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

#417

Abstract. The relative position of sensors placed on specific body parts generates two types of data related to (1) the movement of the body part w.r.t. the body and (2) the whole body w.r.t. the environment. These two data provide orthogonal and complementary components contributing differently to the activity recognition process. In this paper, we introduce an original approach that separates these data and abstracts away the sensors' exact on-body position 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 activities abstracted from these positional biases. We perform a thorough empirical evaluation of our approach on the various datasets featuring on-body sensor deployment in real-life settings. Obtained results show substantial improvements in performances measured by the f1-score and pave the way for developing models that are agnostic to both the position of the data generators and the target users.

1 Introduction

The selection of the sensors' positions in moving targets is a constraint that is encountered in many fields, such as human activity recognition from on-body sensor deployments [6,26,5]. The movements of the area of the target on which the sensors are positioned generate data of two different but complementary natures (see Fig. 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 activity recognition, we notice for example that the kinetics of the hand movements during a race can be decomposed into a circular movement (CM) of the hand relative to the shoulder and a translation movement (TM) associated with the whole body [21]. At least three practical implications can be devised from this: (i) CM data are enough to learn some target concepts, e.g., the hand kinetics movement is enough to determine if a person is at rest or running; (ii) CM data from different positions, e.g., hand and torso, cannot be shared and mixed together. Otherwise, this generates noise and confusion during the learning process; (iii) only TM data can be shared among the different positions as these data are of the exact same nature but taken from different points of view (positions or perspectives).

In this paper, we leverage the data decomposition into universal and position-specific components to improve activity recognition models. These components have distinctive contributions concerning the target concepts to learn. This brings an interesting property that allows us to fuse the universal components

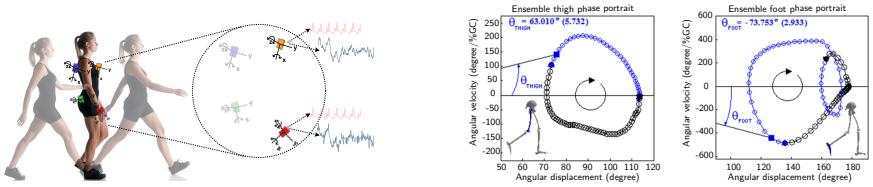

Fig. 1: (left) The hand sensor undergoes two types of movements. One is of the same nature as the torso and linked to the translational movement of the body. The other is linked to the movement of the hand locally relative to the body. (right) Phase plan showing the dynamics of the thigh and foot during gait cycle (GC) (⊸ 1%GC) extracted from the biomechanics works of [7].

as seen from different points of view (positions) while identifying the position-specific components, which could serve as additional knowledge in situations when the position-specific components are not sufficient to recognize an activity. Without this data decomposition process, the local part of the data adds position noise challenging to manage with centralized approaches, e.g., federated learning [20,30]. Indeed, to integrate data from different positions (or clients), 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 impacting the learning process.

Traditional HAR approaches [6,22,34] often consider the sensory inputs to be flattened therefore disregarding the significant impact of the various positional biases. Some approaches consider these problems from the perspective of deployment optimization, mainly focusing on the study of the optimal on-body sensors placement and its impact on the recognition of target activities [4,3,26,5]. There are also rare approaches offering pipelines which include recognition of the position of the data generator followed by the activity recognition [33] or including an explicit model of the context [9,2]. Other approaches, e.g., [17], try to develop heuristics to improve the robustness of activity recognition models to sensors displacements. Regardless of the devised techniques, these approaches rely on centralized processing of the data, which does not match the intrinsic complementary nature of the data, thus limiting their potential capacities.

