Investigating Domain-agnostic Performance in Activity Recognition using Accelerometer Data

APINAN HASTHANASOMBAT, University of Cambridge, UK
ABHIRUP GHOSH, University of Cambridge, UK
DIMITRIS SPATHIS, University of Cambridge, UK
CECILIA MASCOLO, University of Cambridge, UK

Human activity recognition (HAR) models suffer significant performance degradation when faced with data heterogeneity (device, users, environments) at test time. Current approaches to this problem using domain adaptation or transfer learning attempt to improve performance in one specific target domain, often using data from said domain. Requiring access to data from the target domain is limiting and cannot be generally assumed. In addition, there is often no single target domain, but rather multiple ones arising from different sources of data heterogeneity. One way to achieve good performance in this setting would be to gather data from all potential domains the model may encounter at deployment - this is generally infeasible.

This work presents the case for training models which are *domain-agnostic*, i.e., that generalise to unseen test domains. This requires a new way to evaluate models; we discuss a regime called *leave-datasets-out*, and present a starting benchmark for HAR using binary classification. Two state-of-the-art deep models in the literature are tested; they significantly under-perform in unseen domains when compared to their performance on seen domains. It is shown that under this new evaluation regime, a simple model with an appropriate inductive bias performs at least as well as two current deep models on the benchmark, with a p-value of 5.75×10^{-4} and 0.13 when testing for a difference in mean accuracy, whilst being at least 10 times faster to train. Additionally, we provide evidence that domain diversity under certain conditions improves performance on both *seen and unseen* domains. We hope this work provides useful insights to further develop HAR models suitable for real world deployment.

CCS Concepts: • **Computing methodologies** → *Cross-validation*; **Neural networks**; **Supervised learning by classification**.

Additional Key Words and Phrases: Human activity recognition; Generalization; Robustness

ACM Reference Format:
Apinan Hasthanasombat, Abhirup Ghosh, Dimitris Spathis, and Cecilia Mascolo. 2022. Investigating Domain-agnostic Performance in Activity Recognition using Accelerometer Data. In *Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp/ISWC '22 Adjunct), September 11–15, 2022, Cambridge, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3544793.3560398

1 INTRODUCTION

Human Activity Recognition (HAR) using accelerometer sensors have been researched since at least 2004 [3]. The models have moved from hand-crafted features [20, 23], to end-to-end deep models using convolutional and long-short-term memory neural architectures [16, 17]. In contrast, the evaluation setup has not significantly evolved. The models are often tested on the same dataset used for training [10, 12, 13, 19, 27].

This evaluation setup does not accurately reflect the performance of real deployments where training and testing data may be significantly different. This issue has been observed in computer vision where performance

UbiComp/ISWC '22 Adjunct, September 11–15, 2022, Cambridge, United Kingdom

ACM ISBN 978-1-4503-9423-9/22/09.
https://doi.org/10.1145/3544793.3560398

degrades with variations in object pose [1], light [8] and weather [25]. In the HAR context, it is acknowledged in literature that models suffer with changes in users [22], sensors [16] and environments [14]. Testing on the same dataset assumes that these aspects of the training data remain constant during deployment.

Current approaches related to test-time heterogeneity in HAR uses transfer learning or domain adaptation techniques. There have been studies for transferring between users, sensor location, sensor modalities and datasets [5, 9, 16, 18, 28]. In general these approaches use a source dataset $\mathcal{D}_s$ in combination with a small subset, either labelled or unlabelled, from the target (test) domain $\mathcal{D}_t$ for training. A common motivation for these transfer studies is to reuse knowledge gained from a model trained on the source domain, due to limited access to labelled data from the target domain.

The main assumption, inherent in these approaches, is that the researcher has access to data from the target domain. This implies that models are domain specific; for each new target domain, retraining is required. However, it is likely that models will face unseen domains during real deployment, as collecting data for all potential domains is infeasible. Instead, we should strive for models to be *domain-agnostic*, i.e., perform well on seen domains but also generalise to unseen domains, under appropriate conditions. Being able to measure this performance will help researchers build better domain-agnostic models.

This work proposes using a different evaluation setup that may better reflect performance on real deployments, an extension of the leave-one-subject-out regime to the dataset level - leave-datasets-out. An instance of this evaluation method for the task of HAR is given as a binary classification task between two common activities across three openly available HAR datasets. Using this benchmark it is shown that two current state-of-the-art (SotA) deep models [7, 17] face significant performance degradation in unseen domains, even after correcting for factors such as sensor location, sampling rate,e and measurement units.

