Interpretable High-Level Features for Human Activity Recognition

Yale Hartmann[1][a], Hui Liu[1][b], Steffen Lahrberg[1][c] and Tanja Schultz[1][d]
[1]*Cognitive Systems Lab, University of Bremen, Bibliothekstraße 1, 28359 Bremen, Germany*

Keywords: High-level Feature, Human Activity Recognition, Hidden Markov Model, Motion Unit, Few-Shot Learning, Wearable Sensors

Abstract: This paper introduces and evaluates a novel way of processing human activities based on unique combinations of interpretable categorical high-level features with applications to classification, few-shot learning, as well as cross-dataset and cross-sensor comparison, combination, and analysis. Feature extraction is considered as a classification problem and solved with Hidden Markov Models making the feature space easily extensible. The feature extraction is person-independently evaluated on the CSL-SHARE and UniMiB SHAR datasets and achieves balanced accuracies up from 96.1% on CSL-SHARE and up to 91.1% on UniMiB SHAR. Furthermore, classification experiments on the separate and combined datasets achieve 85% (CSL-SHARE), 65% (UniMiB SHAR), and 74% (combined) accuracy. The few-shot learning experiments show potential with low errors in feature extraction but require further work for good activity classification. Remarkable is the possibility to attribute errors and indicate optimization areas easily. These experiments demonstrate the potential and possibilities of the proposed method and the high-level, extensible, and interpretable feature space.

1 INTRODUCTION

When researchers get involved in Human Activity Recognition (HAR), an important research topic to facilitate today's modern life, they will usually incorporate it into mature Machine Learning (ML) algorithms for related tasks, such as automatic segmentation, feature extraction and selection, activity modeling technology, among others. In these, only the research involving ML technology is considered, and many early and subsequent works important to HAR, such as equipment, signal acquisition, Digital Signal Processing (DSP), system evaluation, customization, and practical applications omitted. Therefore, the research goals are often to answer these two questions: Which ML approaches are applicable in HAR? How to adjust the model topologies and parameters to improve the recognition rate of certain tasks?

Many pieces of literature provide solutions to the first question. For example, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Hidden Markov Models (HMMs), and other well-known ML algorithms have shown their capability of solving problems for segmentation, feature study, activity modeling, and other related aspects (Oniga and Sütő, 2014) (Ordóñez and Roggen, 2016) (Yang et al., 2015) (Singh et al., 2017) (Ronao and Cho, 2014). Upgraded architectures for HAR based on these fundamental algorithms, such as Residual Neural Networks (ResNets) (Tuncer et al., 2020) and Hierarchical Hidden Markov Models (HHMM) (Youngblood and Cook, 2007), have also been emerging. Based on the above-listed methods, aiming to improve research results via model improvement, parameter adjustment, and experimental iterations, more works have been proposed to solve the second problem (Ronao and Cho, 2016) (Arifoglu and Bouchachia, 2017) (Uddin et al., 2011) (Rebelo et al., 2013) (Amma et al., 2010).

However, if comparing the rich outcomes of HAR research so far to other ML-based recognition research fields, an aspect of HAR that needs further exploration emerges. In Automatic Speech Recognition (ASR), human vocalization is widely studied and applied to modeling (Schultz, 2002). For example, researchers apply the three-state HMM-based Bakis-model, constructing phonemes by imitating phonetics in segmenting the pronunciation (Bakis, 1976). Each state (begin/middle/end), also called sub-phoneme, models parts of a phoneme, not only enhancing training efficiency and recognition performance but also endowing the model with phonetic and biologic significance. Another well-known example is that in the field of image and video recognition, many ap-

[a] https://orcid.org/0000-0002-1003-0188
[b] https://orcid.org/0000-0002-6850-9570
[c] https://orcid.org/0000-0002-1714-2529
[d] https://orcid.org/0000-0002-9809-7028

proaches are inspired and analogized by the physiological model of how human eyes deal with stereo vision, i.e., from human vision to computer vision (Dickinson et al., 2009) (Scheirer et al., 2014). In contrast, HAR's research so far seems to have been gathering in Artificial Intelligence (AI) itself, and there is very little literature extending the study to somatology. In fact, the concept of "human activity" specifically is practical and essentially linked to physiology and sports science. In order to integrate related knowledge into the research field of computer science, gaps need to be bridged. Once these boundaries are crossed, the HAR research will have a solid foundation for theoretical modeling and experimental optimization than purely discussing the meaning of the topology and parameters in ML models. One such step has been made in the introduction of Motion Units, where human activities have been partitioned into their distinct phases and states/sub-phases based on gait analysis and sport science knowledge (Liu et al., 2021b).

In this paper, we introduce a method to further and more easily incorporate knowledge from other fields into Human Activity Recognition by enabling non-machine learning experts to develop a recognition system. We propose a setup where researchers can define high-level properties of activities with their respective possible values, such as Backwards/Neutral/Forwards on which the activity classification is performed. These features are initially extracted with out-of-the-box classifiers and can later be optimized or transformed into feature functions by machine learning experts. One requirement of said properties is that each activity differs from any other activity in at least one of them, making final classification straightforward, possible errors attributable to the erroneous feature, and the extracted feature space highly interpretable.

2 FEATURE SPACE

The proposed method transforms activities into a high-level feature space with discrete, categorical features, where each combination of features is unique across activities. Thereby breaking the activity classification problem into several sub-problems and a combination task. Instead of classifying a sequence as "Walk" directly, it is recognized as a forward movement (without Left/Right or Up/Down components) at a low speed and includes a cyclic knee movement and, therefore, is classified as "Walk".

