Abstract—Recognizing an activity with a single reference sample using metric learning approaches is a promising research field. The majority of few-shot methods focus on object recognition or face-identification. We follow a metric learning approach to reduce the action recognition problem to a nearest neighbor search in embedding space. We encode signals on a signal level into images and then extract features using a deep residual CNN. Using triplet loss, we learn a feature embedding. The resulting encoder transforms features into an embedding space in which closer distances encode similar actions while higher distances encode different actions. Our approach based on a signal-level formulation remains flexible across a variety of modalities while outperforming the baseline on the large scale NTU RGB+D 120 dataset for the One-Shot action recognition protocol by 4.2%. Further, we show generalization on experiments using the UTD-MHAD dataset for inertial data and the Simitate dataset for motion capturing data. Furthermore, our inter-joint and inter-sensor experiments suggest good capabilities on previously unseen joint and sensor setups.

I. INTRODUCTION

Learning to identify unseen classes from a few samples is an active research topic. Common methods are metric learning [1], [2] and meta-learning [3]. The majority of effort focuses on one-shot object recognition [4], person re-identification or face identification. Only recently few-shot methods for action recognition [5], [6], [7] have gained popularity.

Metric learning has already been used in a variety of computer vision tasks like face recognition [8], person re-identification [9], [11] and image retrieval [10]. One-shot action recognition research focuses on recognition from image sequences [2], [5], [11].

We tackle the problem of learning an action with a single reference demonstration. By considering a representation on a signal level, our approach serves as a general framework for multi-modal and even inter-modal one-shot action recognition. We define inter-modal one-shot recognition by an embedding that is trained on one modality and used for recognition on a new modality with a single reference demonstration per action class.

While classification methods are limited to predicting action labels, metric learning approaches learn an embedding by a similarity function. One-shot action recognition then becomes a nearest neighbor search in embedding space. Fig 1 gives an application example for one-shot action recognition on skeleton sequences using our approach.

In the case of skeleton sequences, we consider each joint axis as a signal sequence. In the case of motion capturing data, we consider each marker position axis from the demonstrator’s hand and the interacting objects as a signal. An image encodes signals discriminatively. Using an image representation yields the benefit of already established network architectures [12] for image classification tasks. The classifier and embedding encoder are jointly trained using triplet loss in conjunction...
with a Multi-Similarity Miner. The nearest neighbor in embedding space defines the most similar action. Fig. [1] gives an illustrative example for skeleton sequences.

Our proposed approach lifts state of the art in One-Shot action recognition on skeleton sequences on the NTU RGB+D 120 dataset for the one-shot evaluation protocol by 4.2% and also shows good performance on auxiliary set reductions. We further show that our approach is useful for one-shot action recognition on motion capturing action sequences on the Simitate dataset.

The main contributions of this paper are as follows:

- We propose a One-Shot action recognition approach based on learning action embeddings.
- We evaluate our proposed approach on the NTU RGB+D 120 dataset.
- We show that the approach generalizes from skeleton sequences to motion capturing sequences on the Simitate dataset and to inertial data on the UTD-MHAD dataset.
- We show, given a similar representation method, our approach is suitable for inter-modal applications.

We claim that our approach based on triplet loss and a common signal-level representation yields high flexibility for applications in one-shot action recognition. We demonstrate good results on one-shot action recognition for conventional sensor modalities (skeleton sequences, inertial measurements, motion capturing). Further, our approach shows functional capabilities when being trained on one modality (e.g., skeleton sequences) and applied on a different modality (e.g., inertial measurements) by a single reference sample per action class. Also, different setups, e.g., training on inertial sensors placed on the wrist and application with a placement on the leg, are possible.

II. RELATED WORK

We give a brief overview of methods related to metric learning and few-shot learning approaches in general. We focus on methods for action embeddings and few-shot action recognition.