We show that under this benchmark, a simple model using an appropriate inductive bias based on our understanding of the data generating mechanism performs at least as well ($p=5.75x10^{-4}$ and 0.13), compared to SotA end-to-end deep learning models, whilst requiring significantly less resources to train. This unexpected result raises questions about using deep end-to-end models as a one-size-fits-all solution in applications with small labelled datasets such as HAR when unseen domain performance is key.

The work makes two further observations, one related to domain-agnostic models, and another about existing transfer techniques. First, it is shown that achieving consistent gains across both seen and unseen domains, across all tested models, is possible when training with data from multiple domains under similar conditions. This observation can be useful not only in improving domain-agnostic performance, but also in our understanding of negative transfer [26]. Additionally, a domain's performance improves without complex transfer or adaptation techniques when a small subset of data is available from said domain, across all tested models. This raises questions about our understanding of existing transfer techniques: do the gains come from the method or from the additional data?

Contributions. In summary, the contributions of this work are the following:

(1) A starting benchmark based on a simple binary classification which measures HAR models' performance on unseen domains corresponding to the leave-datasets-out evaluation regime. This serves as a better proxy to performance in real deployment.
(2) Demonstrate that a model with a simple inductive bias can perform at least as well as current deep models on this new benchmark. This raises further questions about our understanding of deep models in HAR.
(3) Two observations. 1. That performance on both seen and unseen domains improves with multiple domain training under certain conditions (where the additional domain is *not* included in the seen or unseen set). 2. That if the additional domain is already seen, as in the transfer learning setup, this improves the performance without any complex transfer technique, raising questions about our understanding of gains from these methods.

2 DOMAIN-AGNOSTIC PERFORMANCE

Problem Setup. We are interested in the case where we have accelerometer data for a particular participant $\mathbf{x} \in \mathcal{X}$ and activity labels $y \in \mathcal{Y}$. We assume to have access to n datasets $\mathcal{D}_1, \mathcal{D}_2, ...\mathcal{D}_n$ each corresponding to the same activities but in a different domain (corresponding to a different distribution). Each dataset consists of m_k pairs $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)_{i=1}^{m_k}\}$, where each pair corresponds to data from participant i, which in turn is assumed to be independently and identically distributed samples from the corresponding domain. The feature space $\mathcal{X}$ and label space $\mathcal{Y}$ are the same across all datasets. In our experiments, there are three datasets ($n = 3$) with 7, 10 and 9 participants respectively ($m_1 = 7$, $m_2 = 10$ and $m_3 = 9$).

In this work, we often refer to datasets used for training and testing in the following way. Let $\boldsymbol{\mathcal{D}_{tr}}$ denote the set of training dataset(s), and likewise $\boldsymbol{\mathcal{D}_{te}}$ for the testing dataset(s)[1]. For instance the n datasets can be partitioned into two groups, $\boldsymbol{\mathcal{D}_{tr}}$ and $\boldsymbol{\mathcal{D}_{te}}$. The goal is to only use $\boldsymbol{\mathcal{D}_{tr}}$ to train a model that will perform well on $\boldsymbol{\mathcal{D}_{te}}$. i.e., we want to minimise $\mathbb{E}_{\mathbf{x},y\sim\boldsymbol{\mathcal{D}_{te}}}\mathcal{L}(M, (\mathbf{x}, y))$ whilst only having access to $\boldsymbol{\mathcal{D}_{tr}}$, where $\mathcal{L}$ is some loss function, M is the trained model, and $\mathbf{x}, y$ is the data.

Difference to domain adaptation. We note the difference between the setup just described to the typical domain adaptation or transfer leaning setup where there is a designated source $\mathcal{D}_s$ and target $\mathcal{D}_t$ domain, usually corresponding to two different datasets. A model is trained on $\mathcal{D}_s$ and then adapted to work on the target using a subset of data from $\mathcal{D}_t$ [5, 9, 16, 18, 28] i.e., $\boldsymbol{\mathcal{D}}_{tr} = \{\mathcal{D}_s\}$ and $\boldsymbol{\mathcal{D}}_{te} = \{\mathcal{D}_t\}$. The main difference here being that we are *not* interested in the performance of any one particular domain $\mathcal{D}_t$, but rather the performance in domains where the model has not seen any data (i.e., not $\mathcal{D}_t$ or $\mathcal{D}_s$), in addition to the domains where it has already seen data. Related benchmarks have been studied in computer vision under domain generalisation [11]. We also note that a similar setup has been studied previously [24], but not using end-to-end deep learning models, which have become the dominant model paradigm today, and using only single dataset training.