The classification problem to choose between, say 22 different activities, is reduced to multiple classification problems with only a handful of discrete feature values and a final classification in this new feature space. Note that these sub-problems have more data available for each target class but are not necessarily more straightforward to solve as they include vastly different activity signals. Take "Stand" and "Falling Forward", while intuitively being very different, both start in a standing position, have little to no muscle activity, and have no movement to the left or right, and, therefore, share surprisingly many features and would be assigned the same target in several of them.

Setting these ideas into practice and evaluating them requires specific activities and data. In this case, two datasets, namely, CSL-SHARE (Liu et al., 2021a) (called CSL19 in the earlier pieces of literature) and UniMiB SHAR (Micucci et al., 2017). Both datasets mainly contain Activities of Daily Living (ADL) like "walk," "go upstairs," and "sit." In addition, the CSL-SHARE dataset includes several sport-related activities, such as "shuffle-Left/Right" and "V-cut", while the UniMiB SHAR dataset collects eight types of falls. The CSL-SHARE dataset applies four types of wearable sensors (two triaxial accelerometers, two gyroscopes, four EMG sensors, and one electrogoniometer) integrated into a knee bandage, while the UniMiB SHAR uses an accelerometer in an Android smartphone to sense the signal. One particularly challenging aspect of the UniMiB SHAR data is that the participants wore the smartphone in different orientations, making it hard to distinguish directions and other movements.

The initial high-level features are chosen based on (Liu et al., 2021b) and are as follows: *Class*, *Left/Right*, *Up/Down*, and *Back/Front*. *Class* describes the base activity, e.g., "walk", "fall", and "lounge". The other three are based on the Six-Direction Nomenclature (Liu et al., 2021b) and describe entire body translational movement. In contrast to Motion Units, the assignment here is not on sub-sequences but the whole activity level, thinking of these properties more as phonetic features than as phonemes as done with Motion Units. Furthermore, in this paper the "anchored" aspect is replaced by adding a "neutral" option to each of the three directional axes, e.g., Left/Neutral/Right. This neutral option is vital, as all features need to be extractable on every activity for a well-defined feature space.

However, with these four features, the feature combinations are not unique between activities and, therefore, the feature space is extended. On the CSL-SHARE dataset, three features are added: starting *Foot*, muscle *Force*, and *Knee* angle. The *Foot* is required to distinguish between "V-Cut left with the left foot first" and "V-Cut left with the right foot

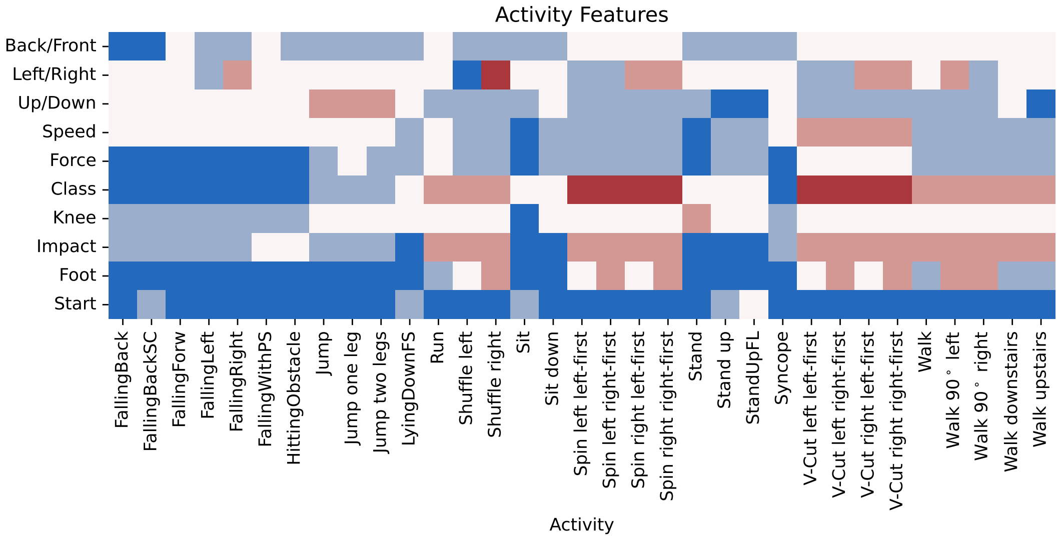


Figure 1: Assigned features per activity. Rows describe features with their respective categorical values per activity. Activities with the same value in a row use the same color. For example, the "FallingBack" and "FallingForward" activities have the same *Start* position marked in dark blue, but differ in the *Back/Front* feature (dark blue and white). No further meaning or order in color is intended, and the color value is solely maintained inside each row, e.g., dark blue does not refer to the same value in the *Start* and *Foot* row.

first", while "Jump with one leg" requires more muscle force to lift the body's weight compared to "Jump with two legs". Lastly, "Sit" and "Stand" can be distinguished when comparing knee angles. On the UniMiB SHAR side, a *Speed* feature to distinguish between "Running" and "Walking" is added. Additionally, a *Start* position and *Impact* feature are added to distinguish between "Stand up from sitting" and "Stand up from laying" as well as "Hitting an obstacle" and "Syncope".

