

Unsupervised Activity Recognition using Trajectory Heatmaps from Inertial Measurement Unit Data

Orhan Konak^a, Pit Wegner, Justin Albert^b and Bert Arnrich^c

*Digital Health – Connected Healthcare,
Hasso Plattner Institute, University of Potsdam, Potsdam, Germany*

Keywords: Activity Recognition, Unsupervised Learning, Inertial Measurement Units.

Abstract: The growth of sensors with varying degrees of integration and functionality has inevitably led to their entry into various fields such as digital health. Here, sensors that can record acceleration and rotation rates, so-called Inertial Measurement Units (IMU), are primarily used to distinguish between different activities, also known as Human Activity Recognition (HAR). If the associations of the motion data to the activities are not known, clustering methods are used. There are many algorithmic approaches to identify similarity structures in the incoming sensor data. These differ mainly in their notion of similarity and grouping, as well as in their complexity. This work aimed to investigate the impact of transforming the incoming time-series data into corresponding motion trajectories and trajectory heatmap images before forwarding it to well-known clustering models. All three input variables were given to the same clustering algorithms, and the results were compared using different evaluation metrics. This work shows that transforming sensor data into trajectories and images leads to a significant increase in cluster assignment for all considered models and different metrics.


1 INTRODUCTION


Human Activity Recognition (HAR), i.e., categorizing physical movements into different activity classes, is becoming increasingly popular in healthcare. In the field of digital health, it holds tremendous potential, such as in the prevention of diseases, the analysis of movements over time according to specific disease progression, the correct execution of activities, and tasks requiring documentation. The emergence of this research field is favored using increasingly accurate and small, and thus portable, Inertial Measurement Units (IMUs) to measure acceleration and angular rate over time. Carrying the sensors on specific body locations leads to certain patterns in the time series, which can then be differentiated into activity classes. Furthermore, in contrast to, e.g., video-based activity recognition, sensor-based classification offers the advantage of privacy protection. This fact makes sensor-based activity classification the preferred method, especially in areas of sensitive data such as healthcare. Methodologically, sensor-based activity recognition


is closely intertwined with the field of machine learning.

Machine learning for activity recognition can be roughly divided into two branches, supervised and unsupervised learning. While the respective input data is labeled in supervised learning, it is unavailable in unsupervised learning. Besides the missing label, the search for unknown similarity patterns in the data is also relevant in unsupervised learning. This allows the data to be divided into clusters of similar patterns. Research in clustering often plays out at improving model-based solution approaches given the same input data. The transformation of the input data with subsequent clustering, on the other hand, is less researched (Ariza Colpas et al., 2020).

Previous work has already shown that transforming the incoming time-series data from IMUs into a motion trajectory and further 2D heatmap image can help to improve the classification accuracy for small datasets in a supervised manner (Konak et al., 2020). The methodological and data basis for the data transformation and the comparison of the results is provided by (Huang et al., 2018). Building upon these works, the contribution of this work is to compare the effect of transforming the incoming IMU data into different modalities on the clustering result.

^a  <https://orcid.org/0000-0003-1884-8029>

^b  <https://orcid.org/0000-0002-6121-792X>

^c  <https://orcid.org/0000-0001-8380-7667>

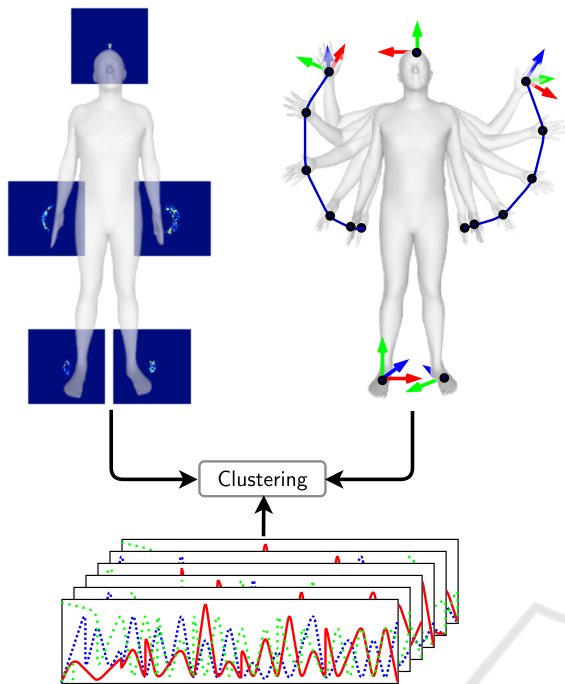


Figure 1: Clustering and evaluation of the three different input modalities: raw sensor data (bottom), trajectory time series (upper right), and heatmap images (upper left).

