Reduction of the Position Bias via Multi-Level Learning for Activity Recognition

# 412

Abstract. The relative position of sensors on the human body generates two types of data: one relating to the movement of the position w.r.t. the body and another relating to the movement of the whole body w.r.t. the environment. These two data give two complementary and orthogonal components contributing differently to the activity recognition process. In this paper, we introduce an original approach that allows us to separate these data and to abstract away the exact on-body position of the sensors from the considered activities. We learn for these two totally orthogonal components (i) the bias that stems from the position and (ii) the actual patterns of the considered activities rid of these positional biases. We perform a thorough empirical evaluation of our approach on the SHL dataset featuring an on-body sensor deployment in a real-life setting. Obtained results show that we are able to substantially improve recognition performances. These results pave the way for the development of models that are agnostic to both the position of the data generators and the target users. Constructed models are found to be robust to evolving environments such as those we are confronted with in Internet of Things applications.

Keywords: Human activity recognition · Multi-level learning · Sensor deployments.

1 Introduction

The selection of the sensors position in the moving targets is a constraint that can be found in many fields, such as the human activity recognition [8,27,6]. The movements of the position of the target on which the sensors are positioned provide data related to two complementary natures (see figure 1). The first concerns the movement of the position relative to the target itself and the second concerns the movement of the target relative to its surroundings. In the case of human activities recognition, for example, we see that the kinetics of the movement of the hand during a race can be decomposed into a circular movement (CM) of the hand in relation to the shoulder and a translation movement (TM) associated with the whole body [24,13].

From this global observation and from this particular case, we can deduce at least three conclusions. The first one is the fact that for some CM concepts, data are enough to learn the target concept (The hand kinetics movement is enough

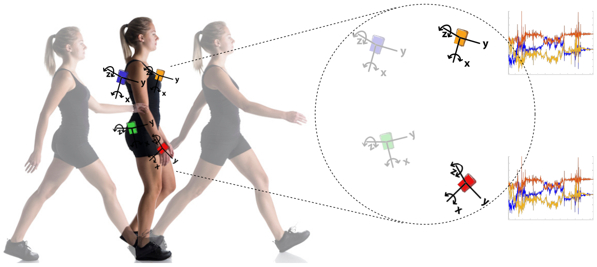


Fig. 1: Two Sensors from our used SHL dataset. The hand sensor undergoes two movements. The first is of the same nature as the torso sensor. It is linked to the translational movement of the body. The second is linked to the movement of the hand locally in relation to the body.

to determine if a person is at rest or running for example). The second one concern the fact that CM data cannot be shared with other data (for example those concerning torso), otherwise they generate noise and confusion rather than improving the recognition process. The CM data must be analyzed locally with local learners. The third conclusion is the fact that from each position in the moving target only TM data can be shared with other data resources. Because these data concern the same kind of data with different point of view (hand, torso, leg, etc.).

The main lesson of this article is to show that the natures of these data are different and obey different laws. Without this data decomposition process, the local part of the data adds position noise difficult to manage with centralized approaches. It shows that to integrate data from several positions (point of view) in approaches like federated learning, it is necessary to separate the data of the same nature (shareable) from the pure local ones linked to the specific kinetics of the position. Similar data can and should be shared to improve recognition rates. However, the specific data must be processed locally, otherwise, it would correspond to the noise and considerably reduce the efficiency of the learning problem. When the decomposition process is well done, we note that for some activities the local model is efficient alone to totally recognize the global activity. It's the case for the following example extracted from Corollo et al. biomechanics works [9]. They show, for example, that independently from what we call TM data, the CM data of the hand are enough and more efficient to recognize running activities (see figure 2).

In this paper, we leverage the separability of the data into universal and position-specific components. These components have distinctive contributions with regards to the target concepts to be learned. Indeed, the universal components of each individual position are assumed to be drawn from an unknown but common distribution whereas the position-specific components are assumed to be drawn from distinct distributions. This brings an interesting property that allows us to fuse the universal components as seen from different points of view (positions) while identifying the position-specific components which could serve

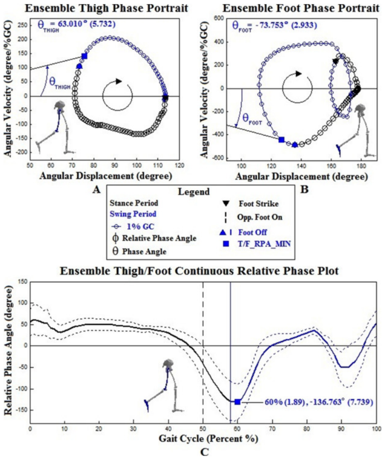


Fig. 2: Typical ensemble thigh (A) and foot (B) phase portraits and continuous thigh/foot relative phase plot (C) during running activities From [9].

as additional knowledge in some situations. Therefore, the problem of learning human activities can be defined by the collaboration of the general model using shared data and local models. It is this original and atypical problem that will be studied in this article.