The activities are grouped on association and activity description of the data publishers in each of the high-level features. Figure 1 displays the assignments for both datasets. The full assignment is required only in the combination experiment (Section 6). The following extraction and classification experiments only use the features developed for each dataset. Making the feature space more concise and resulting in simpler models. For the CSL-SHARE *Class*, *Left/Right*, *Up/Down*, *Force*, *Foot*, and *Knee* are used. For the UniMiB SHAR Activity, *Back/Front*, *Left/Right*, *Up/Down*, *Start* position, *Speed*, and *Impact* are used.

The current choice of features can be improved and should be developed for a broader range of activities. As Table 1 indicates, the 6DN features help distinguish activities across both datasets. The *Speed*, and *Force* features also help distinguish several activities. However, the *Impact*, *Foot*, and *Start* feature are developed mainly for specific datasets and do not provide much information on activities of the other dataset. Developing the features for a larger number of activities and datasets should prove helpful. Especially when paired with a procedure that minimizes the number of features while maximizing activity distances. Clustering comes to mind, either with machine learning techniques or by manually recursively splitting the activity pool. Furthermore, taking inspiration from other fields like the Bewegungslehre from sports science and education (Meinel and Schnabel, 1987) or the Labanotation from dance (Guest, 1977) could provide powerful features.

Note that the *Start* position feature differs from the other features because it requires knowledge of the starting time of an activity. When extracting the features, not into a vector but as a sequence, simply a position feature (values Lay/Sit/Stand) would provide the equivalent information by making *Start* a special case of the *Position* feature.

These choices already demonstrate the extensibility and adjustment possibilities for and to new activities and datasets. However, they also show the difficulty and importance of good feature choices. A poor choice of features might lead to activities not being discriminable in the feature space either due to the same feature values or poor feature extraction. Furthermore, a choice of features that each distinguishes only a few activities will not scale as each feature adds computational cost with little to no benefit in classification. The importance of feature choice also highlights the closeness between feature choice and activity definition. If no feature can distinguish two activities, are they indeed two different activities, and on which granularity should activities be distinguished? The proposed method makes modeling for human activity recognition more accessible for non-machine learning experts and easier to incorporate knowledge from other fields, emphasizing feature choice and design while abstracting the machine learning aspect.

3 FEATURE EXTRACTION

High-level feature extraction here describes the process of extracting attributes like Left/Neutral/Right from a given motion data sequence. The attributes are categorical and lend themselves to treating the extraction as a classification problem. In the following, three extraction methods are explored and evaluated based on Hidden Markov Models, Random Forests, and custom feature engineering. The main benefit of treating feature extraction as a classification problem is the easy extensibility. If a new feature is needed, only annotation of each activity is required to train an out-of-the-box classifier, like a Random Forest. However, Random Forests struggle with sequences of different lengths and, therefore, Hidden Markov Models are configured to work similarly to an out-of-the-box classifier.

The Random Forest is trained on the mean, sum, max, and min low-level features for each channel's first thousand samples. The low-level feature follows the notion that trends like *Left/Right* or *Up/Down* accumulate over time. The limitation to the first thousand samples results in the fact that for UniMiB SHAR, the whole sequence is used, while for CSL-SHARE, the first second of an activity is used (mean duration of most activities is around 1.7 seconds), which should be enough data to extract the chosen features and performed best in a cross-validation experiment.

The Hidden Markov Models can automatically learn to pay attention to essential segments of a sequence as well as channels by choice of topology and Gaussian Mixture training, which helps given the chosen features. For instance, not all parts of walking or falling might be important to determine if the feature value should be Left, Neutral, or Right. Additionally to the one-hot-encoded prediction of each feature, the HMM-based extraction also returns the HMMs confidence in the prediction as additional information to the combination task. Resulting in a feature space dimensionality of 39 (33 from feature values plus six confidences per feature) on the CSL-SHARE dataset and 37 on the UniMiB SHAR dataset (30 unique values plus seven confidences). The setup and hyperparameter choice follow the findings of (Hartmann et al., 2021), including mean-based rotation removal for the UniMiB dataset. Accordingly, a sliding rectangular window is used, the mean, RMS, slope, and max features calculated, and the whole sequence normalized afterward for both datasets. A grid search with a 10-fold person-independent evaluation was performed to determine the best window size, as the high-level features are less time-sensitive compared to activities, and some longer trends, like a left curve, should be more pronounced with longer windows. Accordingly, for the CSL-SHARE dataset, 100ms windows are used (exceptions are the *Left/Right*, *Up/Down*, and *Foot*, where 200ms windows are used) while for UniMiB SHAR, 200ms windows are used (exceptions are *Left/Right* and *Back/Front* with 400ms windows). Each activity is modeled with one of three predetermined topologies: general-purpose, one step, several steps. All three follow the same pattern where the first and last states are called "Random" and are shared across all activities. Thus, trimming the sequence but not influencing the classification. The general-purpose topology contains three consecutive states, while the step topology uses five, and several steps are modeled with three times five, i.e., three steps. The inner three to fifteen states then learn the main distinguishing aspects of the feature. Technically and for future experiments, the best matching topology for a sequence could be chosen automatically.