Wang et al. [13] encode joint trajectory maps into images based on three spatial perspectives. Caetano et al. [14], [15] represent a combination of reference joints and a tree-structured skeleton as images. Their approach preserves spatio-temporal relations and joint relevance. Liu et al. presented a combination of skeleton visualization methods and jointly trained them on multiple streams. In contrast to our approach, their underlying representation enforces custom network architectures and is constrained to skeleton sequences, whereas our approach adds flexibility to other sensor modalities. Kim et al. [16] presented an interpretable visual method for action recognition using temporal convolutional networks. Their approach uses a spatio-temporal representation, which allows visual analysis to understand why a model predicted an action. Especially joint contributions are visually interpretable.

Schroff et al. [8] presented a joint face recognition and clustering approach. They trained a network such that the squared L2 distances in the embedding space directly correspond to face similarity [8]. Triplet loss [17] is used for training the embeddings. The embedding minimizes distances between anchor images and positive images (i.e., same person, different viewpoint) and maximizes distances to negative samples (different person). Yi et al. [9] presented a deep metric learning approach based on a siamese deep neural network for person re-identification. The two sub-nets are combined using a cosine layer. Wojke et al. [1] propose a deep cosine metric learning approach for the person re-identification task. The Cosine Softmax Classifier pushes class samples towards a defined class mean and therefore allows similarity estimation by a nearest neighbor search.

A recent action embedding approach by Hahn et al. [18] takes inspiration from the success of word embeddings in natural language processing. They combine linguistic cues from class labels with spatio-temporal features from sequences. A hierarchical recurrent neural network trains a feature extractor. A joint loss combines classification accuracy and similarity trains a function to encode the input into an embedding. Discriminative embeddings are important for few-shot learning approaches. Jasani et al. [6] proposed a similar approach for skeleton-based zero-shot action recognition. A text-based Spatio Temporal Graph Convolution Network (ST-GCN) [19] extracts features, which are encoded in semantic space by a continuous bag of words method.

Klipner-Gross et al. [2] proposed One-Shot-Similarity Metric Learning. A projection matrix that improves the One-Shot-Similarity relation between the example same and not-same training pairs represents a reduced feature space [2]. Fanello et al. [20] use Histogram of Flow and Global Histogram of Oriented Gradient descriptors with adaptive sparse coding and are classified using a linear SVM. Careaga et al. [5] propose a two-stream model for few-show action recognition on image sequences. They aggregate features from optical flow and the image sequences separately by a Long Short Term Memory (LSTM) and fuse them afterward for learning metrics. Rodriguez et al. [21] presented an one-shot approach based on Simplex Hidden Markov Models (SHMM). Improved dense trajectories are used as base features [22]. A maximum a posteriori (MAP) adoption and an optimized Expectation Maximisation reduce the feature space. A maximum likelihood classification, in combination with the SHMM, allows One-Shot classification. Roy et al. [23] propose a Siamese network approach for discriminating actions by a contrastive loss on a low dimensional representation gathered factory analysis. Mishra et al. [11] presented a generative framework for zero- and few-shot action recognition on image sequences. A probability distribution models classes of actions. The parameters are functions for semantic attribute vectors that represent the action classes.

Liu et al. [7] presented, along with the NTU RGB+D 120 dataset, an approach for one-shot action recognition. They propose an approach named Action-Part Semantic-Relevance aware (APSR). Features are generated by a ST-LSTM [24]. Motivated by word embedding methods, Liu et al. propose to
estimate semantic relevance on body parts and the actions. Similar semantic relevance for the body parts assigns new action instances.

The field of multi-modal few-shot action recognition is entirely unexplored. Al-Naser et al. [25] presented a zero-shot action recognition approach by combining gaze guided object recognition with gesture recognition arm-band. Actions are detected by fusing features of sub-networks per modality and integrating action definitions. Only three actions demonstrate the recognition results.

III. APPROACH

To cover the action recognition task across a variety of sensor modalities we consider the action recognition problem on a signal level. Signals are encoded in a discriminate image representation. An image-like representation allows direct adaption of already established image classification architectures for extracting features. On the extracted features we train a similarity function yielding an action embedding using triplet loss. The triplet loss minimizes embedding distances between similar action samples while maximizing distances between different actions. Finally to solve the one-shot problem, we apply a nearest neighbor search in the embedding space. An approach overview is given in Fig. 2.