To deal with these complementary data sources, we propose an original multi-level model of abstraction of the data generator position encompassing 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. They are responsible for learning (i) the position-dependent patterns of activities and (ii) the movements that link them to the individual. The aim is to abstract the learning examples from the bias arising from the position from which they are generated. The central learner (or set of local generic learners) uses the aggregated universal components from the local learners via a conciliation step based on the efficient federated learning (FL) setting. Ex-

tensive experiments on three representative datasets featuring real-world sensor deployment settings show the effectiveness of abstracting the impact of the data generator's position. We noticeably get substantial improvements in terms of the recognition performances of individual activities and robustness to the evolution of the sensor deployments. We perform a comprehensive comparative analysis of our proposed approach via ablation studies which shows the contribution of the dual interplay between the local and central learners.

2 Problem Formulation

In this section, we briefly characterize the problem of abstracting the exact position of a given sensor. 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, e.g., human body. 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*, distributed according to an unknown generative process. 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. In the case of the SHL dataset, the sensors deployment features data generated from 4 smartphones, carried simultaneously at typical body locations (*hand*, *torso*, *hips*, and *bag*).

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 the data generated by each sensor s_i, of a factorized representations z_i being a composition of (i) position-invariant (abstract or universal) components vector z_{iA}, and (ii) a position-specific (local) components vector z_{iP}. The position-invariant components vector captures the features that are shared across all positions. On the other hand, the position-specific components vector captures specific and complementary insights concerning the target concepts. The first problem to solve in our model is to build automatically this data decomposition process for each sensor automatically. Thanks to this process, 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. Components 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 performances, while traditional HAR approaches often consider the inputs to be flattened and disregard the bias related to the position. We notice that in certain situations the position-specific component alone is enough to recognize the activity, e.g., the circular movements

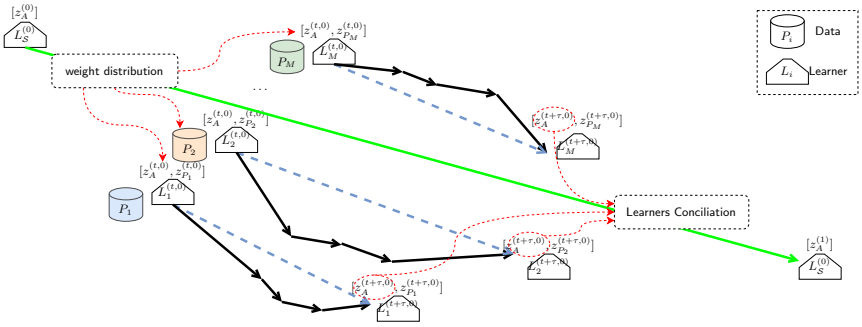

Fig. 2: Framework of the proposed multi-level abstraction architecture. The global learner $L_{\mathcal{S}}$ starts with an initial set of weights which are distributed to the local learners. The local learners L_p, one for each position p, learn the two vector components z_A and z_P, by performing independently a set of gradient steps which allows to get newer versions. These new versions are used during the conciliation step which results in a new version of the global learner, and subsequently a more robust position-independent representation.

of the hand are enough to distinguish between running and walking. In addition, since only position-independent data is shared, this process considerably 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 challenging complementary representations, we propose a model based on multi-level processing to abstract the position as described below. In this model, we suppose that the position-invariant components share the data with a central learner.

3 Source Position Multi-Level Abstraction

Here, we propose an instantiation of the proposed problem formulation composed of local and central learners. To perform the separation of the position-specific components from the universal ones, we use a family of models based on variational autoencoders (VAEs) [16] (§ 3.1). The proposed conciliation step is based on the federated learning (FL)-based aggregation 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. Fig. 2 summarizes the proposed instantiation. Algorithm 1 outlines the complete learning 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 decompose the contents of the data

into different factors of variations, particularly those related to the position itself. The 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 w_p the set of the learner's weights. In particular, the distributions from which are drawn the samples ξ_{p_i} and ξ_{p_j}, $p_i \neq p_j$, can be distinct. At the step t of communication round, each local learner independently runs τ_p iterations of the local solver, e.g., stochastic gradient descent, 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. 2).