Throughout the discussion in this paper, it is often useful to refer to a hypothesised data generating mechanism for HAR data, this is shown in Figure 1. Let $\mathbf{x}$ denote the observed data for a particular domain, a node represents a variable, or a group thereof, and an arrow from node A to be B signifies that A influences the value of B in the data.

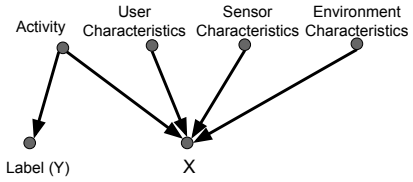


Fig. 1. A possible data generating mechanism for HAR data.

2.1 Measuring domain-agnostic performance

In traditional learning, k-fold cross validation (CV) is often used to estimate the true error of the model (defined as the loss over the unknown distribution the data was drawn from) by taking the average of the loss of each fold. Denote the partitions $1, 2, ..., k$ of a training dataset $\mathcal{D}$ as $\mathcal{D}^1, \mathcal{D}^2, ...\mathcal{D}^3$. The model in the ith fold is trained using all partitions except $\mathcal{D}^i$, i.e., $\boldsymbol{\mathcal{D}_{tr}} = \{\bigcup_{j\neq i} \mathcal{D}^j\}$, and tested on partition i, $\boldsymbol{\mathcal{D}_{te}} = \{\mathcal{D}^i\}$. Denote by $M(\boldsymbol{\mathcal{D}_{tr}})$ a model trained with datasets in $\boldsymbol{\mathcal{D}_{tr}}$. Let $\mathcal{L}(M(\boldsymbol{\mathcal{D}_{tr}}), \boldsymbol{\mathcal{D}_{te}}))$ denote the loss of a model trained on $\boldsymbol{\mathcal{D}_{tr}}$ and tested on $\boldsymbol{\mathcal{D}_{te}}$ for some loss function $\mathcal{L}$. The overall error in k-fold CV of a model M is then approximated by:

[1]A bold font is used to denote a set of datasets whereas a normal font is used when referring to single datasets.

$$\text{Error}(M) = \frac{1}{k} \sum_{i \in 1,\ldots,k} \mathcal{L}(M(\bigcup_{j \neq i} \mathcal{D}^j), \mathcal{D}^i)$$

using a single dataset $\mathcal{D}$. In the context of timeseries analysis, especially in HAR, a variant called leave-one-subject-out CV is often used. This is to avoid the same portion of data appearing in both the training and testing sets, due to the way the timeseries from each participant is split into samples using overlapping windows.

Leave-datasets-out (LDO) cross-validation. In this work, we use a natural extension of this idea to measure domain-agnostic performance, called leave-datasets-out CV. In the simplest setting, given n datasets $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n$, the domain-agnostic error is approximated by:

$$\text{Error}(M) = \frac{1}{n} \sum_{i \in 1,\ldots,n} \sum_{j \in 1,\ldots,n} \mathcal{L}(M(\mathcal{D}_i), \mathcal{D}_j) \tag{1}$$

where if $i = j$ then leave-one-subject-out CV is used, and when $i \neq j$ the model is trained on $\mathcal{D}_i$ and tested on the full dataset $\mathcal{D}_j$. In a later section of this work we will consider the case where we train on multiple datasets instead of a single D_i. This captures the idea that we are interested in the performance of the model on the collection of datasets which could have resulted from the same mechanism, such as the one shown previously.

A starter LDO benchmark for HAR. As there are many possible variations in the user, sensor and environment characteristics in the real-world [6], starting simple before moving on to more complex scenarios will help us understand model failures and hence how to improve them. In this work we use three datasets with as many similar characteristics in the generating mechanism as possible, and two activities which are common across all datasets. This can tell us whether current models are able to deal with a smaller subset of heterogeneity in data.