Learning concepts using data generated from different perspectives (or points of view) is challenging and commonly dealt with in the literature. In the particular field of the recognition of human activity, the proposed approaches mainly focus on the study of the best placement of sensors on the body and its impact on the recognition of target activities [11,20,5,4,27,6]. There are also rare approaches offering pipelines including recognition of the position of the data generator followed by the activity recognition [33,26] or including explicit model of the context like 'sitting in a car', 'taking a shower', etc.[10,3]. Other approaches e.g., [20], try to develop position-agnostic models by using a set of heuristics that significantly increase the robustness of motion sensor-based activity recognition with respect to sensor displacement. More generally, a long line of research related to multi-view learning considers such settings where representations are constructed for each individual view, as a first step, before being fused or aligned together in order to get more robust representations [2,14,21]. However, we are not aware of any work proposing an approach like the one described in this article where data are separated into orthogonal data sources.

To deal with these complementary data sources, we propose a multi-level model of abstraction of the data generator position composed of a central learner (or set of local generic learners) and a set of specific local learners. The local learners (one for each position) use only specific local data concerning the local relativity (position versus the whole target). They are responsible for learning (1) the position-dependent patterns of activities and (2) the movements that link them to the individual. Their goal is to abstract the learning examples from the position from which they were generated. This can be likened to a data projection operation towards a space devoid of positional biases. The central learner (or set of local generic learners) uses the aggregated common part of data coming from the local point of view learners. In the case of local generic learners, each local generic learner uses the part of local data describing the kinetic of the target object without local perturbations. Using a central learner implies aggregating data while using a set of local generic learners implies using collaborative approaches between learners like ensemble learning approaches or federated learning.

In addition to the efficiency of such an approach induced by the separation of data of different natures. It has other advantages such as the reduction of heterogeneity which poses a major problem for learner aggregation techniques such as federated learning [30]. This is the case in the internet of things applications where several sensors are needed to collect data from different points of view. Separating the local relativity between data from the global one gives a simple model to reduce the heterogeneous aspect. In the case of HAR, when the body position is recognized, the locally produced data is normalized to obtain uniform human data regardless of position.

The contributions of our paper can be summarized as follows. (1)We propose an original approach to disregard the impact of the location of the sensor in the target to generate two types of orthogonal data. Local data not having to be shared otherwise it would generate noise and data representing a local point of view of a global situation which can be shared as data after projection in a common space or by doing local learning and aggregation step between these learners. (2) the specific data also give important indications on the activities. Sometimes they are totally sufficient to recognize certain activities as shown for example by the circular movements pattern of a hand during the running activity (see figure 2). Sometimes it is appropriate to combine them with the result of the global situation. In addition to the advantages in terms of efficiency and recognition quality, this approach has other advantages such as reducing the heterogeneity of the data shared between the different sources. (3) Extensive experiments on the SHL dataset featuring a real-world sensor deployment setting show the effectiveness of abstracting the impact of the data generator's position. We noticeably get substantial improvements in terms of: (i) the recognition performances of individual activities; and (iii) the robustness to evolution of the sensors deployments. (4) We perform a comprehensive comparative analysis of the various stages of our proposed approach via ablation studies which show the contribution of the dual interplay between the local and central learners.

This paper is organized as follows: Section 2 recalls the application and gives a formulation of the abstraction of the position problem. Section 3 gives details about the framework of the proposed multi-level abstraction architecture and summarizes the objective functions of both local and global learners. The details of the studied application and the encouraging results obtained are summarized in Section 4. Section 5 concludes the paper and gives some perspectives.

2 Problem Formulation

In this section, we briefly characterize the problem of abstracting the exact position of a given sensor and motivate the need for a multi-level approach to deal with it. We illustrate this problem on a concrete internet of things real-world application. We will use the SHL dataset [12] [1]. It is a highly versatile annotated dataset dedicated to mobility-related human activity recognition. It was recorded over a period of 7 months in 2017 in 8 different modes of transportation in real-life setting in the United Kingdom (*Still*, *Walk*, *Run*, *Bike*, *Car*, *Bus*, *Train*, and *Subway*). The dataset contains multi-modal data from a body-worn camera and from 4 smartphones, carried simultaneously at typical body locations (*Hand*, *Torso*, *Hips*, and *Bag*). The SHL dataset contains 3000 hours of labeled locomotion data in total. It includes 16 modalities such as accelerometer, gyroscope, magnetometer, linear acceleration, orientation, gravity, ambient pressure, cellular networks, etc.

Application details are presented in section 4. From a theoretical point of view, this kind of application can be summarized as follow: we consider settings where a collection $\mathcal{S}$ of M sensors (also called data generators or data sources), denoted $\{s_1, \ldots, s_M\}$, are positioned respectively at positions $\{p_1, \ldots, p_M\}$ in the object of interest. Each sensor s_i generates a stream $\mathbf{x}^i = (x_1^i, x_2^i, \ldots)$ of observations of a certain modality like acceleration or gravity. Furthermore, each observation is composed of channels, e.g. three axes of an accelerometer. The goal is to continuously recognize a set of target concepts $\mathcal{Y}$ like *running* or *biking* in the case of the human activity recognition according to all sensor's positions.