For this purpose, we make use of the Deep Inertial Poser (DIP-IMU) dataset, the currently largest publicly available IMU dataset (Huang et al., 2018), and classical clustering algorithms for the three modalities (1) raw input data from six sensors placed on different body regions, (2) motion mapping 3D trajectories of the sensors over time, and (3) 2D heatmaps of the motion trajectories for the respective sensors, by using various performance metrics, as shown in Figure 1.

The remainder of this work proceeds as follows: In section 2, the work is placed in the field of previous research on the effect of sensor modality transformation for unsupervised HAR. Section 3 is concerned with the algorithmic approach to data transformation and the methodology used for this study. The results are subject of section 4. The findings are further discussed in section 5. Finally, the conclusions are part of section 6.

2 RELATED WORK

Research on HAR based on IMU data with no ground truth is mainly focused on the comparison of different clustering techniques (Chen et al., 2021). There are only a few works that researched the impact of data transformation on the clustering outcome.

The idea to project human activities into an embedding space in which similar activities are located more closely was proposed in (Sheng and Huber, 2020). Using subsequent clustering algorithms can benefit from the embeddings that represent the distinct activities performed by a person. The evaluation was carried out on three labeled benchmark datasets. With improved performance in grouping the underlying human activities compared to unsupervised methods applied directly to the original dataset, they showed the framework’s capability.

(Bai et al., 2019) proposed a deep learning variational autoencoder activity recognition model for the representation of the activities in distinct time periods. By applying the proposed method on a publicly available dataset, they showed that transforming the IMU data (raw accelerometer and gyroscope data) to an encoded 128-dimensional vector has led to an improvement in grouping the activities. Three traditional clustering methods were used for evaluation.

Similarly, (Abedin et al., 2020) proposed a deep learning paradigm for unsupervised activity representations for sensory data with strong semantic correspondence to different human activities. Comparisons were made with closely related approaches, including traditional clustering methods for three diverse HAR datasets, and the effectiveness of the proposed approach could be shown. The proposed method is inspired by techniques, which are more common in image clustering (Xie et al., 2016; Min et al., 2018; Li et al., 2018).

Although these methods achieve good results, they share the common idea of a lower-dimensional representation of the data. The architecture aims to reproduce the input through a bottleneck of the target dimensionality, letting the network decide how best to compress the given information into a latent space (Wang et al., 2015). The method is often used for image feature extraction because convolutional layers are particularly powerful in detecting significant structures (Bishop, 2007). In contrast, and to the best of our knowledge, this is the first work examining the impact of data transformation from IMUs into interpretable motion trajectories and 2D heatmap images before feeding it to the clustering algorithm.

3 METHODS

For this work, the clustering evaluation has been performed for three kinds of input data: raw sensor data, trajectory time series, and trajectory images. Also, for each type of data, Principal Component Analysis (PCA) was applied to observe the effect of dimen-