In particular, we use three open HAR datasets which share the walking and stair climbing activities, where the data was collected under controlled environments, and where the sensor was worn on the same body position. This leaves heterogeneity in the user, which is expected in real-world deployments, and any other heterogeneity in the sensors that are not related to its placement. If we are unable to perform well with these more restrictive heterogeneity, then it is worthwhile to understand why before moving on to tackle more complex scenarios, such as location independent models [5] and scenarios with several activities.

Datasets. There are three datasets, MHEALTH [2], PAMAP2 [21], and WHARF [4], which contain data from sensors located on the right wrist of the participants. There are only two overlapping activities across all datasets: walking, and ascending stairs.

Preprocessing. For each dataset, samples were filtered for the two common activities (walking, stairs), and only for readings captured from a sensor on the right-wrist of the participant. Invalid values and anomalies were removed. The time-series for each participant was normalised to a common sampling rate of 50Hz, amplitude normalised, and values converted to a common unit (ms^2). Any participants with corrupted data is discarded. The subject-timeseries is then segmented into 5 second windows (250 samples at 50Hz) with an overlap of 2.5 seconds (125 samples at 50Hz).

Current model performance. We test two state-of-the-art deep neural network models from the literature. One is attributed to [7], a convolutional model, which has shown to consistently outperform in a standardised test [15]. Another is the DeepConvLSTM model which uses both convolutional and LSTM layers [17]. Since current models in the literature are multiclass classifiers, i.e., they are able to distinguish between many different activites, it is expected that they should perform well on a binary classification task.

The model was evaluated according to equation 1. The loss function used is the average binary classification accuracy with a 0.5 threshold. The performance of the two SotA deep models are shown in the first two left violin plots, labelled DeepConvLSTM and DeepConv, in Figure 2a. Each datapoint in the plot is the average accuracy of training the model on dataset i and testing the model on dataset j. Given the three datasets used, there are a total

of 9 combinations. If $i = j$ then normal leave-one-subject-out CV is used. If $i \neq j$ then the full dataset j is used for testing and this is repeated 10 times.

3 IMPROVING DOMAIN-AGNOSTIC PERFORMANCE

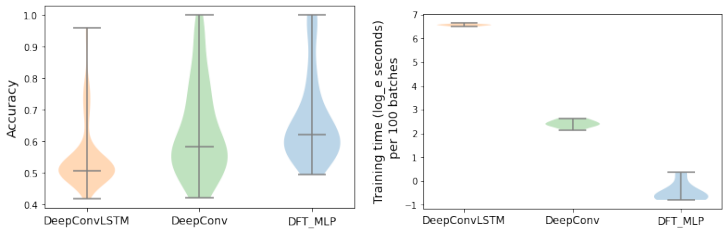
Revisiting fundamentals. By considering the data generating mechanism (Figure 1) we can see that the observed data can be influenced by a number of different factors other than the activity performed by the user. This raises an important point: models can easily be fooled by confounding factors which may be predictive of the activity label.

Let us take a concrete example. It may be the case that in one particular dataset, data collection for the walking activity was performed only on the elderly, whereas in more strenuous activities, such as running, data was collected on younger participants. This would suggest that a model would, in theory, be able learn to discriminate the walking activity by only using user characteristics that are present in elderly participants. When using this model on a different dataset where walking may also be performed by younger participants, the model would face performance degradation.

The two simplest ways to reduce the likelihood that a model is fooled by confounding factors is to reduce the size of the hypothesis class, and by incorporating the researcher's knowledge about the problem in the form of an inductive bias. In this particular case, based on our understanding of human activity, we know that motions associated with an activity is performed at a relatively low frequency i.e., at most a couple of times per second. We further know that we are not so interested in features that do not affect the general shape of the motion, such as the amplitude, since the general shape is what determines the activity rather than the range in which they are performed.

As the simplest implementation of this idea, the discrete fourier transform (DFT) power spectrum of low frequencies was used as features through a multi-layer perceptron (MLP) network. Albeit its simplicity, it fulfils the two criteria: reducing the hypothesis class, and incorporating an inductive bias.