A. Problem Formulation

We consider the One-Shot action recognition problem as a metric learning problem. First we encode action sequences on a signal level into an image representation. The input in our case is a signal matrix \( S \in \mathbb{R}^{N \times M} \) where each row vector represents a discrete 1-dimensional signal and each column vector represents a sample of all sensors at one specific time step. The matrix is transformed to an RGB image \( I \in \{0, \ldots, 255\}^{H \times W \times 3} \) by normalizing the signal length \( M \) to \( W \) and the range of the signals to \( H \). The identity of each signal is encoded in the color channel.

Resulting in a dataset \( D = \{(I_i, y_i)\}_{i=1}^N \) of \( N \) training images \( I_i \) with labels \( y_i \in \{1, \ldots, C\} \). Our goal is to train a feature embedding \( \vec{x} = f_\Theta(I) \) with parameters \( \Theta \) which projects input images \( I \in \{0, \ldots, 255\}^{H \times W \times 3} \) into a feature representation \( \vec{x} \in \mathbb{X}^d \). The feature representation reflects minimal distances for similar classes.

B. Representation

To allow a CNN based classifier to discriminate well between the action classes, we aim to find a discriminative representation in the first place. For encoding the signal identity we sample discriminative colors in the HSV color space depending on the number of signals. Temporal relations are represented by the position in the image. Signal changes are encoded spatially and joint relations are preserved. A limitation of this approach is that only lower dimensional signals can be encoded. Image sequences or their transformations like optical flow, motion history images have to many dimensions to encode on a signal level by using our representation. Extracted human pose estimates, hand- and/or object estimates
from image sequences are adequate signals for encoding in this representation. Example representations for different sensor modalities are given in Fig. 3.

C. Feature Extraction

Most action recognition approaches based on CNNs present custom architecture designs in their pipelines. A benefit is the direct control over the number of model parameters that can be specifically engineered for data representations or use cases. Recent advances in architecture design can not be transferred directly. Searching good hyper-parameters for training is then often an empirical study. Minor architecture changes can result in a completely different set of hyper-parameters. He et al. suggested the use of residual layers during training to tackle the vanishing gradient problem. We take advantage of the recent development in architecture design and use an already established architecture for image classification. We decided to use a Resnet18 architecture and take advantage of pre-trained weights. The feature extractor serves as input for learning a metric on the auxiliary set.

D. Metric Learning

Metric learning aims to learn an embedding space, where the embedding vectors of similar samples are encouraged to be closer, while dissimilar ones are pushed apart from each other. We use a triplet loss in combination with a Multi-Similarity-Miner for mining good triplet candidates during training.

While the triplet loss has been used in image ranking, face recognition, person re-identification, and for complex event detection, it has only rarely been used for inter- and cross-modal ranking to improve action recognition and for complex event detection. Given a triplet of an anchor image \( I_0 \), a positive data sample, representing the same action class image \( I_1 \), and a negative sample, representing a different action class \( I_4 \), the triplet loss can be formulated as:

\[
\mathcal{L}_t(I_0, I_1, I_4) = \max\left( \|f(I_0) - f(I_1)\|_2^2 - \|f(I_0) - f(I_4)\|_2^2 + \alpha, 0 \right)
\]

where \( \alpha \) describes an additional distance margin. Finding good candidate pairs is crucial. Therefore we use a Multi-Similarity Miner to mine positive and negative pairs that are assumed to be difficult to push apart in the embedding space. That means positive pairs are constructed by an anchor and positive image pair \( \{I_0, I_1\} \) and it’s embedding \( f(I_0) \), preferring pairs with a low distance in embedding space with the following condition:

\[
\|f(I_0) - f(I_1)\|_2 > \min_{y_k \neq y_i} \|f(I_0) - f(I_k)\|_2 - \epsilon,
\]

likewise, negative pairs \( \{I_0, I_4\} \) are mined by the highest distance in embedding space:

\[
\|f(I_0) - f(I_1)\|_2 < \max_{y_k \neq y_i} \|f(I_0) - f(I_k)\|_2 + \epsilon,
\]

where \( \epsilon \) is a given margin. Finally, we yield the total loss by:

\[
\mathcal{L} = \alpha \mathcal{L}_t + \beta \mathcal{L}_c,
\]

such that the influences of the loss can be weighted using the scalars \( \alpha \) for the triplet loss and \( \beta \) for the classifier loss. Finding an action class by a query and set of references is now reduced to a nearest-neighbor search in the embedding space.