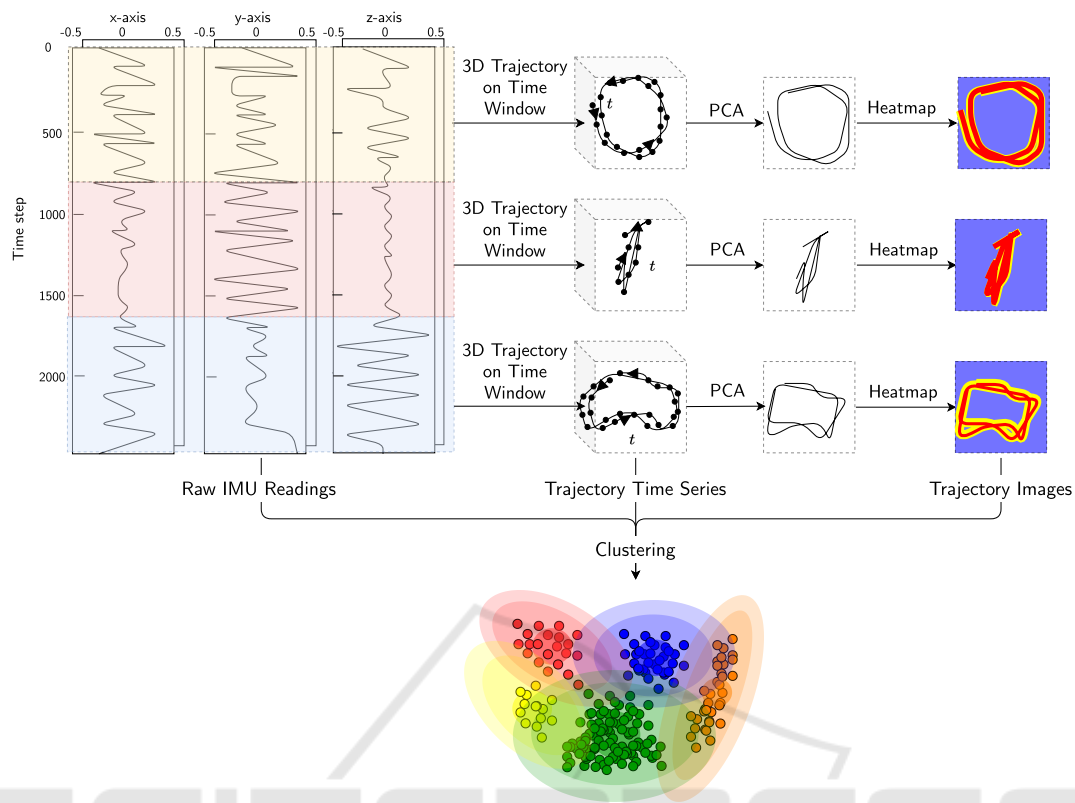


Figure 2: Overview: *Top*: Implementation pipeline of modality transformation from IMU data for given time windows into 3D trajectories and further 2D heatmap images via deep inertial poser. *Bottom*: Clustering for the three different input modalities raw sensor data, trajectory time series, and trajectory images.

sionality reduction on the clustering results (Pearson, 1901). As depicted in Figure 2 the whole pipeline starts with raw IMU readings. For short time windows, the data is transformed into motion trajectories, which are an exact reproduction of the sequence of the movements with the help of a Skinned Multi-Person Linear Model (SMPL), a realistic 3D model of the human body based on skinning and blend shapes (Loper et al., 2015). The last step consists of reducing the dimensionality of the trajectory and highlighting more frequently passed pixels in a heatmap (trajectory image). All three modalities are passed to the clustering algorithms k-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for comparison with different distance and evaluation metrics (Lloyd, 2019; Ester et al., 1996). In the following, the methodological approach is described in more depth.

3.1 Dataset

In order to evaluate the impact of the proposed approach, the largest publicly available dataset DIP-IMU was used (Huang et al., 2018). The dataset con-

sists of 17 IMU readings, containing approximately 90 minutes of real IMU data in conjunction with ground-truth poses for ten subjects in 64 sequences with 330,000 time instants and 13 different activities. The performed activities are listed in Table 1.

3.2 Data Transformation

Transforming IMU readings into corresponding motion trajectories is challenging as it requires an algorithmic detour because of the given error drift coming along with sensors (Konak et al., 2020). Therefore, we use SMPL, a skinned vertex-based model that represents a wide variety of body shapes. (Huang et al., 2018) have shown that a Bidirectional Recurrent Neural Networks (BiRNN) with Long Short-Term Memory (LSTM) cells can map the IMU readings consisting of acceleration and orientation onto the SMPL pose parameters.

The SMPL model is composed of $r = 6890$ vertices in three dimensions, which leads to a vector representation of dimension \mathbb{R}^{3r} . For activity recognition, the recorded time range denoted as S , is divided into smaller time windows S_i . In each time window i ,

Input : $S = \{S_1, \dots, S_m\}$ - List of m equally sized time windows where each time window consists of k time series of measurement values $S_i = \{S_{i,1}, \dots, S_{i,k}\}$;
 $A \in \mathbb{N}^{m \times l}$ - List of activity labels $\{a_1, \dots, a_l\}$ for each time window;
 $vertices = 1, \dots, r$ - List of vertices to track, e.g., wrist, ankle;
 $d \in \mathbb{N}$ - Dimension for heatmap;
 $n \in \mathbb{N}$ - Time frames in each time window