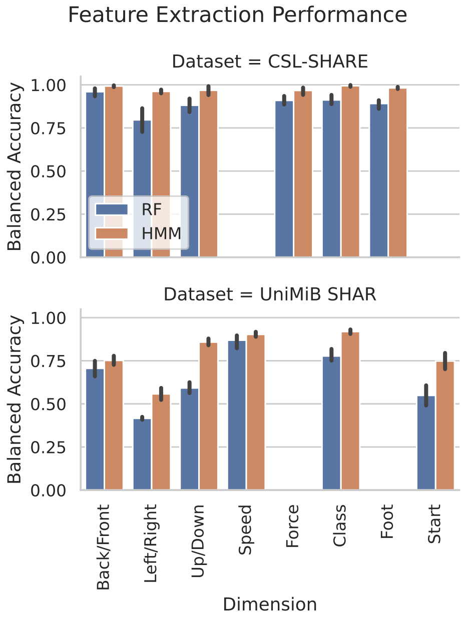
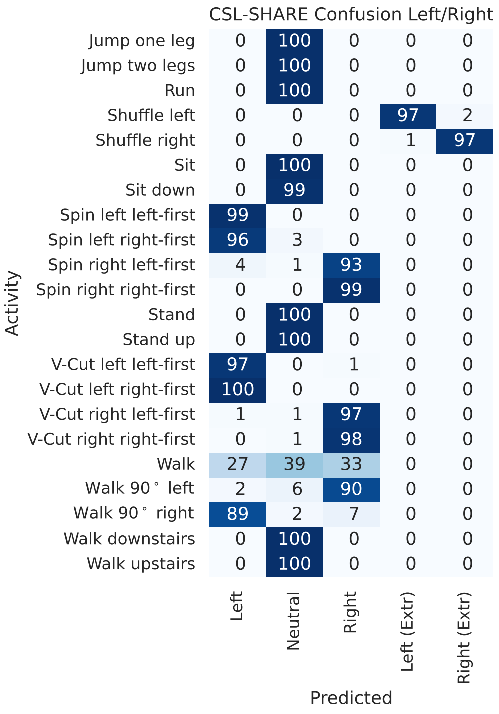


Figure 2: Balanced accuracies for feature extraction in a 10-fold person independent cross-validation.

The feature extraction with HMMs and Random Forests are then evaluated in a 10-fold person-independent cross-validation. The balanced accuracy is chosen, as the feature targets can be highly imbalanced (see Figure 1). The results are depicted in Figure 2. The HMM-based extraction outperforms the Random Forest one consistently. Accordingly, the following will focus on the HMM-based extraction. The features on the CSL-SHARE dataset are extracted with balanced accuracies in the very high ninety percent, with the lowest accuracy in the *Left/Right* feature at 96.1% balanced accuracy. This performance is encouraging, especially when compared to the activity classification accuracy of 93.7% in a leave-one-person-out cross-validation that is state of the art (Hartmann et al., 2021). On the UniMiB SHAR dataset, similar observations can be made. The balanced accuracies are around eighty percent, with *Left/Right* being an exception at 55.7% balanced accuracy. While these accuracies are promising for the overall classification, it is crucial to note that errors on the feature level propagate and add up in the final classification. Specifically, the extraction of the *Left/Right* feature requires further attention.

A closer look into the *Left/Right* evaluation in Figure 3 reveals that while the features are correctly extracted for almost all activities, the three activities “Walk”, “Walk 90° left”, and “Walk 90° right” are causing problems. Note that “Walk 90°” refers to walking a curve in three steps at which end a 90° turn is finished, rather than walking in a straight line. As *Left/Right* is the only feature distinguishing these activities in the high-level feature space, this error will propagate in the activity level classification (see Section 4).

The classifier-based feature extraction allows for effortless feature extension but adds a computational cost. For this reason, it is imperative to investigate if the features can also be extracted with hand-crafted functions. Furthermore, investigating what the classifiers base their decision on is of interest. As seen above, one of the main challenges and opportunities on the CSL-SHARE dataset is correctly identifying the direction in “Walk”, “Walk 90° left”, and “Walk 90° right”. Taking a closer look at the characteristics of these three activities shows that they have very distinct slopes in the readings of the lower gyroscope in the *Left/Right* direction, specifically during the swing phase, which ought to be further exploited. Therefore, a cross-channel feature was developed using the goniometer to determine the swing phase combined with extracting said slope from the lower gyroscope. The cross-channel feature, named *AvgSlopeSwing* in the following, is then classified by a random forest requiring only five estimators and the whole setup evaluated in a leave-one-person-out cross-validation scheme where it performed on par with the HMM-based extraction at 72.6% accuracy in the three-class problem. While the observed behavior extracted with this feature fits most participants, it mixes directions with the neutral case for others. Figure 4 illustrates this. The feature demonstrates that the computational cost of classifier-based feature extraction can be reduced with feature engineering and should be further pursued.


CSL-SHARE Confusion Left/Right

Activity	Left	Neutral	Right	Left (Extr)	Right (Extr)
Jump one leg	0	100	0	0	0
Jump two legs	0	100	0	0	0
Run	0	100	0	0	0
Shuffle left	0	0	0	97	2
Shuffle right	0	0	0	1	97
Sit	0	100	0	0	0
Sit down	0	99	0	0	0
Spin left left-first	99	0	0	0	0
Spin left right-first	96	3	0	0	0
Spin right left-first	4	1	93	0	0
Spin right right-first	0	0	99	0	0
Stand	0	100	0	0	0
Stand up	0	100	0	0	0
V-Cut left left-first	97	0	1	0	0
V-Cut left right-first	100	0	0	0	0
V-Cut right left-first	1	1	97	0	0
V-Cut right right-first	0	1	98	0	0
Walk	27	39	33	0	0
Walk 90 ° left	2	6	90	0	0
Walk 90 ° right	89	2	7	0	0
Walk downstairs	0	100	0	0	0
Walk upstairs	0	100	0	0	0

Figure 3: Confusion matrix for the *Left/Right* feature broken down into errors per activity.