IV. Experiments

To show the multi-modal One-Shot recognition performance we applied our methods to three datasets containing three different modalities. We used skeleton sequences from the NTU RGB+D 120 dataset for large scale One-Shot action recognition. With 100 auxiliary classes and 20 validation classes it is the largest dataset that we applied to our approach. To show the multi modal capabilities of our approach we also used the UTD-MHAD dataset (inertial and skeleton data) and the Simitate dataset (motion capturing data).

The datasets are split into an auxiliary set, representing action classes that are used for training, and a validation set. In our experiments the validation set does contain novel actions or actions from a novel sensor modality. One sample of each validation class serves as reference demonstration.
Fig. 4: Result graphs for the NTU RGB+D 120 dataset (a), the UTDMHAD dataset (b) and the Simitate dataset (c). Skl denotes skeleton data, IMU denotes inertial data, s. val denotes a static validation set.

| Approach                               | Accuracy |
|----------------------------------------|----------|
| Attention Network [35]                 | 41.0%    |
| Fully Connected [33]                   | 42.1%    |
| Average Pooling [36]                   | 42.9%    |
| APSR [7]                               | 45.3%    |
| Ours                                   | 49.5%    |

TABLE I: One-Shot action recognition results on the NTU RGB+D 120 dataset.

| #Train Classes | APSR [7] | Ours |
|----------------|----------|------|
| 20             | 29.1%    | 33.6%|
| 40             | 34.8%    | 36.9%|
| 60             | 39.2%    | 40.1%|
| 80             | 42.8%    | 44.1%|
| 100            | 45.3%    | 49.5%|

TABLE II: Results for different auxiliary training set sizes for One-Shot recognition on the NTU RGB+D 120 dataset.

This protocol is based on the one proposed by [7] for the NTU RGB+D 120 dataset. We conducted similar experiments with the remaining two data sets. In depth descriptions are given below in Section IV-A per dataset. Results are discussed after the dataset presentation in Section IV-A0c. First we trained a model on the auxiliary set. The resulting model estimates embeddings for the reference actions and then for the evaluation actions. We then calculate the nearest neighbour from the evaluation embeddings to the reference embeddings. This yields to which action from the reference set the current evaluation sample comes closest.

A. Datasets

a) NTU RGB+D 120: The NTU RGB+D 120 [7] dataset is a large scale action recognition dataset containing RGB+D image streams and skeleton estimates. The dataset consists of 114,480 sequences containing 120 action classes from 106 subjects in 155 different views. We follow the One-Shot protocol as described by the dataset authors. The dataset is split into two parts: an auxiliary set and an evaluation set. The action classes of the two parts are distinct. 100 classes are used for training, 20 classes are used for testing. We follow the unseen classes and reference samples as documented in the accompanied dataset repository [7] A1, A7, A13, A19, A25, A31, A37, A43, A49, A55, A61, A67, A73, A79, A85, A91, A97, A103, A109, A115 are previously unseen. As reference the demonstration for filenames starting with S001C003P008R001* are used for actions with IDs below 60 and S018C003P008R001* for actions with IDs above 60. One-Shot action recognition results are given in Table I. Like Liu et al. [7] we also experimented with the effect of the auxiliary set reduction. Results are given in Fig. [4] (a) and Table II.