The objective function $f_p(w_p)$ is constructed using a family of models based on VAEs for their ability to deal with entangled representations. The task here is to learn these factors of variation, commonly referred to as learning a disentangled representation. It corresponds to finding a representation where each of its dimensions is sensitive to the variations of exactly one precise underlying factor and not the others.

Depending on the availability of explicit knowledge about the underlying factors of variation, different strategies are pursued to learn the disentangled representation. For example, in video prediction [8,14], temporal-invariance is often leveraged with a content representation which captures structure that is shared across all video frames and a pose representation capturing content that varies over time. These strategies require devising complex architectures and intricate loss functions to enforce prior knowledge. Alternatively, the disentanglement can be performed using separate representations for each factor of variation, which are jointly learned by different encoders, e.g. [25,24]. Although the representations are explicitly separated and learned by different encoders, getting exact correspondence with the factors of variation, i.e., non-overlapping dimensions, is not ensured and can lead to identical representations. Recent advances in unsupervised disentangling based on VAEs demonstrated noticeable successes in many fields using the β-VAE, which leads to improved disentanglement [13]. It uses a unique representation vector and assigns an additional parameter ($\beta > 1$) to the VAE objective, precisely, on the Kullback Leibler (KL) divergence between the variational posterior and the prior, which is intended to put implicit independence pressure on the learned posterior. The improved objective becomes:

$$\mathcal{L}(x; \theta, \varphi) = \mathbb{E}_{q_\varphi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\varphi(z|x)||p(z)) - \alpha D_{KL}(q_\varphi(z)||p(z)),$$

where the term controlled by α allows to specify a much richer class of properties and more complex constraints on the dimensions of the learned representation other than independence. Indeed, the proposed conciliation step is challenging due to the dissimilarity of the data distributions across the local learners, leading to discrepancies between their respective learned representations. One way to deal with this issue is by imposing sparsity on the latent representation in a way that only a few dimensions get activated depending on the learner and activities. We ensure the emergence of such sparse representations using the appropriate

structure in the prior $p(z)$ such that the targeted underlying factors are captured by precise and homogeneous dimensions of the latent representation. We set the sparse prior as $p(z) = \prod_d (1-\gamma)\mathcal{N}(z_d; 0, 1) + \gamma\mathcal{N}(z_d; 0, \sigma_0^2)$ with $\sigma_0^2 = 0.05$. This distribution can be interpreted as a mixture of samples being either activated or not, whose proportion is controlled by the weight parameter γ [19].

3.2 Referential (or Central) Learner

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

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

where α_p is used to weigh the contribution of every learner to the universal representation. After a predefined number of local update steps, we conduct a conciliation step (see the dotted red arrows in Fig. 2). Each conciliation step t produces a new version of the referential learner $L_{\mathcal{S}}^{(t)}$ and, a new version of the referential universal representation $z_A^{(t)}$. The conciliation step has to be performed on the learned representations $z_{pA}^{(t)}$ via regularization, for example. In our approach, the conciliation step is performed via representation alignment, e.g., correlation-based alignment [1]. More formally, we instrument the objective function of the local learners with an additional term derived from the representation alignment [29].

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

where R is a regularization term responsible for aligning the locally learned universal components with the ones learned by the referential learner and $\lambda \in [0, 1]$ is a regularization parameter that balances between the local objective and the regularization term. Note that in this setting, it is required that, at conciliation step t, a copy of the referential learner's weights be available locally to perform the generative step. 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 [20], federated normalized averaging [30]. The shared global model is updated based on the federated averaging 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. η is the client learning rate and g_p represents the stochastic gradient over a mini-batch of samples.