Output : Cluster for different input signals

Procedure:
 Initialize $SMPL \in \mathbb{R}^{r \times n \times 3}$, $T \in \mathbb{R}^{vertices \times n \times 3}$, $T' \in \mathbb{R}^{vertices \times n \times 2}$, $H \in \mathbb{R}^{m \times vertices \times d \times d}$
for $i \leftarrow 1$ **to** m **do**
 $SMPL \leftarrow PoseEstimation(S_i)$;
 $T \leftarrow SMPL^{r=vertices}$;
 $T' \leftarrow PCA(T, n_components = 2)$;
 $H_i \leftarrow Heatmap(T')$;
end
return $Cluster(\{S_i, T', H, PCA(S_i, n_components = 2), PCA(T', n_components = 2), PCA(H, n_components = 2)\})$

Algorithm 1: Clustering on the DIP-IMU dataset for different input modalities.

Table 1: Description of activities performed.

Category	# Frames	Minutes
Motion		
Upper Body	116,817	32.45
Arm Chest Crossings		
Arm Circles		
Arm Head Crossings		
Arm Raises		
Arm Stretches Up		
Lower Body	70,743	19.65
Leg Raises		
Squats		
Lunges		
Locomotion	73,935	20.54
Walking		
Sidesteps		
Crosssteps		
Freestyle	18,587	5.16
Jumping Jacks		
Sumos		

a pose estimation is predicted for n time frames by the BiRNN for the incoming acceleration and orientation data $\mathbf{a}, \boldsymbol{\omega}$:

$$f : IMU(\mathbf{a}, \boldsymbol{\omega}) \rightarrow SMPL$$

After reconstructing 3D human body poses from IMU readings in each time frame, the position of an arbitrary point over time can be tracked, hence allowing to reconstruct the motion trajectory T .

Further, PCA is applied on the resulting trajectory for dimensionality reduction from $\mathbb{R}^3 \rightarrow \mathbb{R}^2$. To preserve temporal relation, a heatmap $\mathbf{H} \in \mathbb{R}^{d \times d}$ of height and width d is generated. The sum $h_{ij} =$

$\sum_{i,j} PCA$ determines each pixel's color of the trajectories projection into the lower 2-dimensional space boundaries of the pixel. With each modality's first two principal components - trajectory heatmap images, trajectory time series, and IMU readings - the data was passed to different clustering algorithms for further processing.

3.3 Clustering

As described in algorithm 1, the whole pipeline from data acquisition to clustering consists of different algorithmic layers in a prescribed order. To examine the effect of data transformation, we used three different clustering techniques for all three modalities and different evaluation metrics. As a well-known clustering algorithm, k-means clustering was applied to determine the activities classification for the given number of clusters, and the distance measures euclidean distance and Dynamic Time Warping (DTW) (Müller, 2007). DTW aims to find a non-linear mapping between two time-series of different lengths.

Contrary to k-means clustering, DBSCAN was used as a second algorithm, which does not require a predetermined number of clusters as parameter input. Instead, it relies on connectivity between points in a particular range. Hence, more densely connected areas are grouped in one cluster, while outliers are detected as noise. The algorithm thus requires two parameters, a threshold range for the examined neighborhood of a selected sample (ϵ) and a measure for density, i.e., the number of points in that range (ρ) (Ester et al., 1996). Like other clustering algorithms, the distance measure can be chosen arbitrarily, applying it to various input types, such

as n-dimensional data points, time series, or images. We used accuracy, Adjusted Rand Index (ARI), and Adjusted Mutual Information (AMI) for the evaluation. Since DIP-IMU also incorporates the ground truth data, the accuracy can be determined by calculating the proportion of correct predictions among the total number of the examined cases. Furthermore, we used AMI and ARI as a metric (Vinh et al., 2009). AMI is defined as:

$$AMI(U, V) = \frac{MI(U, V) - E\{MI(U, V)\}}{\max\{H(U), H(V)\} - E\{MI(U, V)\}}$$

where MI denotes the mutual information between two partitions, $E\{MI(U, V)\}$ the expected mutual information between two random clusterings, and $H(U), H(V)$ the entropies associated with the partitionings U, V . Using the permutation model, the ARI Index is calculated as follows:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}$$

where n_{ij}, a_i, b_j are values from the contingency table, a table where the overlap between two groupings can be summarized.