2.1 Abstraction of the Position

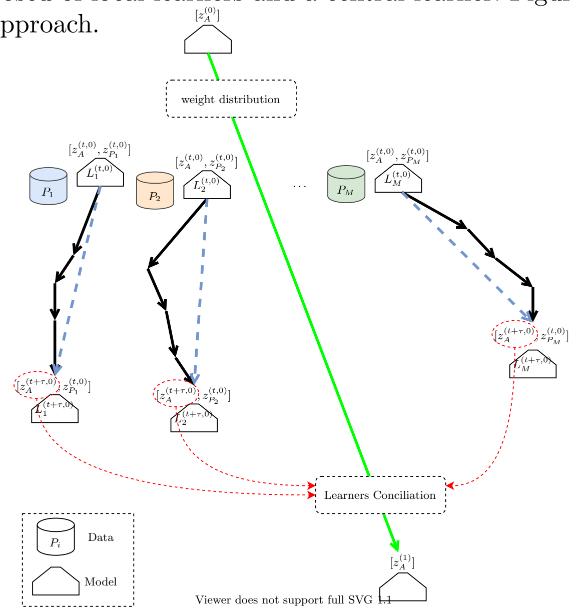
As described in the previous section, each sensor produces two types of orthogonal data. This problem can be formally defined as the construction, for each sensor s_i ($i \in \{1, \ldots, M\}$) of a latent representations z_i being a composition of (i) a universal (or position-invariant) components vector z_{iA}, and (ii) a position-specific components vector z_{iP}. The position-invariant components vector captures the information that is shared across all positions. On the other hand, the position-specific components vector captures specific and complementary insights with regard to the target concepts. The first problem to solve, in our model, is to build automatically this data decomposition process for each sensor. Thanks to this transformation each sensor $s \in \mathcal{S}$ will disentangle the data interlaced between the local and universal component $\mathbf{x}$ by projecting them into two separate representations z_A and z_P. Data z_P will be used only in a local learner, and z_A can be used in the local learner or shared with the same data coming from all other sensors in a global learner. This process allows us to have fine-grained control on the inference process where one can leverage different configurations in order to get optimal recognition performances. We notice, for example, that in certain situations the position-specific component alone is

[1] http://www.shl-dataset.org/

enough to recognize the activity (e.g., the circular movements of the hand are enough to distinguish between running and walking). In addition, since only position-independent data is shared, this process greatly reduces data heterogeneity. It, therefore, improves data aggregation techniques or learners such as federated learning [32] by sharing only the position-invariant data. When the data are not decomposed the position-specific part of the data represents noise for the global system. To deal with these two complementary representations, we propose a model based on multi-level processing to abstract the position as described bellow. In this model, we suppose that the position-invariant components share the data with central learner.

3 Multi-Level Abstraction of the Source Position

In this section, we propose an instantiation model of the proposed problem formulation composed of local learners and a central learner. Figure 3 summarizes the proposed approach.



$[z_A^{(0)}]$
weight distribution
$[z_A^{(t,0)}, z_{P_1}^{(t,0)}]$
P_1
$L_1^{(t,0)}$
$[z_A^{(t,0)}, z_{P_2}^{(t,0)}]$
P_2
$L_2^{(t,0)}$
…
$[z_A^{(t,0)}, z_{P_M}^{(t,0)}]$
P_M
$L_M^{(t,0)}$
$[z_A^{(t+\tau,0)}, z_{P_M}^{(t,0)}]$
$L_M^{(t+\tau,0)}$
$[z_A^{(t+\tau,0)}, z_{P_1}^{(t,0)}]$
$L_1^{(t+\tau,0)}$
$[z_A^{(t+\tau,0)}, z_{P_2}^{(t,0)}]$
$L_2^{(t+\tau,0)}$
Learners Conciliation
P_i
Data
Model
$[z_A^{(1)}]$
Viewer does not support full SVG 1.1

Fig. 3: Framework of the proposed multi-level abstraction architecture. The global learner $L_{\mathcal{S}}$ starts with an initial set of weights which are distributed to the local learners. The local learners L_p, one for each position p, learn the two vector components z_A and z_P, by performing independently a set of gradient steps which allows to get a newer version. These new versions are used during the conciliation step which gives us a new version of the global learner, and subsequently a more robust position-independent representation.

To perform the separation of the position-specific components from the universal ones we use a family of models based on variational autoencoders (VAEs)

(§ 3.1). We then propose a conciliation step based on the federated learning (FL) setting where the position-specific learners in our formulation are assimilated to the decentralized clients in FL (§ 3.2). This instantiation is described in the following. The Algorithm 1 outlines the complete process.

3.1 Position-Specific (or Local) Learners

The position-specific learners L_p pursue their own learning steps locally using their own generated data. Their goal is to be able to decompose the contents of the data into different factors of variations, in particular, those related to the position itself. This objective of the local learner L_p can be formalized as the expected loss over the data distribution of the position p:

$$f_p(w_p) = \mathbb{E}_{\xi_p}[\tilde{f}_p(w_p; \xi_p)]$$

where ξ_p is a random data sample drawn according to the distribution of position p and $\tilde{f}_p(w_p; \xi_p)$ is a loss function corresponding to this sample and w_p the set of weights. In particular, the positions can have non-i.i.d. data distributions, i.e., the distributions of ξ_{p_i} and ξ_{p_j}, $p_i \neq p_j$, are distinct. At the step t of communication round, each local learner independently runs τ_p iterations of the local solver base on the stochastic gradient descent (SGD) starting from the current global model $L_p^{(t,0)}$ until the step $L_p^{(t,\tau_p)}$ to optimize its own local objective (see the black arrows depicted in Fig. 3).