b) UTDM-HAD: The UTDM-HAD [33] contains 27 actions of 8 individuals performing 4 repetitions each. RGB-D camera, skeleton estimates and inertial measurements are included. The RGB-D camera is placed frontal to the demonstrating person. The IMU is either attached at the wrist or the leg during the movements. No One-Shot protocol is defined therefore we defined custom splits. We started with 23 auxiliary classes and evaluated with reduced training sets. We then proceeded two-fold. First we evaluated with a static validation set of the four actions with the highest ID (24-27) and incrementally reduced the training set. Per experiment we removed the top four action classes until the validation classes exceeded the auxiliary classes. Second we proceeded similar for the auxiliary classes but moved the reduced training instances over into the validation set. By this we decreased the training set while increasing the validation set. Results for the two experiments executed on skeleton and inertial data are given in Table III. Fig. [4] (b) show visually the influence of the auxiliary set. In a third experiment we evaluated the inter-joint One-Shot learning abilities of our approach. For actions with ids up to 21 the inertial unit was placed on the subjects wrist and for the remaining ids from 22-27 the sensor was placed on the subjects leg. This allows us to inspect the One-Shot action recognition transfer to other sensor positions by learning on wrist sequences and recognize on leg sequences with one reference example. We always used the first trial of the first subject as reference sample and the remainder for testing. Results for the inter-joint experiment on inertial data are given in Table IV.

https://github.com/shahroudy/NTURGB-D
We further evaluate on the Simitate dataset. The Simitate benchmark focuses on robotic imitation learning tasks. Hand and object data are provided from a motion capturing system in 1932 sequences containing 26 classes of 4 different complexities. The individuals execute tasks of different kinds of activities from drawing motions with their hand over to object interactions and more complex activities like ironing. We consider one action class of each complexity level as unknown. Namely, *zickzack* from basic motions, *mix* from motions, *close* from complex and *bring* from sequential. Resulting in an auxiliary set of 22 classes and 4 evaluation classes. The corresponding first sequence by filename is used as reference sample. Results are given in Table VI.

| #Train Classes | #Val Classes | Skeleton | Inertial |
|----------------|--------------|----------|----------|
| 23             | 4            | 97.6%    | 81.3%    |
| 19             | 4            | 95.9%    | 78.0%    |
| 15             | 4            | 92.7%    | 83.7%    |
| 11             | 4            | 92.7%    | 87.0%    |
| 7              | 4            | 82.1%    | 84.6%    |
| 3              | 4            | 82.9%    | 91.1%    |
| 19             | 8            | 88.2%    | 74.0%    |
| 15             | 12           | 71.6%    | 66.8%    |
| 11             | 16           | 62.1%    | 55.0%    |
| 7              | 20           | 53.0%    | 45.2%    |
| 3              | 24           | 47.1%    | 42.3%    |

**TABLE III:** One-Shot action recognition results on the UTD-MHAD dataset.

| #Train Classes | #Val Classes | Skeleton | Inertial |
|----------------|--------------|----------|----------|
| 21             | 6            | Left wrist | 72.8% |
| 6              | 21           | Left leg  | 29.5% |
| 6              | 6            | Left wrist | 83.7% |
| 6              | 6            | Left leg  | 64.0% |

**TABLE IV:** Inter-joint One-Shot action recognition results on the UTD-MHAD dataset.

c) *Simitate:* We further evaluate on the Simitate dataset. The Simitate benchmark focuses on robotic imitation learning tasks. Hand and object data are provided from a motion capturing system in 1932 sequences containing 26 classes of 4 different complexities. The individuals execute tasks of different kinds of activities from drawing motions with their hand over to object interactions and more complex activities like ironing. We consider one action class of each complexity level as unknown. Namely, *zickzack* from basic motions, *mix* from motions, *close* from complex and *bring* from sequential. Resulting in an auxiliary set of 22 classes and 4 evaluation classes. The corresponding first sequence by filename is used as reference sample. Results are given in Table VI.

| Train Modality | Val. Modality | Accuracy |
|----------------|---------------|----------|
| Skeleton       | Inertial      | 41.13%   |
| Inertial       | Skeleton      | 43.53%   |

**TABLE V:** Inter-sensor One-Shot action recognition results on the UTD-MHAD dataset.

| #Train Classes | #Val Classes | Accuracy |
|----------------|--------------|----------|
| 22             | 4            | 93.2%    |
| 18             | 4            | 91.0%    |
| 14             | 4            | 93.2%    |
| 10             | 4            | 91.0%    |
| 18             | 8            | 79.2%    |
| 14             | 12           | 56.0%    |
| 10             | 16           | 36.2%    |