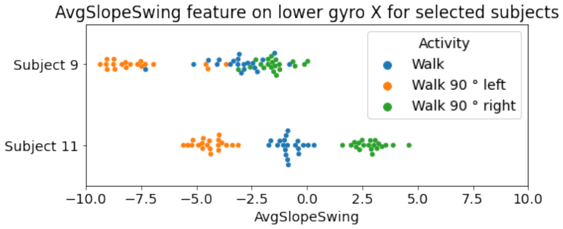
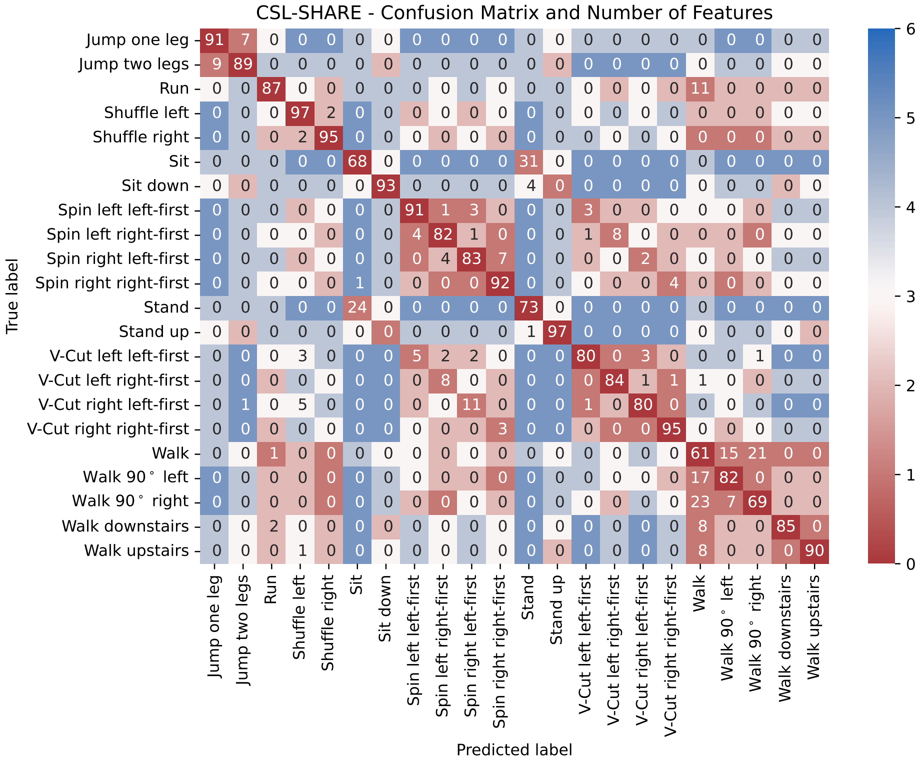


Figure 4: Average slope during swing phase feature illustrated for Left/Neutral/Right walking with two select subjects.



	Jump one leg	Jump two legs	Run	Shuffle left	Shuffle right	Sit	Sit down	Spin left left-first	Spin left right-first	Spin right left-first	Spin right right-first	Stand	Stand up	V-Cut left left-first	V-Cut left right-first	V-Cut right left-first	V-Cut right right-first	Walk	Walk 90 ° left	Walk 90 ° right	Walk downstairs	Walk upstairs
Jump one leg	91	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jump two legs	9	89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Run	0	0	87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0
Shuffle left	0	0	0	97	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Shuffle right	0	0	0	2	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sit	0	0	0	0	0	68	0	0	0	0	0	31	0	0	0	0	0	0	0	0	0	0
Sit down	0	0	0	0	0	0	93	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0
Spin left left-first	0	0	0	0	0	0	0	91	1	3	0	0	0	3	0	0	0	0	0	0	0	0
Spin left right-first	0	0	0	0	0	0	0	4	82	1	0	0	0	1	8	0	0	0	0	0	0	0
Spin right left-first	0	0	0	0	0	0	0	0	4	83	7	0	0	0	0	2	0	0	0	0	0	0
Spin right right-first	0	0	0	0	0	1	0	0	0	0	92	0	0	0	0	0	4	0	0	0	0	0
Stand	0	0	0	0	0	24	0	0	0	0	0	73	0	0	0	0	0	0	0	0	0	0
Stand up	0	0	0	0	0	0	0	0	0	0	0	1	97	0	0	0	0	0	0	0	0	0
V-Cut left left-first	0	0	0	3	0	0	0	5	2	2	0	0	0	80	0	3	0	0	0	1	0	0
V-Cut left right-first	0	0	0	0	0	0	0	0	8	0	0	0	0	0	84	1	1	1	0	0	0	0
V-Cut right left-first	0	1	0	5	0	0	0	0	0	11	0	0	0	1	0	80	0	0	0	0	0	0
V-Cut right right-first	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	95	0	0	0	0	0
Walk	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	61	15	21	0	0
Walk 90 ° left	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	82	0	0	0
Walk 90 ° right	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	7	69	0	0
Walk downstairs	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	85	0
Walk upstairs	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	90

Figure 5: CSL-SHARE confusion matrix and activity distance. Color indicates activity distance in high-level feature space by the number of different feature values. The number in each cell denotes confusion in percent.

4 CLASSIFICATION

The discriminatory power of the extracted high-level feature space combined with the Hidden Markov Models confidences is investigated using a 10-fold person-independent cross-validation. The feature extractors are trained on the same training data as the combining classifier to avoid overestimating the method's actual performance. A Random Forest is used for classification in the high-level feature space.