The objective function $f_p(w; \xi_p)$ can be implemented in several ways, we chose a family of models based on VAEs to perform the separation. In a long line of research [15,19], this family of models demonstrated the ability to deal with entangled representations. The authors in [22] go further by incorporating an additional term in the objective function for a better latent space decomposition.

These important results summarized in the following are used in our objective function: The data $\mathbf{x}^i$ captured by each sensor $s_i, i \in \{1, \ldots, M\}$ are generated from two underlying factors $c = (c_1, c_2)$, the one related to the position-specific c_1 and the position-invariant components c_2. These observations are modeled using a real-valued latent/code vector $z \in \mathbb{R}^d$, interpreted as the representation of the data. The generative model is defined by the standard Gaussian prior $p(z) = \mathcal{N}(0, I)$, intentionally chosen to be a factorized distribution, and the decoder $p_\theta(x|z)$ parameterized by a neural net. The variational posterior for an observation is $q_\varphi(z|x) = \prod_{j=1}^{d} \mathcal{N}(z_j|\mu_j(x), \sigma_j^2(x))$, with the mean and variance produced by the encoder, also parameterized by a neural net. The variational posterior can be seen as the distribution of the representation corresponding to the data point x. The distribution of representations for the entire data set is then given by an inference model: $\mu = W_\mu h_\varphi^{enc}(x) + b_\mu$ and $\log \sigma = W_\sigma h_\varphi^{enc}(x) + b_\sigma$ and generative model: $z \sim q_\phi(z|x) = \mathcal{N}(\mu, \sigma)$. The loss function for training the model is given by: $\log p_\varphi(x|z) = D_{KL}(q(z|x)||p(z)) + L(\theta, \varphi; x, z)$, where $D_{KL}(||)$ stands for the non-negative Kullback–Leibler divergence between the true and the approximate posterior. Hence, maximizing $L(\theta, \varphi; x, z)$ is equivalent to maximizing the lower bound to the true objective in: $\log p_\theta(x|z) \geq L(\theta, \varphi; x, z) = \mathbb{E}_{q_\varphi}(z|x)[\log p_\theta(x|z)] - \beta D_{KL}(q_\varphi(z|x)||p(z)) - \alpha D_{KL}(q_\varphi(z)||p(z))$

3.2 Referential (or Central) Learner

Each local learner pursue its own ”version” of the universal representation z_{iA}, but has not to diverge from the *referential* universal representation z_A, which constitutes a consensus among the local learners. In our setting, we build the *referential* universal representation by making every learner contributes to it via a weighted aggregation defined in the following. Given the objectives $f_p(w)$ of the local learners L_p, the referential learner objective function is formulated as:

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) := \sum_{p=1}^{M} \alpha_p \times f_p(w_p) \right\} \; with \; \sum_{p=1}^{M} \alpha_p = 1 \tag{1}$$

Where α_p are used to weigh the contribution of every learner to the referential universal representation. After a predefined number of local update steps, we conduct a conciliation step (see the dotted red arrows in Fig. 3). Each conciliation step t produces a new version of the referential learner $L_S^{(t)}$ and, subsequently, a new version of the referential universal representation $z_A^{(t)}$. The conciliation step has to be performed on the learned representations $z_{pA}^{(t)}, p \in \{1, \ldots, M\}$ via regularization, for example. In our approach, the conciliation step is performed via representation alignment, e.g., Correlation-based alignment [2], Concatenation-based fusion [21], etc.) or minimization of information-theoretic-based quantities (e.g., mutual information neural estimation [7], etc.). More formally, we instrument the objective function of the local learners with an additional term derived from the representation alignment [29]

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \frac{1}{M} \sum_{p=1}^{M} F_p(w_p) \right\}, \; F_p(w_p) = \min_{w \in \mathbb{R}^d} \left\{ f_p(w_p) + \lambda R(z_{iA}, z_A^{(t)}) \right\} \tag{2}$$

where R is a regularization term responsible of aligning the locally learned universal components with the ones learned by the referential learner and $\lambda \in (0, 1)$ is a regularization parameter that balance between the local objective and the regularization term. Note that in this setting, it is required that a copy of the referential learner's weights, at conciliation step t, be available locally in order to perform the generative step which produces.

The conciliation process can also be carried out without this proposed alignment. Position-specific and universal components will still be learned separately but locally. Then, the conciliation can be performed via the standard FL setting where the weights of the local universal components learners are aggregated and used to update the referential learner. In this regard, the conciliation step can be implemented with any federated learning algorithm (e.g. *Federated Averaging* [23], Federated Normalized Averaging [30], etc.).