Algorithm 1: Multi-level abstraction of sensor position

Input : $\{\mathbf{x}^p\}_{p=1}^{M}$ streams of annotated observations from the sensors

$w \leftarrow$ initWeights() ; % *Init. referential learner weights*

distributeWeights(w, $\mathcal{S}$) ; % *Weights distribution*

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 $\nabla_{w_p}\mathcal{L}(w_p)$ with respect to the mini-batch

Compute adapted parameters: $w_p^{(t)} \leftarrow w_p^{(t-1)} - \eta\nabla_{w_p}\mathcal{L}(w_p)$

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 Eq. 3

end

Result: $L_{\mathcal{S}}$ and $L_p, p \in \{1, \ldots, M\}$, the trained referential and local learners

4 Experiments and Results

We perform an empirical evaluation of the proposed approach, consisting of two major stages: (1) 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 contributes individually, with and without the conciliation process, to the recognition performances (§ 4.2); (2) we then evaluate various inference configurations where the position-specific and universal components are combined to improve the performances. We also provide a comparative analysis against baselines (§ 4.3). Code and supplementary material can be found in https://github.com/alphaequivalence/positionAbstraction (anon.)

4.1 Experimental setup

We evaluate our proposed approach on three large-scale real-world wearable benchmark datasets featuring multi-location and heterogeneous sensors: SHL [11], HHAR [28], and Fusion [27] datasets (see § A.1 for a detailed description). Implementation details can be found in § A.2. We compare our approach with the following closely related baselines.

- **DeepConvLSTM** [22]: a model encompassing 4 convolutional layers responsible of extracting features from the sensory inputs and 2 long short-term memory (LSTM) cells used to capture their temporal dependence.
- **DeepSense** [34]: a variant of the DeepConvLSTM model combining convolutional and a Gated Recurrent Units (GRU) in place of the LSTM cells.
- **AttnSense** [18]: features an additional attention mechanism on top of the DeepSense model forcing it to capture the most prominent sensory inputs both in the space and time domains to make the final predictions.

For the ablation study, we compare our approach with two baselines which do not perform the separation nor conciliation steps. These models consist 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 (see § A.3). To make these baselines comparable with our models, we make sure to get the same complexity, i.e., comparable number of parameters. We use the f1-score in order to assess performances of the architectures. We compute this metric following the method recommended in [10] to alleviate bias that could stem from unbalanced class distribution (see § C). In addition, to alleviate performance overestimation problem, we rely in our experiments on the meta-segmented partitioning proposed in [12] (see § D).

4.2 Evaluation of the 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 measure the sparsity of a given representation using the *Hoyer* extrinsic metric [15] 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 one. Table 1 summarizes the average normalized sparsity of the obtained representations. Fig. 3 illustrates the average latent magnitude computed for each dimension of the learned representations.

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*. 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, we can observe that during conciliation some dimensions of the central learner's latent representation are getting 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.) corresponding to the universal components, while the remaining dimensions having low activation and some noticeable picks (e.g., at 3, 12, 18, and 24) corresponding to the position-specific components.

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

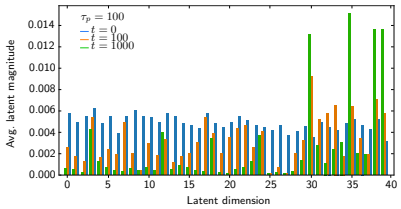

Fig. 3: Average latent encoding magnitude computed over different steps of the conciliation process.

Config.	Average normalized sparsity±std.			
	Bag	*Hand*	*Hips*	*Torso*
w/o concil.	$0.42_{\pm.072}$	$0.77_{\pm.002}$	$0.71_{\pm.029}$	$0.68_{\pm.024}$
w/ concil.	$0.44_{\pm.0145}$	$0.91_{\pm.0521}$	$0.87_{\pm.038}$	$0.727_{\pm.033}$

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

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. To alleviate any effect that could be attributed to the model's complexity, the additional dense layer has low VC-dimension so that we ensure it has no capacity to improve the representation by itself. Table 2 compares the obtained performances with the baseline models on the considered representative datasets. 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 3, 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.