Finally, we made use of a clustering technique that relies on neural networks. Most neural network architectures require manually labeled data, limiting their use in unsupervised scenarios. However, convolutional layers pose a robust tool for pattern recognition and feature extraction in image segmentation and classification. Invariant Information Clustering (IIC) makes these applicable to the unsupervised domain by generating artificially transformed images of the training data as paired samples for similar information in the image (Ji et al., 2019). Via convolutional representation learning, the mutual information is extracted to a latent space representing the cluster probability distribution. With each epoch, the representation is refined, and the optimal cluster mapping is extracted.

4 EVALUATION

Unsupervised activity clustering enormously benefits from the image representation to measure intra-cluster similarity. Using trajectory images for unsupervised classification significantly improves classification accuracy over raw sensor data for the k-means and DBSCAN algorithms.

Table 2: Clustering results for k-means performed on the DIP-IMU dataset for different input modalities and distance metrics.

Input modality	Quality metric		
	AMI	ARI	Acc
Raw sensor data			
Euclidean	0.135	0.045	0.220
DTW	0.239	0.033	0.280
Raw sensor data (2D)			
Euclidean	0.099	0.011	0.194
Trajectory time series			
Euclidean	0.375	0.167	0.314
DTW	0.462	0.298	0.408
Trajectory time series (2D)			
Euclidean	0.322	0.177	0.328
Trajectory image			
Euclidean	0.462	0.264	0.366
Trajectory image (2D)			
Euclidean	0.396	0.230	0.377

To evaluate this unsupervised approach to the HAR problem, three different models, k-means, ARI-optimized DBSCAN, and AMI-optimized DBSCAN were trained on DIP-IMU in six different formats using all applicable distance metrics. The formats include the raw sensor data, the generated trajectory time series, trajectory heatmap images, and the first two principal components for each. The input dimensions for the clustering algorithms are as follow: raw sensor inputs from six IMUs for acceleration and rotation in three dimensions and for a time window of 5s with 60 Hz leads to input size of $6 \times 6 \times 300$; trajectory time series with the positional coordinate for all six sensors in each time frame produces the dimension $6 \times 3 \times 300$; the trajectory image of size $6 \times 64 \times 64$ for all six sensors. Additionally, for comparison purposes, the metrics of a random classifier and the accuracy of an IIC network trained on the trajectory images were recorded.

As can be seen in Table 2, the DTW distance metric produces the best clustering results for time series data, compared to euclidean distance. This holds for both k-means and DBSCAN clustering. Also, both clustering algorithms profit from the representation as trajectories, improving all three observed metrics. The transformation to trajectory images leads to slightly worse k-means results than the trajectory time series but still outperforms the clustering on raw sensor data. PCA, for dimensionality reduction, does not provide additional value, as almost all measured evaluation metrics are lower than their original counterpart.

The DBSCAN algorithms optimized by either ARI or AMI show similar results regarding the dif-

Table 3: Clustering results for AMI-optimized DBSCAN on the DIP-IMU dataset for different input modalities and distance metrics.

Input modality	Quality metric			# of clusters	Parameters	
	AMI	ARI	Acc		ϵ	ρ
Distance metric						
Raw sensor data						
Euclidean	0.143	0.066	0.256	13	9.0	2
DTW	0.247	0.077	0.188	13	4.0	2
Raw sensor data (2D)						
Euclidean	0.122	0.033	0.206	9	1.4	4
Trajectory time series						
Euclidean	0.278	0.104	0.221	13	1.0	2
DTW	0.352	0.129	0.320	13	0.6	8
Trajectory time series (2D)						
Euclidean	0.323	0.180	0.292	14	0.4	5
Trajectory image						
Euclidean	0.171	0.053	0.243	13	1.6	4
Trajectory image (2D)						
Euclidean	0.363	0.243	0.360	12	0.2	7

Table 4: Clustering results for ARI-optimized DBSCAN on the DIP-IMU dataset for different input modalities and distance metrics.