Global updates. The shared global model is updated as follows:

$$w^{(t+1,0)} - w^{(t,0)} = \sum_{p=1}^{M} \alpha_p \Delta_p^{(t)} = - \sum_{p=1}^{M} \alpha_p \cdot \eta \sum_{k=0}^{\tau_p - 1} g_p(w_p^{(t,k)}) \tag{3}$$

where $w_p^{(t,k)}$ denotes client p's model after the k-th local update in the t-th communication round and $\Delta_p^{(t)} = w_p^{(t,\tau_p)} - w_p^{(t,0)}$ denotes the cumulative local progress made by client p at round t. Also, η is the client learning rate and g_p represents the stochastic gradient over a mini-batch of B samples. When the number of clients M is large, then the central server may only randomly select a subset of clients to perform computation at each round.

Algorithm 1: Multi-level abstraction of sensor position

```
Input : (i) ${\mathbf{x}^p}_{p=1}^{M}$ streams of annotated observations generated by the data
        sources, (ii)
$w \leftarrow$ initWeights() ;   % Initialize the referential learner's weights
foreach position $p \in \mathcal{S}$ do
    $w_p \leftarrow w$ ;                                    % Weights distribution
end
while not converged do
    ; % Local updates
    foreach position $p \in \mathcal{S}$ do
        for $t \in \tau_p$ steps do
            Sample mini-batch $\{x_i^p\}_{i=1}^{n_p}$ from the stream of data $\mathbf{x}^p$
            Evaluate $\nabla_{w_p} R(w_p)$ with respect to the mini-batch
            Compute adapted parameters with gradient descent:
              $w_p^{(t)} \leftarrow w_p^{(t-1)} - \eta \nabla_{w_p} R(w_p)$
        end
    end
    ; % Central updates
    Update central model's weights $L_{\mathcal{S}}$ by aggregating the incoming weights
      from the local models $L_p, p \in \{1, \ldots, M\}$ using Eqn. 3
end
Result: $L_{\mathcal{S}}$ the trained referential learner and $L_p, p \in \{1, \ldots, M\}$ the trained
        local learners
```

4 Experiments and Results

In our experiments, we use the SHL dataset dedicated to mobility-related human activity recognition. This dataset provides multimodal and multilocation locomotion data recorded in real-life settings (as described in § 2). We perform an empirical evaluation of the proposed approach, consisting of two major stages. In the first stage, we evaluate the quality of the data separation into position-specific and universal components which is performed by the local learners and how each of these components contribute individually, with and without the conciliation process, to the recognition performances (§ 4.1); We, then, evaluate various inference configurations where the position-specific and universal components are combined together to improve the recognition performances. We also provide a comparative analysis against baselines (§ 4.2). The code to reproduce the experiments is publicly available [2].

[2] Software package and code to reproduce empirical results are publicly available at https://github.com/alphaequivalence/positionAbstraction (anonymized)

Experimental setup. We use Tensorflow [1] for building the architecture of the VAE used to model the learners in our proposed approach. This architecture is illustrated in Figure 4. As a preprocessing step, the annotated input streams from the SHL dataset are segmented into sequences of 6000 samples which correspond to a duration of 1 min. given a sampling rate of 100 Hz. To model the temporal dependencies in the considered sequences, we use long short-term memory (LSTM) cells [16]. For weight optimization, we use stochastic gradient descent with Nesterov momentum of 0.9 and a learning-rate of 0.1. Weight decay is set to 0.0001. The number of update steps τ_p performed by each local learner before the conciliation step is set to 100.

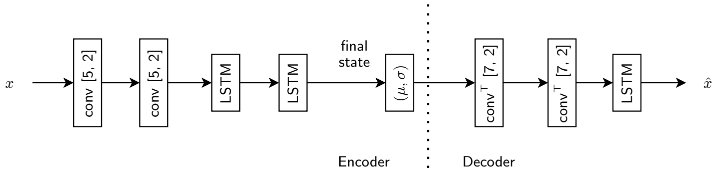


Fig. 4: Architecture of the variational autoencoder used to model the learners in our proposed approach. All convolutions are 1D with their hyperparameters (kernel size and stride) shown. These hyperparameters are further optimized. All layers are preceded by batch normalization and a ReLU activation. conv$^\top$ stands for transposed convolution. LSTM cells are used to capture the temporal dependencies in the considered sequences. The final states generated by the LSTM cell are used to model the latent distribution's mean and variance.

4.1 Evaluation of Data Decomposition Process

In this part, we evaluate the ability of the local learners to decompose the sensor data into the position-specific components and the universal ones. We evaluate this process with and without the conciliation phase, then we show the impact of this step on the recognition performances. We distinguish two possibilities to achieve this decomposition process. The first one referred to as *explicit separation*, deals with two separate vectors, one for the position-specific components and the other for the universal one. The second one uses only one representation with additional constraints, referred to as *constrained separation* as described in the VAE formulation. The main additional constraints are the independence and the sparsity (i.e. a significant proportion of the dimensions to be inactive) commonly considered to be a good evaluation criterion in the literature [22]. As in their work, we measure the sparsity of a given representation using the *Hoyer* extrinsic metric [17] which is formally defined for a vector $\mathbf{y} \in \mathbb{R}^d$ to be

$$Hoyer(\mathbf{y}) = \frac{\sqrt{d} - \|\mathbf{y}\|_1 / \|\mathbf{y}\|_2}{\sqrt{d} - 1} \in [0, 1]$$

yielding 0 for a fully dense vector and 1 for a fully sparse vector. Table 1 summarizes the average normalized sparsity of the obtained representations. Fig. 5

illustrates the average latent magnitude computed for each dimension of the learned representations.

