required by the HAR models to make predictions on a mini-batch of 128 segments. Clearly, our framework demonstrates a significant advantage over other HAR models for real-time activity recognition using sparse data-streams by removing the need for prior interpolation pre-processing.

SparseSense Model Behaviour

We visualize the learned feature spaces for both models in 2D space using t-distributed stochastic neighbor embedding (t-SNE) [Maaten and Hinton, 2008] in Fig. 4. In the absence of significant data sparsity, the segment embeddings belonging to each activity category are clustered together while different activities are separated in the feature space. However, while SparseSense is able to maintain this cluster separation for severely missed sample ratios and incomplete observation sets, DeepConvLSTM clearly struggles to discriminate between the interpolated segments. Technically, the symmetric max pooling operation in the aggregation layer of SparseSense incentivises our HAR model to summarize sensory segments using only the most informative readings present in the segment. We refer to these information bearing samples as the *contributing samples*.

In Fig. 5, we provide density plots for the number of sensor readings that ultimately contribute to the aggregated segment embeddings for each activity category of the WISDM dataset. We observe that SparseSense intelligently summarizes the segments through discarding potentially redundant information in the neighboring samples when windows of fully received samples ($m = 200$) are presented to the network—see the density plots where the tails towards 200 contributing samples have a probability of zero. More interestingly, the network displays a clear distinction in its behavior towards learning embeddings for static activities (*i.e.*, sitting and standing) as opposed to dynamic activities (*i.e.*, walking, jogging and climbing stairs) by exploiting far fewer number of sensor observations out of the $m = 200$ received samples in the window. This can be intuitively understood as static activities reflect signal patterns with small changes in sensor measurements of a timed window as compared with dynamic activities and thus, can be summarized with smaller number of observations.

4 Conclusions

We present an end-to-end human activity recognition framework to learn directly from temporally sparse data-streams using set-based deep neural networks. In contrast to previous studies that rely on interpolation pre-processing to synthesize sensory partitions with fixed temporal context, our proposed *SparseSense* network seamlessly operates on sparse segments with potentially varying number of sensor readings and delivers highly accurate predictions in the presence of missing sensor observations. Through extensive experiments on publicly available HAR datasets, we substantiate how our novel treatment for sparse data-stream classification results in recognition models that significantly outperform state-of-the-art deep learning based HAR models while incurring notably lower real-time prediction delays. We believe that our work will provide a new method for understanding human motion data from passive wearables for health care applications.

References

[Alsheikh *et al.*, 2016] Mohammad Abu Alsheikh, Ahmed Selim, Dusit Niyato, Linda Doyle, Shaowei Lin, and Hwee-Pink Tan. Deep activity recognition models with triaxial accelerometers. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Bulling *et al.*, 2014] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3):33, 2014.

[Chen *et al.*, 2015] Shengjian Jammy Chen, Christophe Fumeaux, Damith Chinthana Ranasinghe, and Thomas Kaufmann. Paired snap-on buttons connections for balanced antennas in wearable systems. *IEEE Antennas and Wireless Propagation Letters*, 14:1498–1501, 2015.

[Gövercin *et al.*, 2010] Mehmet Gövercin, Y Költzsch, M Meis, S Wegel, M Gietzelt, J Spehr, S Winkelbach, M Marschollek, and E Steinhagen-Thiessen. Defining the user requirements for wearable and optical fall prediction and fall detection devices for home use. *Informatics for health and social care*, 35(3-4):177–187, 2010.

[Gu *et al.*, 2018] Fuqiang Gu, Kourosh Khoshelham, Shahrokh Valaee, Jianga Shang, and Rui Zhang. Locomotion activity recognition using stacked denoising autoencoders. *IEEE Internet of Things Journal*, 5(3):2085–2093, 2018.

[Guan and Plötz, 2017] Yu Guan and Thomas Plötz. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):11, 2017.

[Hammerla *et al.*, 2016] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1533–1540, 2016.

[Kwapisz *et al.*, 2011] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[Lemey *et al.*, 2016] Sam Lemey, Sam Agneessens, Patrick Van Torre, Kristof Baes, Jan Vanfleteren, and Hendrik Rogier. Wearable flexible lightweight modular rfid tag with integrated energy harvester. *IEEE Transactions on Microwave Theory and Techniques*, 64(7):2304–2314, 2016.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9:2579–2605, 2008.

[Ordóñez and Roggen, 2016] Francisco Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017.

[Qi *et al.*, 2017] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[Tieleman and Hinton, 2012] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

[Torres *et al.*, 2013] Roberto L Shinmoto Torres, Damith C Ranasinghe, Qinfeng Shi, and Alanson P Sample. Sensor enabled wearable rfid technology for mitigating the risk of falls near beds. In *IEEE International Conference on RFID*, pages 191–198, 2013.

[Torres *et al.*, 2017] Roberto L Shinmoto Torres, Renuka Visvanathan, Derek Abbott, Keith D Hill, and Damith C Ranasinghe. A battery-less and wireless wearable sensor system for identifying bed and chair exits in a pilot trial in hospitalized older people. *PloS one*, 12(10):1–25, 2017.

[Wang *et al.*, 2019] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3 – 11, 2019.

[Wickramasinghe and Ranasinghe, 2015] Asanga Wickramasinghe and Damith Ranasinghe. Recognising activities in real time using body worn passive sensors with sparse data streams: To interpolate or not to interpolate? In *proceedings of the 12th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 21–30, 2015.

[Yang *et al.*, 2015] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3995–4001, 2015.

[Zaheer *et al.*, 2017] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401, 2017.

[Zeng *et al.*, 2014] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *6th International Conference on Mobile Computing, Applications and Services*, pages 197–205, 2014.