Model	HHAR	Fusion	SHL
DeepConvLSTM	$70.1_{\pm.0018}$	$68.5_{\pm.002}$	$65.3_{\pm.0206}$
DeepSense	$72.0_{\pm.0022}$	$69.1_{\pm.0017}$	$66.5_{\pm.006}$
AttnSense	$76.2_{\pm.0074}$	$70.3_{\pm.0027}$	$68.4_{\pm.03}$
Feature fusion	$72.9_{\pm.004}$	$68.7_{\pm.001}$	$66.8_{\pm.009}$
Corr. align.	$75.8_{\pm.0014}$	$70.2_{\pm.04}$	$69.1_{\pm.015}$
Proposed	$78.3_{\pm.0045}$	$72.8_{\pm.002}$	$74.5_{\pm.0133}$

Table 2: Recognition performances (f1-score) of the baseline models on different representative related datasets. Evaluation based on the meta-segmented cross-validation. Experiments were averaged over 7 repetition runs.

We observe from Table 3 that, overall, the obtained performances 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). Surpris-

Config.	*Bag*	*Hand*	*Hips*	*Torso*
No sep.	63±.0089	63±.0014	65±.0126	60±.0072
Universal				
w/o concil.	66±.0224	65±.0147	66±.0035	62±.013
w/ concil.	66±.016	67±.0015	67±.0354	63±.01
Pos.-specific				
w/o concil.	64±.3	66±.007	67±.0026	61±.087
w/ concil.	65±.029	68±.03	70±.07	61±.029

Table 3: Performances obtained using either the universal or the position-specific components.

Class	Best Config.	Overall
Still	$z_{hi}; z_t$(85.77)	83.26±0.7
Walk	$z_A; z_{ha}$(88.54)	86.74±0.058
Run	z_{ha}(90.51)	89.46±0.03
Bike	$z_A; z_{hi}$(85.62)	83.22±0.086
Car	$z_A; z_{ha}$(78.24)	77.14±0.2
Bus	z_{ha}(78.08)	75.17±0.004
Train	$z_{hi}; z_{hi}$(76.13)	74.88±0.08
Subway	$z_A; z_{ha}; z_t$(75.89)	74.07±0.006

Table 4: Per-class performances obtained using various inference configurations.

ingly, this has a mild impact on the 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*.

4.3 Inference Configurations

Here we evaluate the robustness of the proposed approach to the evolution of the sensors deployments via the flexibility that it offers for the inference step. 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 performances by assessing the optimal configuration which allows the correct prediction of each of the individual activities. For this, we evaluate the predictions obtained using basic inference configurations, i.e., the combination of the universal components with *torso* $[z_A; z_t]$; *hand* $[z_A; z_{ha}]$; *bag* $[z_A; z_b]$; and *hips*-specific $[z_A; z_{hi}]$ components. Compared to the baseline models, the evaluated inference configurations yield better 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 (approx. 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 [23]. This issue could potentially be circumvented by using more featured configurations where other position-specific representations, rather than a single one, can be leveraged to infer these problematic or hard-to-infer activities.

Table 4 summarizes the evaluation results of the inference configurations featuring the combination of various position-specific components. We observe an increase in 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 configurations, which yield the highest performances for

these activities, use genuine combinations like z_A alone in the case of *bus* or a combination of z_A, z_{ha}, and z_t 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 z_{ha} and z_t (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 z_A and z_{ha} for *walk* and z_{hi} for *bike*, as in the previous setting. For *run*, the highest scores are achieved using only z_{ha}, which supports the observations presented in § 1.

5 Summary and Future Work

This paper proposes an original approach for abstracting the impact of the specific position of the sensory data generators. Our approach is based on multi-level processing, starting with the disentanglement of the position-specific and universal components at a local level and the conciliation of the universal components at a global level. Experimental results show that the proposed approach improves recognition rates and has many advantages, including reducing the data sources' heterogeneity impact. The decomposition process allows a better recognition rate in several ways: (i) 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; (ii) by aggregating only data of the same nature presenting different points of view and; (iii) for certain activities, the local component alone is sufficient to ensure recognition, e.g., hand movement during run. Future work follows two axes. (1) Improving the quality of the model, in particular, having a fine-grained control on the data decomposition process using additional domain knowledge, e.g., expliciting the dynamics of the body movements in the latent space like in [31,7]. (2) Improving federated multi-source approaches where the sources are entangled with local components. Sharing only mutualisable components has a promising potential.