Table 1: Summary of the per-position average normalized sparsity measured using the *Hoyer* extrinsic metric [17]. Results with and without the conciliation step are shown.

	Average normalized sparsity±std.			
Config.	*Bag*	*Hand*	*Hips*	*Torso*
w/o conciliation	0.42 ± .072	0.77 ± .002	0.71 ± .029	0.68 ± .024
w/ conciliation	0.44 ± .0145	0.91 ± .0521	0.87 ± .038	0.727 ± .033

From Table 1, we can observe, as expected, that the representations learned by the local learners of the *Hand* and *Hips* have high sparsity compared to *Bag* and *Torso*. The sparsity increases further when the conciliation is performed as the dimensions that are less important are being pushed more and more towards zero. Regarding the latent magnitudes, during conciliation, we can observe that some dimensions of the latent representation are getting more an more activated (e.g., dimensions 30, 35, 39, and 40 with an average magnitude of 0.0134, 0.146, 0.0138, and 0.138, resp.) while these same dimensions remain relatively small when there is no conciliation.

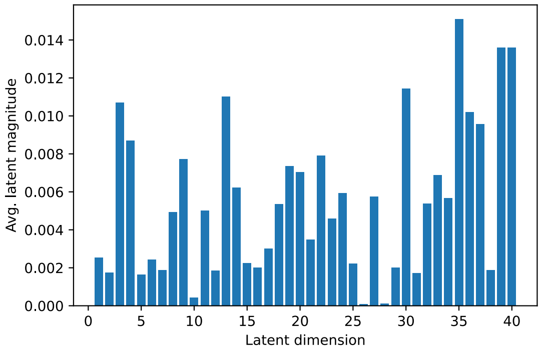


Fig. 5: Average latent encoding magnitude computed over the four positions.

Evaluation of the recognition performances As demonstrated above, the dimensions of the learned representations have meaningful interpretation with regards to the activities that we seek to recognize. To further assess the usefulness of the separated components per se (without a conciliation step), we leverage them in a traditional discriminative setting. In other words, we take the learned representation and add, on top of it, a simple dense layer. This additional layer is trained to minimize classification loss while the rest of the circuit is kept frozen. Note that the additional Dense layer has low VC-dimension so that we ensure it has no capacity to improve the representation by itself. Furthermore, to better understand how the process of conciliation among the learners, attached to the different positions, impacts the quality of both the universal and position-specific components, we leverage similarly the separated components but this time, after performing the conciliation process. Table 2, summarizes obtained results. We compare the results with baseline models trained on data generated from specific positions without applying the separation nor the conciliation processes. This configuration is referred to as *"Baseline (no sep.)"* in Table 2.

Table 2: Summary of the recognition performances obtained using either the universal or the position-specific components learned in each position by the local learners. Recognition performances with and without the conciliation process are reported. For reference, the recognition of a baseline model which do not perform separation (nor conciliation) are additionally shown.

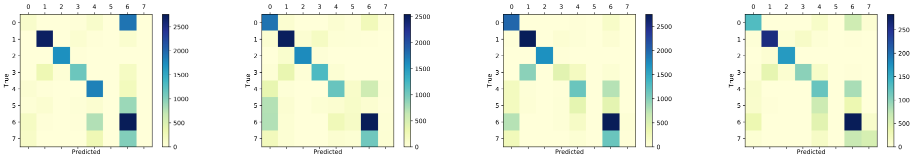
	Recognition Performances±std.			
Config.	*Bag*	*Hand*	*Hips*	*Torso*
Baseline (no sep.)	63.79 ± .0089	63.86 ± .0014	65.70 ± .0126	60.61 ± .0072
Universal comp.				
w/o conciliation	66.17 ± .0224	65.26 ± .0147	66.12 ± .0035	62.47 ± .013
w/ conciliation	66.97 ± .016	67.8 ± .0015	67.84 ± .0354	63.12 ± .01
Pos.-specific comp.				
w/o conciliation	64.2 ± .3	66.17 ± .007	67.9 ± .0026	61.32± .087
w/ conciliation	65.66 ± .029	68.94 ± .03	70.45 ± .07	61.15 ± .029

We observe from Table 2 that, overall, the recognition performances obtained using the position-specific and universal components are better than those obtained using the baseline (without separation nor conciliation). In theory, with the conciliation step, optimal representations would emerge in particular for the universal components. Indeed, this is achieved by the additional alignment term in Eq. 2 which should make them interchangeable regardless of the position from which they have been generated. This should nevertheless be harder in the case of the position-specific components which may activate very diverse dimensions of the learned representation (as described in the experimental results above). Surprisingly, this has a mild impact on the recognition performances which stay comparable. This could potentially be explained by the importance of the position-specific components for the recognition of many of the activities that are considered in the SHL dataset. It is worth noticing though that the universal components achieve remarkable improvements in the case of *Bag* and *Torso*.