On the CSL-SHARE dataset, overall performance of 85% is achieved, compared to 93.6% reported previously (Hartmann et al., 2021). The main problems in classification occur where there are few distinguishing features or the distinguishing features' extraction accuracy is lacking. Figure 5 shows that most confusions occur, where the distance between activities is low. In that sense, the figure highlights areas of attention where high confusion is expected due to low distance in red and areas with high distance and low expected confusion in blue. Confusion in low-distance areas is further increased if the extraction of features struggles, which can be taken from Figure 2 and 1. For instance, the difference between the four "Spins" to their "V-Cut" counterparts is only in the *Force* feature (also the *Speed* feature, which is omitted here). Another example is the "Walk" and "Walk 90°" activities that differ in the *Left/Right* feature, which, as shown in Figure 3, has problems with determining the direction of these three activities.

A similar result is found in the UniMiB SHAR dataset, with a 64% accuracy compared to 77.0% reported previously (Hartmann et al., 2021). The confusion matrix reveals that the different falls are hard to discriminate due to only a few distinguishing features. "FallingLeft" and "FallingRight", for instance, are mainly differing in the *Left/Right* feature from each other as well as the other falls. However, as shown in Figure 2, the *Left/Right* feature is not extracted very well, possibly due to only one example activity for each feature value in the UniMiB SHAR dataset. Similarly, the *Start* feature has some problems, which manifests when comparing "Stand up from sitting" to "Stand up from laying" and again when comparing "FallingBackSC" and "FallingBack". A closer look into the projected high-level feature space shows that it errs 50% of the time in the *Left/Right* and 70% in the *Start* feature, leaving the final classifier little chance to correctly predict the activity.

These classification evaluations show that while

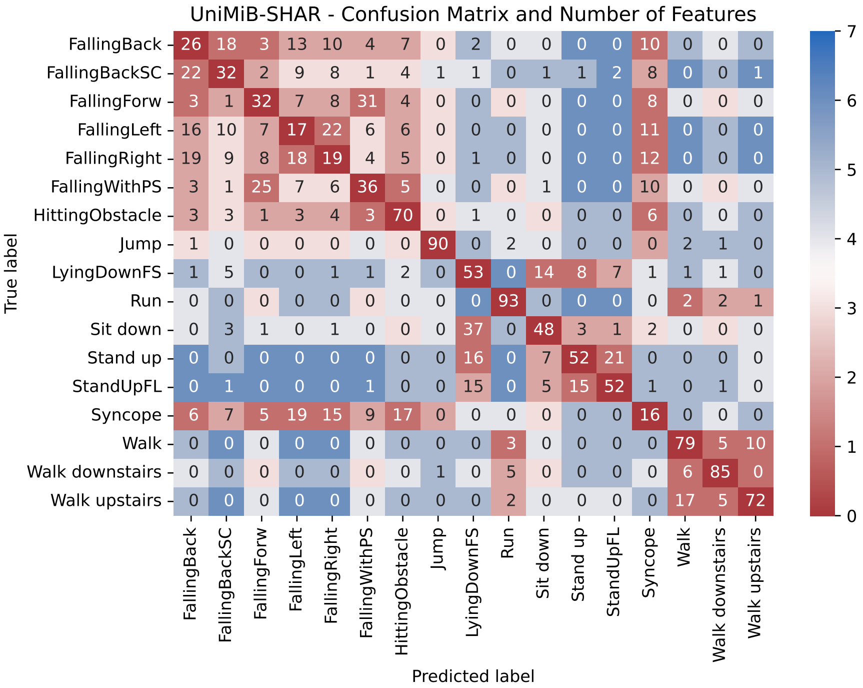


Figure 6: UniMiB SHAR confusion matrix and activity distance. Color indicates activity distance in high-level feature space by the number of different feature values. The number in each cell denotes confusion in percent.

this setup trails the reported state-of-the-art performances by ten points on both datasets, there is potential in the method. Specifically, due to its high interpretability, errors can be attributed and considered when further developing the features and feature extraction.

5 IMBALANCED DATA AND FEW-SHOT LEARNING

A significant advantage of the proposed method is that data from different activities is combined in the high-level feature extraction, resulting in a higher number of data points for each feature value. This data sharing allows features to be extracted from unseen activities and such with a low occurrence in the dataset while at the same time maintaining interpretability. Few-shot learning and imbalanced data are especially interesting in human activity recognition as there is a vast amount of different activities, which could be picked up quickly with a few-shot learning algorithm. Furthermore, some activities are hard to record and have little training data available but require accurate recognition. Falls are a common example, which in current datasets are typically simulated but are crucial to recognize accurately.