Input modality	Quality metric			# of clusters	Parameters	
	AMI	ARI	Acc		ϵ	ρ
Distance metric						
Raw sensor data						
Euclidean	0.077	0.087	0.138	13	8.0	1
DTW	0.186	0.160	0.057	13	4.0	1
Raw sensor data (2D)						
Euclidean	0.095	0.058	0.148	28	0.6	3
Trajectory time series						
Euclidean	0.246	0.175	0.168	13	1.0	1
DTW	0.322	0.223	0.176	13	0.4	1
Trajectory time series (2D)						
Euclidean	0.283	0.215	0.152	248	0.2	1
Trajectory image						
Euclidean	0.086	0.105	0.188	13	1.2	1
Trajectory image (2D)						
Euclidean	0.362	0.253	0.366	13	0.2	6

ferences in input modality and distance metric. Each step in the transformation (sensor data \rightarrow trajectory time series \rightarrow trajectory images) leads to improved clustering results across all distance and quality metrics, as can be seen from the data in Table 3 and Table 4. In particular, the clustering of the dimensionality-reduced trajectory images performed best, in contrast to the poor results achieved on the full images. In general, it can be noted that the number of DBSCAN-found clusters often closely matched the number of ground truth classes. Comparing the results of the two different optimizations, the AMI-maximization leads to higher accuracy and more stable cluster numbers.

The differences between all three clustering results on the two principal components of the image dataset and a complete overview of 2D clustering results are shown in Figure 3. It is noticeable that the distribution of the first two principal components of the different activities are widely spread and grouped in clusters for the trajectory images. The data from different activities are cluttered for the first two principal components of the raw sensor data.

A classifier choosing randomly from the given number of classes achieved an accuracy of 12.8% and AMI and ARI scores very close to zero. An IIC network trained on the trajectory images achieved an accuracy of 35.8%.

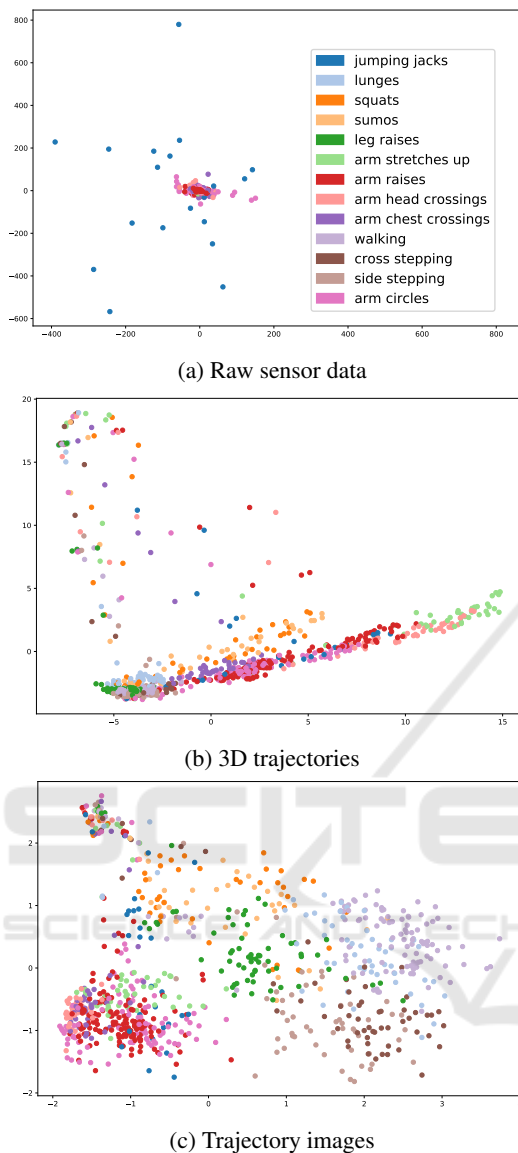


Figure 3: Plot of the top two principle components of the different data inputs. The ground truth classes are represented by distinct colors. Each type of data leads to a different spread.

5 DISCUSSION

In this work, an unsupervised learning approach for HAR based on sensor data was presented. The higher performance of the DTW distance metric over euclidean distance is expected for time series data, as similar activities are performed at different speeds by different subjects, which DTW can compensate for by mapping similar sections.

The significant difference between the results of clustering the full trajectory images and their two

principal components is most likely due to the improved sample distribution, as shown in Figure 3. Looking at the labels, the principle components can separate the arm from leg activities and overlapping patches of activities within these areas. The failure cases of the clustering algorithms are mainly due to this cluster overlap, which conventional clustering methods cannot detect. As visible in Figure 4, this problem increases in severity for trajectories and even more for raw sensor data.