4.2 Inference Configurations

Depending on the activity, the right prediction can be achieved by using either components z_A or z_{iP} taken individually, or a combination of the universal component z_A and the most appropriate position-specific component. In this part, we take a fine-grained look at the previously obtained recognition performances by assessing the optimal configuration which allows the correct prediction of each of the individual activities we are interested in. For this, we evaluate the predictions obtained using basic inference configurations, i.e., the combination of the universal components with *Torso*-specific components $[z_A; z_{Torso}]$; *Hand*-specific components $[z_A; z_{Hand}]$; *Bag*-specific components $[z_A; z_{Bag}]$; and with *Hips*-specific components $[z_A; z_{Hips}]$. Figure 6 shows the confusion matrices obtained using each of these configurations.

For reference, we compare the obtained results here with the performances of two activity recognition baselines which do not perform the separation nor conciliation steps. The architecture of these basic models consists of convolution-based circuits for each position which are then fused together and trained jointly. We implemented two types of fusion schemes: concatenation-based and alignment-based fusion. These models constitute the baselines against which we compare the recognition performances obtained using the various inference configurations. To make these baselines comparable with the models based on our proposed inference configurations, we make sure to get the same complexity, i.e. comparable number of parameters. We also use Bayesian optimization based on Gaussian processes as surrogate models to select the optimal hyperparameters of the baselines models [28]. Additional details about the baselines and the ranges of their optimized hyperparameters are available in the code repository.



(a) $z_A; z_{Torso}$ (b) $z_A; z_{Hand}$ (c) $z_A; z_{Bag}$ (d) $z_A; z_{Hips}$

Fig. 6: Confusion matrices obtained using different inference configurations. Combination of the universal components z_A and: (a) *Torso*-specific components; (b) *Hand*-specific components; (c) *Bag*-specific components; (d) *Hips*-specific components. The activities are numbered as *1:Still*, *2:Walk*, *3:Run*, *4:Bike*, *5:Car*, *6:Bus*, *7:Train*, and *8:Subway*.

Compared to the baseline models, the evaluated inference configurations yield better recognition performances in general. For example, the combination of the universal and most of the position-specific components help discriminate efficiently activities like *Walk*, *Run*, and *Bike*. On the other hand, some activities like *Car*, *Bus*, or *Train* suffer from confusion and do no show significant improvements over the baseline ($\sim 2\%$ on avg.). Also, activity *Subway* exhibits the same behavior with less proportion suggesting that this ”on-wheels” group of activities need elaborate combination of points of views as demonstrated in [25]. This issue could potentially be circumvented by using more featured inference configurations where other position-specific representations (or learners), rather than a single one, can be leveraged to infer these problematic or hard-to-infer activities.

Table 3: Summary of the evaluation of inference configurations. Recognition performances (mean and std.) of the best inference configuration is shown along with the recognition performances (mean and std.) averaged over all evaluated configurations. Evaluations repeated 7 times. The ubscripts of the position-specific representations are shortened as z_b (*Bag*), z_{ha} (*Hand*), z_{hi} (*Hips*), and z_t (*Torso*). Performance of the baseline models are also displayed.

Config.	*Best Config.*	*Recogn. Perf.±std.*	*mean±std.*
Baselines			
Concat. fusion	-	-	60.24 ± .014
Corr. Alignment	-	-	63.79 ± .032
Activities			
Still	$z_{hi}; z_t$	85.77±0.016	83.26±0.7
Walk	$z_A; z_{ha}$	88.54±0.07	86.74±0.058
Run	z_{ha}	90.51±0.016	89.46±0.03
Bike	$z_A; z_{hi}$	85.62±0.2	83.22±0.086
Car	$z_A; z_{ha}$	78.24±0.058	77.14±0.2
Bus	z_{ha}	78.08±0.022	75.17±0.004
Train	$z_{hi}; z_{hi}$	76.13±0.175	74.88±0.08
Subway	$z_A; z_{ha}; z_t$	75.89±0.009	74.07±0.006

Table 3 summarizes the evaluation results of the inference configurations featuring the combination of various position-specific components. We observe an increase in terms of the correct predictions for most of the activities compared to the previous setting. In particular, the ”on-wheels” group of activities, i.e., *car*, *bus*, *train*, and *subway*, get improved substantially. At the same time, as expected, we see now that the inference configurations, which yield the highest recognition performances for these activities, use genuine combinations like *Hand*-specific components alone in the case of *Bus* or a combination of the universal, *Hand*- and *Torso*-specific components in the case of *Subway*. On the other hand, *Still* gets the least improvement compared to the previous setting while the best configuration to infer it is a combination of the *Hand*- and *Torso*-specific components (85.77±0.016). It is worth noticing that activities like *Walk* and *Bike* still achieve competitive performances (88.54±0.07 and 85.62±0.2, resp.) while using the same inference configuration, i.e., a combination of the universal and *Hand*-specific components for *Walk* and *Hips*-specific for *Bike*, as in the previous setting. In the case of *Run*, the highest recognition performances are achieved using only the *Hand*-specific components, which supports the observations presented in § 1.

5 Summary and Future Work

In this paper, we propose an original approach for abstracting the impact of the specific position where sensory data generator are located. Our approach is based on a multi-level processing starting by the disentanglement of the position-specific and universal components. Experimental results show that it substantially improves recognition rates and has many advantages including reducing the heterogeneity of the data. The data decomposition process allows a better recognition rate in several ways. (1) by reducing the noise induced by the data linked to the position itself (e.g. the local component of the movement of the hand constitutes noise for the local component of the movement of the feet for example), (2) by aggregating only data of the same nature presenting different points of view and,(3) for certain activities, the local component alone is sufficient to ensure recognition (e.g. the movement of a hand during the race, for example).