**TABLE VI:** One-Shot action recognition results on the Simitate dataset.

On the NTU RGB+D 120 dataset we compare against the proposed baseline APSR by Liu et al. [7]. Table I shows the results with a auxiliary set size of 100 action classes and a validation set size of previously unseen 20 action classes. Our proposed approach performs 4.2% better than the first follow up [7] and 6.6% better than the second follow up [8]. Fig. 4(a) and Table II shows results for an increasing amount of auxiliary classes (100 auxiliary classes and 20 validation classes are considered as the standard protocol). Overall our approach performs better on all conducted auxiliary set experiments as the baseline approach. Especially with a lower and higher amount of auxiliary classes our metric learning based approach performs better. In comparison to the baseline the accuracy of our approach increased more linearly by increasing the amount of auxiliary classes. In Fig. 5 we show a UMAP [37] visualizations that give an insight about the discriminate capabilities. Distances in embedding space capture the amount of identities well. This is the case for the three top clusters containing the actions (grab other person’s stuff, take a photo of other person and hugging other person). The two clusters at x-axis around 7.5 correspond to the actions arm circles and throw, suggesting that actions with clear high joint-relevance can also be clustered well. The most right cluster corresponds to the class falling and supports this hypothesis. In the top center we have a quite sparse cluster reflecting highly noisy skeleton sequences from multiple classes. Mainly sequences with multiple persons, especially with close activity like hugging, resulted in noisy data.

The UTD-MHAD dataset was used to show the generalization capabilities of the proposed approach across different modalities. By considering a signal level action representation we could compare skeleton and inertial results and also perform inter-joint and inter-sensor experiments. Fig. 4(b) shows the effect on the resulting One-Shot accuracy with increasing auxiliary sets. Interesting to note is that these experiment series we could observe that not necessarily a higher amount of classes used for training will lead to a higher accuracy. This was the case for our experiments on inertial data, where training on only three classes shows results in a more similar action embedding. This observation could not be transferred to the skeleton experiments on this dataset. The selection of auxiliary classes used for training should be well chosen. Adding more classes does not necessarily mean higher similarity in the embedding but can also add more confusion. Our inter-joint experiments yielded more transferable embeddings by training on data from the wrist and validating on the leg as shown in Table IV. This holds true for our conducted experiments, but we do want not exclude the possibility of finding a subdivision of the wrist auxiliary set that results in a higher transferable embedding. A key-insight among our experiments is that balanced classes for training and testing yielded mostly higher accuracy for lower dimensional modalities like IMU (see Table III) and motion capturing (see Table IV). This is especially visible in our inter-joint experiments as shown in Table IV. In comparison, our experiments applied to skeleton...
sequences benefited from more auxiliary classes (see Table III and Fig. 4 (a,c)). In our inter-sensor experiments we used all action from one modality as auxiliary set and evaluated the other modality with a single reference sample. Results for this experiment are given in Table V. The resulting One-Shot recognition for inertial to skeleton performs slightly (+2.4%) better than the reverse direction. In approximately 40% of the time an action trained on a different data modality on the UTD-MHAD dataset could be recognized with just one reference sample. This is an interesting observation and shows the flexibility of our proposed approach. We observed that the inertial measurements have a relation to the arm and hand movements of the skeleton, which explains the good transferability across the modalities.

Finally, we evaluated our approach on the Simitate dataset with motion capturing data. Results are given in Table VI and Fig. 4 (c). The amount of classes is comparable to the one from the UTD-MHAD dataset. The effects of the auxiliary set reduction are en par with the experiments conducted on the UTD-MHAD dataset. The results that the proposed approach transfers also good to motion capturing data. The class-distances from the motion capturing experiments are higher in embedding space then the ones gathered by the inertial experiments (see Fig. 4(b) & (c)).