The increased performance of trajectories over raw sensor data can most likely be attributed to the fact that the fixed-size 3D body model in some way standardizes the paths. Thus, similar motions create trajectories in similar regions, regardless of the subject's size or physique. It also appears that the missing temporal information in the trajectory images is not critical for unsupervised classification performance, leading to the conclusion that the computationally expensive DTW distance metric on time series could be replaced with euclidean distance of trajectory images for some use cases.

Comparing the k-means and DBSCAN models, it is interesting to note that they perform very differently on the trajectory image representation but not on their two principal components. Although DBSCAN finds a similar number of clusters as the ground truth classes, they do not necessarily represent each other. Also, they do not correspond to the clusters found by k-means, which is why the number of clusters can be outruled as a reason for this finding. The reason might be that k-means clustering creates more evenly shaped clusters on an almost evenly spaced dataset, whereas DBSCAN would group areas of similar density. In the case of trajectory images, euclidean distance is calculated in $64 \times 64 = 4096$ dimensions, whereas trajectories pose 12 and raw sensor data 60 dimensions. In high dimensions, the ratio between the nearest and farthest points approaches 1, i.e., the points essentially become uniformly distant from each other (Aggarwal et al., 2001), making clustering much more difficult for DBSCAN.

The accuracy of the IIC model is similar to the best results of both DBSCAN and k-means on the trajectory images. This implies that the features extracted by the convolutional layers do not provide much additional information for better clustering. Comparing the accuracy rates across all models and input modalities, ARI-maximizing DBSCAN performs by far the worst, often only slightly outperforming a random classifier.

6 CONCLUSION AND FUTURE WORK

The effect of using trajectories over raw sensor data in unsupervised classification for HAR is striking. Clustering algorithms using either trajectory time series or trajectory images outperform the sensor-based variants. The potential power of unsupervised classification in activity recognition for videos was already indicated by (Niebles et al., 2008). The approaches used in such methods could be powerful tools for trajectory image clustering and should be explored further. Slightly different sensor setups or using sensors from a different manufacturer can be achieved through transfer learning from the original synthetic Archive of Motion Capture as Surface Shapes (AMASS) dataset (Mahmood et al., 2019). Furthermore, euclidean distance in high-dimensional space should be mitigated, for example, by using L norms (Aggarwal et al., 2001).

REFERENCES

- Abedin, A., Motlagh, F., Shi, Q., Rezatofighi, H., and Ranasinghe, D. (2020). Towards Deep Clustering of Human Activities from Wearables. In *Proceedings of the 2020 International Symposium on Wearable Computers*, ISWC '20, page 1–6, New York, NY, USA. Association for Computing Machinery.
- Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Van den Bussche, J. and Vianu, V., editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Ariza Colpas, P., Vicario, E., De-La-Hoz-Franco, E., Pineres-Melo, M., Oviedo-Carrascal, A., and Patara, F. (2020). Unsupervised Human Activity Recognition Using the Clustering Approach: A Review. *Sensors*, 20(9).
- Bai, L., Yeung, C., Efstratiou, C., and Chikomo, M. (2019). Motion2Vector: Unsupervised Learning in Human Activity Recognition Using Wrist-Sensing Data. 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*, UbiComp/ISWC '19 Adjunct, page 537–542, New York, NY, USA. Association for Computing Machinery.
- Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition.
- Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2021). Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.*, 54(4).
- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press.
- Huang, Y., Kaufmann, M., et al. (2018). Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15. Two first authors contributed equally.
- Ji, X., Henriques, J. F., and Vedaldi, A. (2019). Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9865–9874.
- Konak, O., Wegner, P., and Arnrich, B. (2020). IMU-Based Movement Trajectory Heatmaps for Human Activity Recognition. *Sensors*, 20(24).
- Li, F., Qiao, H., and Zhang, B. (2018). Discriminatively Boosted Image Clustering with Fully Convolutional Auto-Encoders. *Pattern Recognition*, 83:161–173.
- Lloyd, S. (2019). Least Square Quantization in PCM. In *Proceedings of the IEEE Transactions on Information Theory*, volume 28, pages 129–137.
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16.
- Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*, pages 5442–5451.
- Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., and Long, J. (2018). A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture. *IEEE Access*, 6:39501–39514.
- Müller, M. (2007). Dynamic Time Warping. *Information Retrieval for Music and Motion*, 2:69–84.
- Niebles, J., Wang, H., and Fei-Fei, L. (2008). Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. *International Journal for Computer Vision*, 3(79):299–318.
- Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2:559–572.
- Sheng, T. and Huber, M. (2020). Unsupervised Embedding Learning for Human Activity Recognition Using Wearable Sensor Data.
- Vinh, N. X., Epps, J., and Bailey, J. (2009). Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 1073–1080, New York, NY, USA. Association for Computing Machinery.
- Wang, Y., Yao, H., and Zhao, S. (2015). Auto-Encoder Based Dimensionality Reduction. *Neurocomputing*, 184.