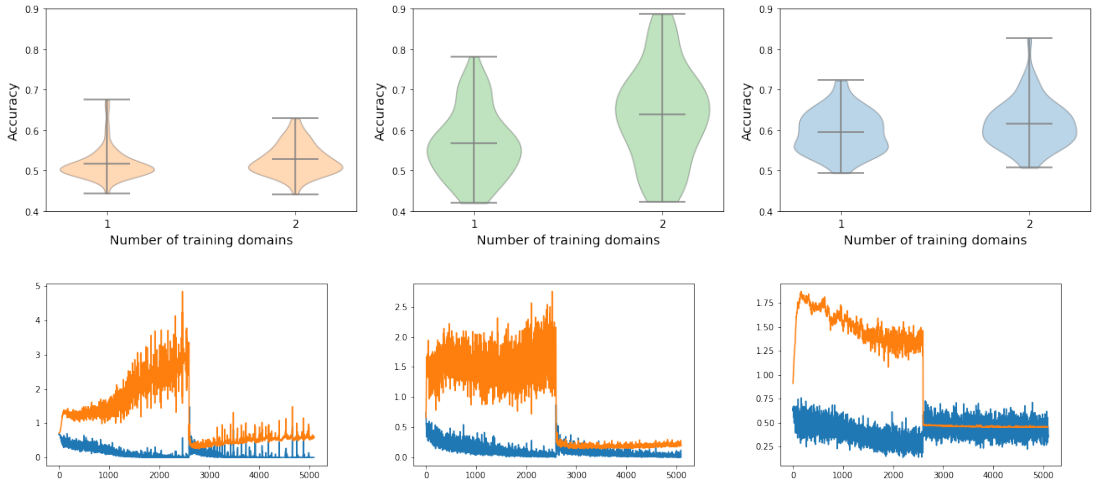
Simple model LDO results. This simplified model was compared with the deep models using the LDO benchmark; results are shown in the rightmost plot of Figure 2a, labelled DFT_MLP, and a comparison of (log) training time is shown in Figure 2b, using commodity hardware on an Intel Core i7-8650. A two-sided statistical test to detect whether the average performance of the DFT-MLP model is different to the DeepConvLSTM and DeepConv model yields a p-value of $5.75\text{x}10^{-4}$ and 0.131 respectively. The difference in training time across all models are statistically significant at a 0.01 threshold level.



(a) (b)

Fig. 2. (a) Average accuracy on the benchmark of each model. Each datapoint in the plot is the average accuracy of training the model on dataset i and testing the model on dataset j - a total of 9 combinations. If $i = j$ then normal leave-one-subject-out CV is used. If $i \neq j$ then the full dataset j is used for testing and this is repeated 10 times. (b) Training time for each model on commodity hardware, note y-axis is a log scale.

Using more than one domain. If our assumption that the domains are connected by the same data generating mechanism is true, we should in theory improve domain-agnostic performance by training on more than one domain. This section briefly investigates this idea.



(a) DeepConvLSTM (b) DeepConv (c) DFT-MLP

Fig. 3. Top: Overall performance on the *unseen* dataset based on training with one or two domains across all considered models. Bottom: All models see a noticeable drop in validation loss (orange) on the original training domain $\mathcal{D}_{tr,1}$, when a small sample of data from an additional domain $\mathcal{D}_{tr,2}$ is introduced. Training loss is shown in blue. The training and validation loss is based only on data from $\mathcal{D}_{tr,1}$.

Setup. The models were trained and evaluated according to Eq. 1 as before. However, instead of using only one training dataset $\mathcal{D}_i$, two were used, and we are interested only in the performance on the unseen dataset. i.e., $\mathcal{D}_{tr} = \{\mathcal{D}_{tr,1}, \mathcal{D}_{tr,2}\}$. For the first training dataset $D_{tr,1}$, the full dataset is used, then a small random sample of 128 windows is selected from three random participants in the second dataset $D_{tr,2}$ and included into the training data halfway through training time. Performance is then measured on the remaining (third) unseen dataset. The reason for such a small sample of the second training domain is to see the effect of performance based on data diversity rather than the effect of data quantity.

The results comparing single domain to two domain training is shown in the top portion of Figure 3. The left violin in each plot shows average accuracy from training with a single domain, similar to the previous setup, but where performance is shown only on the unseen dataset. The right violin shows average accuracy by training with two domains, $D_{tr,1}$ and $D_{tr,2}$, and testing on the remaining unseen dataset. A similar statistical test is performed across all models to test whether the average performance using one or two domains are different with p-values of 0.139, $5.75\text{x}10^{-4}$ and $6.75\text{x}10^{-2}$ for DeepConvLstm, DeepConv and the DFT-MLP model respectively.