B. Implementation

Our implementation is based on Pytorch [32], [38]. The representations were generated as described in [29]. A batch size of 32 was used on a single Nvidia GeForce RTX 2080 Ti with 11GB GDDR-6 memory. We trained for 50 epochs using a pre-trained Resnet18 [12]. Embeddings and models were saved for the epoch with the highest accuracy. The classification and metric loss were weighted by 0.5 each. For the multi similarity miner we used an epsilon of 0.05 while we used a margin of 0.2 for the triplet margin loss. A stochastic gradient decent optimizer with a learning rate of 0.01 was used in all optimizers. The embedding model outputs a 256 dimensional embedding and the classifier yielded a 200 dimensional feature vector.

V. Conclusion

We presented a One-Shot action recognition approach by employing a signal level representation in conjunction with metric learning using a triplet margin loss. By considering a representation on a signal level, our approach remains flexible across different sensor modalities like skeleton, inertial measurements and motion capturing data. Our approach allows one-shot recognition on all of the modalities we experimented with and further indicated to serve as a flexible framework for inter-joint and even inter-sensor experiments. We evaluate our approach on three different, publicly available, datasets. Most importantly, we showed an improvement of the current state of the art for one-shot action recognition on the large scale NTU RGB+D 120 dataset. To show the transfer capabilities, we also verified our results using the UTD-MHAD dataset for skeleton and inertial data and the Simitate dataset for motion capturing data. Inter-joint experiments show inertial sensor attached to the wrist and the leg from the UTD-MAHD dataset. On the UTD-MHAD dataset inter-sensor experiments between the skeleton and inertial data were executed. We found that more classes used during training for lower variate sensor data like IMUs and motion capturing systems do not necessarily improve the one-shot recognition accuracy in our experiments. A good selection of training classes and a balanced training and validation set improved results across all modalities.

REFERENCES

[1] N. Wojke and A. Bewley, “Deep cosine metric learning for person re-identification,” in 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018, pp. 748–756.
[2] O. Kliper-Gross, T. Hassner, and L. Wolf, “One shot similarity metric learning for action recognition,” in International Workshop on Similarity-Based Pattern Recognition. Springer, 2011, pp. 31–45.
[3] L.-Y. Gui, Y.-X. Wang, D. Ramanan, and J. M. Moura, “Few-shot human motion prediction via meta-learning,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 432–450.
[4] Z. Fu, T. Xiang, E. Kodirov, and S. Gong, “Zero-shot object recognition by semantic manifold distance,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2635–2644.
[5] C. Careaga, B. Hutchinson, N. Hodas, and L. Phillips, “Metric-based few-shot learning for video action recognition,” arXiv preprint arXiv:1909.09602, 2019.
[6] B. Jasani and A. Mazagonwalla, “Skeleton based zero shot action recognition in joint pose-language semantic space,” arXiv preprint arXiv:1911.11346, 2019.
D. Roy, C. K. Mohan, and K. S. R. Murty, “Action recognition based on
S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional
networks for skeleton-based action recognition,” in Thirty-second AAAI conference on artificial intelligence, 2018.
S. R. Fanello, I. Gori, G. Metta, and F. Odone, “One-shot learning for real-time action recognition,” in Iberian Conference on Pattern Recognition and Image Analysis. Springer, 2013, pp. 31–40.
M. Rodriguez, C. Orrite, C. Medrano, and D. Makris, “Fast simplex-lmm for one-shot learning activity recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 41–48.
H. Wang and C. Schmid, “Action recognition with improved trajectories,” in Proceedings of the IEEE Conference on computer vision, 2013, pp. 3551–3558.
D. Roy, C. K. Mohan, and K. S. R. Murty, “Action recognition based on discriminative embedding of actions using siamese networks,” in 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018, pp. 3473–3477.
J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, “Skeleton-based action recognition using spatio-temporal lstm network with trust gates,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 12, pp. 3007–3021, 2017.
M. Al-Naser, H. Ohashi, S. Ahmed, K. Nakamura, T. Akiyama, T. Sato, P. X. Nguyen, and A. Dengel, “Hierarchical model for zero-shot activity recognition using wearable sensors,” in ICAART (2), 2018, pp. 478–485.
M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” Pattern Recognition, vol. 68, pp. 346–362, 2017.
X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, “Multi-similarity loss with general pair weighting for deep metric learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5022–5030.