APPENDIX

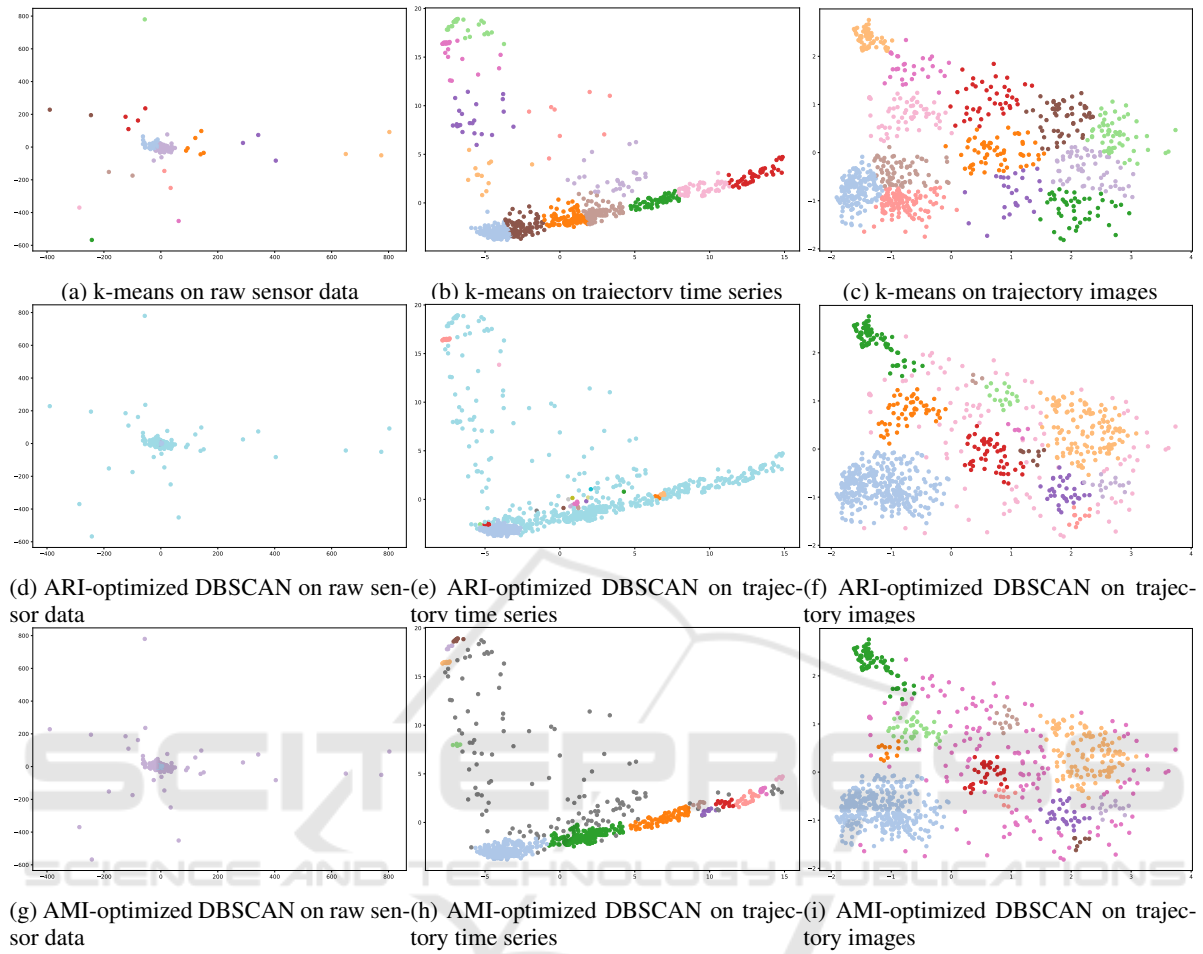


Figure 4: Clustering results on the top two principle components across all tested datasets and clustering algorithms.