The following experiment evaluates the high-level classifier on the full data, except for one activity, from which only one sample is included in the training data in a 5-fold person-independent cross-validation. This procedure is repeated for each activity and with varying amounts of given samples. The activity occurrences in the CSL-SHARE dataset are mostly equally distributed at roughly 20 recordings per activity and participant. Therefore, four sample sizes are evaluated, namely: 1 sample (0.3% percent of otherwise available samples in the 5-fold scheme), 8 (2.5%), 16 (5%), and 32 (10%). In the UniMiB SHAR dataset the occurrences of each activity differ, ranging from 153 samples (“Standing up from sitting”) up to 1985 (“Running”), with the median at roughly 500 samples. Accordingly, the following numbers of samples were evaluated: 1 sample (0.25% mean), 10 (2.5% mean), 20 (5% mean), 40 (10% mean). As the training data is highly imbalanced and includes only a few training examples, the high-level classifier is switched from a Random Forest to a Balanced Random Forest utilizing undersampling (Chen et al., 2004). All other parameters remain the same, including the feature extraction. Four metrics are evaluated: the overall accuracy, the f1 score of the low sample activity, the percentage of samples with fully correct extracted features, and the amount of errors in

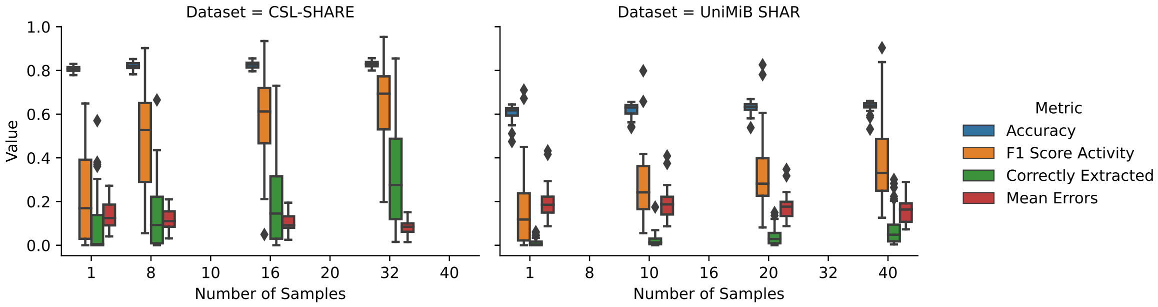


Figure 7: Performance on CSL-SHARE and UniMiB SHAR with four evaluated metrics. Averaged across 5-fold person-independent cross-validation and rotating low-resourced activity.

the extracted features (calculated as mean difference between the one-hot-encoded feature vectors without confidences). The overall accuracy can be compared with the classification results in Section 4 and indicate the overall performance implication if one activity is low-resourced. The f1 score indicates how well the low-resourced activity is recognized after training. The percentage of samples with correctly extracted features indicates how well the feature extraction works. This percentage might differ from the f1 score, as the activity-level classifier might compensate for wrongly extracted features but might also ignore correctly extracted features due to a bias learned in training. The error metric complements this by indicating how far off the feature extraction was from the correct feature values.

Figure 7 shows the results of this evaluation. The plot shows both the f1 score and correct extraction increasing and the mean extraction error decreasing with the number of shown samples, while the overall accuracy remains stable and is close to the scores found in the classification experiments (see Section 4). The latter, combined with the low f1 score, indicates that all other activities perform well independently of the low-resourced activity. The low f1 score shows that this setup struggles with one-shot learning, but can perform well with low-resourced activities beyond one-shot learning and can come close to full data with only 10% of the original data. Note that the mean error of the extracted feature is at less than 20% for both datasets (CSL-SHARE one-shot is at 14.2% and UniMiB SHAR at 19.5%). Recall that the one-hot encoded feature space has 33 (CSL-SHARE) and 30 (UniMiB SHAR) dimensions, meaning that on average less than three features are extracted incorrectly (dimensionality times the error divided by two, as an error in a feature results in two errors in the one-hot encoding) in both datasets in the one-shot experiment. Concluding, that while the classification of low-resourced activities should be improved, the feature extraction errs in few, but activity distinguishing features and improving these would improve few-shot learning.

6 DATASET COMBINATION

Another major advantage of the proposed method is the universality and independence of the extracted features from the sensor setup. Opening up the opportunity to combine and compare datasets in this high-level feature space. Of course, there are limitations, and not every feature makes sense for each dataset. However, this can be accounted for when choosing the features. The following experiment is a proof of concept and combines the aforementioned datasets CSL-SHARE and UniMiB SHAR in a stratified person-dependent evaluation. The scheme was chosen to ensure each activity occurring in the training and test set, which is challenging in a person-independent evaluation across datasets with different participants. The features listed in Section 2 for each of the datasets are combined as displayed in Figure 1. Therefore, more but not necessarily information-increasing features are extracted compared to the classification and few-shot learning experiments. The feature extraction is trained individually for the different datasets but similar to the previous experiments within the evaluation scheme. The activity-level classification is then done in the resulting feature space. Not unlike an interpretable equivalent to swapping the first layers of a Neural Network depending on input while keeping the last layers fixed. The activity-level classification is, therefore, independent from the dataset, which means, that now “V-Cuts” (CSL-SHARE) need to be distinguished from “Falls” (UniMiB SHAR). At the same time, both datasets have shared activities like “Walk” resulting in more samples compared to each individual dataset. The resulting 74% accuracy is in the middle between the 85% on the CSL-SHARE and the 64% on the UniMiB SHAR dataset (see Section 4). The confusion matrix is omitted here for brevity. The errors made during classification are the same as in the previous individual dataset experiments, except the "Walk" activity. "Walk" is now less confused with the "Walk 90°" activities and seems to have benefited from the additional samples or the additional features not present in previous experiments. The effects and possible benefits of the data combination will be investigated in future work. The dataset combination demonstrates that combining datasets with this method is possible, which creates a foundation to find commonalities and suited features for activities across different settings.