Additionally, it is interesting to note that when the small sample from the second domain ($\mathcal{D}_{tr,2}$) is introduced we see a noticeable drop in validation loss in the original training domain, $\mathcal{D}_{tr,1}$. This is shown in the bottom portion of Figure 3. This suggests that the additional data diversity provided by the second domain also increases performance not only in unseen domains, but also in seen domains (excluding itself). We additionally observe that performance in its own domain also improve (as expected), but without using any transfer technique.

4 CONCLUSION

Limitations. The proposed benchmark is constrained under a number of dimensions. A more realistic version of the task would be multiclass over as many datasets as possible. Few publicly available HAR datasets have overlapping classes with the same sensor placement. This paper is nonetheless a start, as models which perform well on multiclass should also do well in the simpler binary classification setting. In the future we hope that this benchmark can be extended when more data is available to the community.

The aim of this work was to present the case for training models which are domain-agnostic i.e generalise to unseen domains of the same activity. This will bring us closer to robust real-world deployment of HAR models. To do so we have presented three main points.

First, we proposed using the leave-datasets-out cross-validation regime, which we argue is a better way to measure domain-agnostic performance than current evaluation methods. We present a starting point of this in the HAR context using a binary classification across three publicly open HAR datasets. We evaluate current state-of-the-art deep models for HAR, and find that they face significant performance degradation when tested against this new benchmark. We show that by using a simple inductive bias from our knowledge of the problem, we can instead use a model that achieves similar, if not better performance than current deep models ($p=5.75x10^{-4}$ and 0.131) and that is 10-100 times faster to train. Finally, we show that training with even a small amount of data from an additional domain improves performance on unseen, seen (excluding the same domain), and in the same domain without complex transfer or adaptation techniques, across all models considered.

These results suggest that end-to-end deep models may not always be a one-size-fits-all solution in HAR applications where large-scale training data is hard to come by, when deployment is likely to be on resource constrained devices, and where real deployed models face multiple sources of heterogeneity.

ACKNOWLEDGMENTS

A. Hasthanasombat is supported by the Cambridge Trust and King's College Cambridge. A. Ghosh, D. Spathis and C. Mascolo are supported by ERC through project 833296 (EAR). C.Mascolo is additionally supported by Nokia Bell Labs.

REFERENCES

[1] Michael A. Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 2019. Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4840–4849. https://doi.org/10.1109/CVPR.2019.00498 ISSN: 2575-7075.

[2] Oresti Banos, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014. mHealthDroid: A Novel Framework for Agile Development of Mobile Health Applications. In *Ambient Assisted Living and Daily Activities (Lecture Notes in Computer Science)*, Leandro Pecchia, Liming Luke Chen, Chris Nugent, and José Bravo (Eds.). Springer International Publishing, Cham, 91–98. https://doi.org/10.1007/978-3-319-13105-4_14

[3] Ling Bao and Stephen S. Intille. 2004. Activity Recognition from User-Annotated Acceleration Data. In *Pervasive Computing (Lecture Notes in Computer Science)*, Alois Ferscha and Friedemann Mattern (Eds.). Springer, Berlin, Heidelberg, 1–17. https://doi.org/10.1007/978-3-540-24646-6_1

[4] Barbara Bruno, Fulvio Mastrogiovanni, Antonio Sgorbissa, Tullio Vernazza, and Renato Zaccaria. 2013. Analysis of human behavior recognition algorithms based on acceleration data. In *2013 IEEE International Conference on Robotics and Automation*. 1602–1607. https://doi.org/10.1109/ICRA.2013.6630784 ISSN: 1050-4729.

[5] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A Systematic Study of Unsupervised Domain Adaptation for Robust Human-Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (March 2020), 1–30. https://doi.org/10.1145/3380985

[6] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2020. Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities. *arXiv:2001.07416 [cs]* (Jan. 2020). http://arxiv.org/abs/2001.07416 arXiv: 2001.07416.

[7] Yuqing Chen and Yang Xue. 2015. A Deep Learning Approach to Human Activity Recognition Based on Single Accelerometer. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*. 1488–1492. https://doi.org/10.1109/SMC.2015.263

[8] Dengxin Dai and Luc Van Gool. 2018. *Dark Model Adaptation: Semantic Image Segmentation from Daytime to Nighttime*. Technical Report arXiv:1810.02575. arXiv. https://doi.org/10.48550/arXiv.1810.02575 arXiv:1810.02575 [cs] type: article.