Future work follows two axes: (1) improving the quality of the model in particular, having a fine-grained control on the data decomposition process by adding additional domain knowledge-based models, e.g., representing explicitly the dynamics of the body movements in the latent space as done in [18,31] in the case of human activity recognition case. (2) the conducted research rises interesting questions to pursue noticeably the improvement of multi-sources approaches where the various perspectives are entangled with local components, which could improve federated approaches by reducing the noise especially by sharing only the mutualisable components.

References

1. Abadi, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: ICML. pp. 1247–1255 (2013)
3. Asim, Y., et al.: Context-aware human activity recognition (cahar) in-the-wild using smartphone accelerometer. IEEE Sensors Journal **20**(8), 4361–4371 (2020)
4. Attal, F., Mohammed, S., Dedabrishvili, M., et al.: Physical human activity recognition using wearable sensors. Sensors **15**(12), 31314–31338 (2015)
5. Banos, O., Toth, M.A., Damas, M., et al.: Dealing with the effects of sensor displacement in wearable activity recognition. Sensors **14**(6), 9995–10023 (2014)
6. Barshan, B., Yurtman, A.: Classifying daily and sports activities invariantly to the positioning of wearable motion sensor units. IEEE Internet of Things J. (2020)
7. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: ICML. pp. 531–540 (2018)
8. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Computing Surveys (CSUR) **46**(3), 1–33 (2014)
9. Carollo, J.J., Worster, K., Pan, Z., Ma, J., et al.: Relative phase measures of intersegmental coordination describe motor control impairments in children with cerebral palsy who exhibit stiff-knee gait. Clinical Biomechanics **59**, 40–46 (2018)

10. Ehatisham-Ul-Haq, M., et al.: Coarse-to-fine human activity recognition with behavioral context modeling using smart inertial sensors. IEEE Access **8** (2020)
11. Förster, K., et al.: Evolving discriminative features robust to sensor displacement for activity recognition in body area sensor networks. In: ISSNIP. IEEE (2009)
12. Gjoreski, H., et al.: The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. IEEE Access (2018)
13. Hamidi, M., Osmani, A.: Data generation process modeling for activity recognition. In: ECML-PKDD. pp. 374–390. Springer (2020)
14. Hamidi, M., Osmani, A., Alizadeh, P.: A multi-view architecture for the shl challenge. In: UbiComp-ISWC '20. p. 317–322. ACM (2020)
15. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., et al.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
17. Hurley, N., Rickard, S.: Comparing measures of sparsity. IEEE Transactions on Information Theory **55**(10), 4723–4741 (2009)
18. Karl, M., Soelch, M., et al.: Deep variational bayes filters: Unsupervised learning of state space models from raw data. arXiv preprint arXiv:1605.06432 (2016)
19. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML. pp. 2649–2658 (2018)
20. Kunze, K., Lukowicz, P.: Dealing with sensor displacement in motion-based onbody activity recognition systems. In: UbiComp. pp. 20–29 (2008)
21. Li, Y., Yang, M., Zhang, Z.: A survey of multi-view representation learning. IEEE TKDE **31**(10), 1863–1883 (2018)
22. Mathieu, E., Rainforth, T., Siddharth, N., Teh, Y.W.: Disentangling disentanglement in variational autoencoders. In: ICML. pp. 4402–4412 (2019)
23. McMahan, B., Moore, E., Ramage, D., et al.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS. pp. 1273–1282 (2017)
24. Melendez-Calderon, A., Shirota, C., Balasubramanian, S.: Estimating movement smoothness from inertial measurement units. bioRxiv (2020)
25. Osmani, A., Hamidi, M., Alizadeh, P.: Hierarchical learning of dependent concepts for human activity recognition. In: PAKDD. Springer (2021)
26. Osmani, A., Hamidi, M., Bouhouche, S.: Monitoring of a dynamic system based on autoencoders. In: IJCAI. pp. 1836–1843 (2019)
27. Shi, J., Zuo, D., Zhang, Z., Luo, D.: Sensor-based activity recognition independent of device placement and orientation. Transactions on Emerging Telecommunications Technologies **31**(4), e3823 (2020)
28. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. In: NIPS. pp. 2951–2959 (2012)
29. T Dinh, C., Tran, N., Nguyen, T.D.: Personalized federated learning with moreau envelopes. NeurIPS **33** (2020)
30. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. NeurIPS **33** (2020)
31. Watter, M., Springenberg, J.T., et al.: Embed to control: a locally linear latent dynamics model for control from raw images. In: NeurIPS. pp. 2746–2754 (2015)
32. Woodworth, B.E., Patel, K.K., Srebro, N.: Minibatch vs local sgd for heterogeneous distributed learning. In: NeurIPS. vol. 33, pp. 6281–6292 (2020)
33. Yang, R., Wang, B.: Pacp: a position-independent activity recognition method using smartphone sensors. Information **7**(4), 72 (2016)