7 CONCLUSIONS

This paper introduces a novel way of classifying human activities based on unique combinations of interpretable high-level features for each activity. The extracted features are interpretable and easily extensible, and allow comparison and combination of different datasets in the high-level feature space. The three approaches using HMMs, Random Forests, and custom high-level feature functions for feature extraction were proposed and evaluated, of which the Hidden Markov Model performed best across the two datasets CSL-SHARE and UniMiB SHAR. Classification experiments investigated how well the extraction combined with a Random Forest for final classification could perform. In a person-independent 10-fold cross-validation, they performed well, at 85% on the CSL-SHARE compared to 93.6% state-of-the-art and 64% compared to 77.0% state-of-the-art on the UniMiB SHAR dataset (Hartmann et al., 2021). Furthermore, few-shot learning experiments were conducted, where one-shot learning did not succeed, but low extraction error rates and increasing f1 scores in few-shot learning are encouraging. Additionally, an experiment combining the two datasets showed the potential and promise of developing human activity recognition systems across data sources. Remarkable is that the errors in both classification experiments and few-shot learning experiments are attributable, and the next steps for increased performance are clear: deepened development and choice of features along with their extraction methods.

Furthermore, the next two major steps are clear: further investigate and develop high-level features and extract these as sequences rather than as vectors to enable online recognition. High-level features will be developed borrowing from previous HAR work, sports knowledge, and even utilizing findings and criteria from dance from decades of previous work. The main challenges for sequence extraction are creating the ground truth and addressing varying sampling rates across datasets. The ground truth creation likely requires manual data annotation and might only be possible for certain high-level features and if the dataset provides video examples. A sliding window approach could address the sequence problem when extracting features for a single dataset, but it does not scale easily across datasets due to different sampling rates. The slow nature of high-level features might enable re-sampling and should be investigated in future work. These and other topics, including estimating a performance ceiling with Neural Networks and extending to further datasets and modalities, are future work for these high-level features.

REFERENCES

Amma, C., Gehrig, D., and Schultz, T. (2010). Airwriting recognition using wearable motion sensors. In *First Augmented Human International Conference*, page 10. ACM.

Arifoglu, D. and Bouchachia, A. (2017). Activity recognition and abnormal behaviour detection with recurrent neural networks. *Procedia Computer Science*, 110:86–93.

Bakis, R. (1976). Continuous speech recognition via centisecond acoustic states. *The Journal of the Acoustical Society of America*, 59(S1):S97–S97.

Chen, C., Liaw, A., and Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. Technical report.

Dickinson, S. J., Leonardis, A., Schiele, B., and Tarr, M. J. (2009). *Object categorization: computer and human vision perspectives*. Cambridge University Press.

Guest, A. H. (1977). *Labanotation: Or, Kinetography Laban : the System of Analyzing and Recording Movement*. Number 27. Taylor & Francis.

Hartmann, Y., Liu, H., and Schultz, T. (2021). Feature space reduction for human activity recognition based on multi-channel biosignals. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 215–222. INSTICC, SciTePress.

Liu, H., Hartmann, Y., and Schultz, T. (2021a). CSL-SHARE: A multimodal wearable sensor-based human activity dataset. *Frontiers in Computer Science*.

Liu, H., Hartmann, Y., and Schultz, T. (2021b). Motion Units: Generalized sequence modeling of human activities for sensor-based activity recognition. In *EUSIPCO 2021 - 29th European Signal Processing Conference*. IEEE.

Meinel, K. and Schnabel, G. (1987). *Bewegungslehre - Sportmotorik: Abriß einer Theorie der sportlichen Motorik unter pädagogischem Aspekt.* Meyer Meyer Verlag, Aachen, 12., ergänzte auflage edition.

Micucci, D., Mobilio, M., and Napoletano, P. (2017). UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences*, 7(10):1101.

Oniga, S. and Sütő, J. (2014). Human activity recognition using neural networks. In *Proceedings of the 15th International Carpathian Control Conference*, pages 403–406. IEEE.

Ordóñez, F. J. and Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115.

Rebelo, D., Amma, C., Gamboa, H., and Schultz, T. (2013). Human activity recognition for an intelligent knee orthosis. In *BIOSIGNALS 2013 - 6th International Conference on Bio-inspired Systems and Signal Processing*, pages 368–371.

Ronao, C. A. and Cho, S.-B. (2014). Human activity recognition using smartphone sensors with two-stage continuous hidden markov models. In *ICNC 2014 - 10th International Conference on Natural Computation*, pages 681–686. IEEE.

Ronao, C. A. and Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244.

Scheirer, W. J., Anthony, S. E., Nakayama, K., and Cox, D. D. (2014). Perceptual annotation: Measuring human vision to improve computer vision. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1679–1686.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In *ICSLP 2002 - 7th International Conference on Spoken Language Processing*.

Singh, D., Merdivan, E., Psychoula, I., Kropf, J., Hanke, S., Geist, M., and Holzinger, A. (2017). Human activity recognition using recurrent neural networks. In *CD-MAKE 2017 - International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 267–274. Springer.

Tuncer, T., Ertam, F., Dogan, S., Aydemir, E., and Pławiak, P. (2020). Ensemble residual network-based gender and activity recognition method with signals. *The Journal of Supercomputing*, 76(3):2119–2138.

Uddin, M. Z., Thang, N. D., Kim, J. T., and Kim, T.-S. (2011). Human activity recognition using body joint-angle features and hidden markov model. *Etri Journal*, 33(4):569–579.

Yang, J., Nguyen, M. N., San, P. P., Li, X., and Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI*, volume 15, pages 3995–4001. Buenos Aires, Argentina.

Youngblood, G. M. and Cook, D. J. (2007). Data mining for hierarchical model creation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(4):561–572.