References

1. Andrew, G., et al.: Deep canonical correlation analysis. In: ICML (2013)
2. Asim, Y., et al.: Context-aware human activity recognition (cahar) in-the-wild using smartphone accelerometer. IEEE Sensors Journal **20**(8), 4361–4371 (2020)
3. Attal, F., Mohammed, S., Dedabrishvili, M., et al.: Physical human activity recognition using wearable sensors. Sensors **15**(12), 31314–31338 (2015)
4. 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)
5. Barshan, B., Yurtman, A.: Classifying daily and sports activities invariantly to the positioning of wearable motion sensor units. IEEE Internet of Things J. (2020)
6. 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)
7. 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)
8. Denton, E.L., Birodkar, V.: Unsupervised learning of disentangled representations from video. In: NIPS (2017)

9. Ehatisham-Ul-Haq, M., et al.: Coarse-to-fine human activity recognition with behavioral context modeling using smart inertial sensors. IEEE Access **8** (2020)
10. Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies. ACM SIGKDD **12**(1), 49–57 (2010)
11. Gjoreski, H., et al.: The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. IEEE Access (2018)
12. Hammerla, N.Y., Plötz, T.: Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In: UbiComp'15. pp. 1041–1051 (2015)
13. Higgins, I., et al.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)
14. Hsieh, J.T., et al.: Learning to decompose and disentangle representations for video prediction. arXiv preprint arXiv:1806.04166 (2018)
15. Hurley, N., Rickard, S.: Comparing measures of sparsity. IEEE Transactions on Information Theory **55**(10), 4723–4741 (2009)
16. Kingma, D., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
17. Kunze, K., Lukowicz, P.: Dealing with sensor displacement in motion-based onbody activity recognition systems. In: UbiComp. pp. 20–29 (2008)
18. Ma, H., Li, W., Zhang, X., Gao, S., Lu, S.: Attnsense: Multi-level attention mechanism for multimodal human activity recognition. In: IJCAI. pp. 3109–3115 (2019)
19. Mathieu, E., Rainforth, T., Siddharth, N., Teh, Y.W.: Disentangling disentanglement in variational autoencoders. In: ICML. pp. 4402–4412 (2019)
20. McMahan, B., Moore, E., Ramage, D., et al.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS. pp. 1273–1282 (2017)
21. Melendez-Calderon, A., Shirota, C., Balasubramanian, S.: Estimating movement smoothness from inertial measurement units. bioRxiv (2020)
22. Ordóñez, F.J., Roggen, D.: Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors **16**(1), 115 (2016)
23. Osmani, A., Hamidi, M., Alizadeh, P.: Hierarchical learning of dependent concepts for human activity recognition. In: PAKDD. Springer (2021)
24. Qian, H., et al.: Latent independent excitation for generalizable sensor-based cross-person activity recognition. In: AAAI. vol. 35, pp. 11921–11929 (2021)
25. Sadeghi, M., et al.: Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM TASLP **28**, 1788–1800 (2020)
26. Shi, J., Zuo, D., Zhang, Z., Luo, D.: Sensor-based activity recognition independent of device placement and orientation. Transactions on ETT **31**(4), e3823 (2020)
27. Shoaib, M., Bosch, S., et al.: Fusion of smartphone motion sensors for physical activity recognition. Sensors **14**(6), 10146–10176 (2014)
28. Stisen, A., et al.: Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In: SenSys'15. pp. 127–140 (2015)
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)
34. Yao, S., et al.: Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In: WWW'17. pp. 351–360 (2017)