[9] Martin Gjoreski, Stefan Kalabakov, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2019. Cross-dataset deep transfer learning for activity recognition. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. ACM, London United Kingdom, 714–718. https://doi.org/10.1145/3341162.3344865

[10] Yu Guan and Thomas Plötz. 2017. Ensembles of Deep LSTM Learners for Activity Recognition using Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (June 2017), 11:1–11:28. https://doi.org/10.1145/3090076

[11] Ishaan Gulrajani and David Lopez-Paz. 2020. In Search of Lost Domain Generalization. *arXiv:2007.01434 [cs, stat]* (July 2020). http://arxiv.org/abs/2007.01434 arXiv: 2007.01434.

[12] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi. 2015. Multi-modal Convolutional Neural Networks for Activity Recognition. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*. 3017–3022. https://doi.org/10.1109/SMC.2015.525

[13] Nils Y Hammerla, Shane Halloran, and Thomas Plotz. 2016. Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables. *IJCAI* (2016), 8.

[14] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom '18)*. Association for Computing Machinery, New York, NY, USA, 289–304. https://doi.org/10.1145/3241539.3241548

[15] Artur Jordao, Antonio C. Nazare Jr., Jessica Sena, and William Robson Schwartz. 2019. Human Activity Recognition Based on Wearable Sensor Data: A Standardization of the State-of-the-Art. *arXiv:1806.05226 [cs]* (Feb. 2019). http://arxiv.org/abs/1806.05226 arXiv: 1806.05226.

[16] Francisco Javier Ordóñez Morales and Daniel Roggen. 2016. Deep convolutional feature transfer across mobile activity recognition domains, sensor modalities and locations. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers (ISWC '16)*. Association for Computing Machinery, New York, NY, USA, 92–99. https://doi.org/10.1145/2971763.2971764

[17] Francisco Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (Jan. 2016), 115. https://doi.org/10.3390/s16010115

[18] Xin Qin, Yiqiang Chen, Jindong Wang, and Chaohui Yu. 2019. Cross-Dataset Activity Recognition via Adaptive Spatial-Temporal Transfer Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (Dec. 2019), 148:1–148:25. https://doi.org/10.1145/3369818

[19] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal Deep Learning for Activity and Context Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4 (Jan. 2018), 157:1–157:27. https://doi.org/10.1145/3161174

[20] Nishkam Ravi. 2005. Activity Recognition from Accelerometer Data. (2005), 6.

[21] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring. In *2012 16th International Symposium on Wearable Computers*. 108–109. https://doi.org/10.1109/ISWC.2012.13 ISSN: 2376-8541.

[22] Seyed-Ali Rokni, Marjan Nourollahi, and Hassan Ghasemzadeh. 2018. Personalized Human Activity Recognition Using Convolutional Neural Networks. *AAAI* (2018), 2.

[23] Jianqiang Shen. 2004. Machine Learning for Activity Recognition. (2004), 15.

[24] Pekka Siirtola, Heli Koskimäki, and Juha Röning. 2018. Experiences with Publicly Open Human Activity Data Sets - Studying the Generalizability of the Recognition Models:. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science and Technology Publications, Funchal, Madeira, Portugal, 291–299. https://doi.org/10.5220/0006553302910299

[25] Georg Volk, Stefan Müller, Alexander von Bernuth, Dennis Hospach, and Oliver Bringmann. 2019. Towards Robust CNN-based Object Detection through Augmentation with Synthetic Rain Variations. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. 285–292. https://doi.org/10.1109/ITSC.2019.8917269

[26] Jindong Wang, Vincent W. Zheng, Yiqiang Chen, and Meiyu Huang. 2018. Deep Transfer Learning for Cross-domain Activity Recognition. In *Proceedings of the 3rd International Conference on Crowd Science and Engineering (ICCSE'18)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3265689.3265705

[27] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition. *IJCAI* (2015), 7.

[28] Jiachen Zhao, Fang Deng, Haibo He, and Jie Chen. 2021. Local Domain Adaptation for Cross-Domain Activity Recognition. *IEEE Transactions on Human-Machine Systems* 51, 1 (Feb. 2021), 12–21. https://doi.org/10.1109/THMS.2020.3039196 Conference Name: IEEE Transactions on Human-Machine Systems.