2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Gimme Signals: Discriminative signal encoding for multimodal activity recognition

Raphael Memmesheimer Nick Theisen Dietrich Paulus

***Abstract*— We present a simple, yet effective and flexible method for action recognition supporting multiple sensor modalities. Multivariate signal sequences are encoded in an image and are then classified using a recently proposed EfficientNet CNN architecture. Our focus was to find an approach that generalizes well across different sensor modalities without specific adaptions while still achieving good results. We apply our method to 4 action recognition datasets containing skeleton sequences, inertial and motion capturing measurements as well as Wi-Fi fingerprints that range up to 120 action classes. Our method defines the current best CNN-based approach on the NTU RGB+D 120 dataset, lifts the state of the art on the ARIL Wi-Fi dataset by +6.8%, improves the UTD-MHAD inertial baseline by +14.4%, the UTD-MHAD skeleton baseline by +0.5% and achieves 96.1% on the Simitate motion capturing data (80/20 split). We further demonstrate experiments on both, modality fusion on a signal level and signal reduction to prevent the representation from overloading.**

I. Introduction

Action (also referred to as activity or behaviour) recognition is a well studied field and enables application in many different areas like elderly care [5], [6], [7], [8], smart homes [7], [8], surveillance [9], [10] robotics [11], [12] and driver behaviour analysis [13], [14], [15].

Action recognition can be defined as finding a mapping that assigns a class label to a sequence of signals. The input data can, for instance, be measurements from Inertial Measurement Units (IMU), skeleton sequences, motion capturing sequences or image streams. We tackle the action recognition problem on a signal level as this is a common basis for a variety of input modalities or features that can be transformed into multivariate signal sequences. A common basis is important for the generalization across different modalities.

Some sensors like IMUs, Wi-Fi receivers yield multivariate signals directly, other sensors like RGB-D cameras provide skeleton estimates indirectly. Skeleton estimates can be transformed easily into multivariate signals by considering their joint axes. This also holds for human poses that can be estimated on camera streams using recent methods [16]. Predicting the action class from multivariate signal sequences can then be seen as finding discriminative representations for signals.

Convolutional neural networks have shown great performance in classification tasks. We, therefore, propose a representation that transforms multivariate signal sequences into images. Recent proposed Convolutional Neural Network (CNN) architectures use architecture search conditioned on maximizing the accuracy while minimizing the floating-point operations [17], [18]. Therefore they are good candidates for use in robotic systems. Figure 1 gives an exemplary overview of the variety of modalities that our proposed representation can be used for. We evaluated the approach on 5 datasets containing different modalities. Many proposed fusion approaches rely on custom-engineered sub-models per sensor modality which are usually combined in multi-stream architectures. In contrast, we fuse the modalities on a representation level. This has the huge benefit of having a

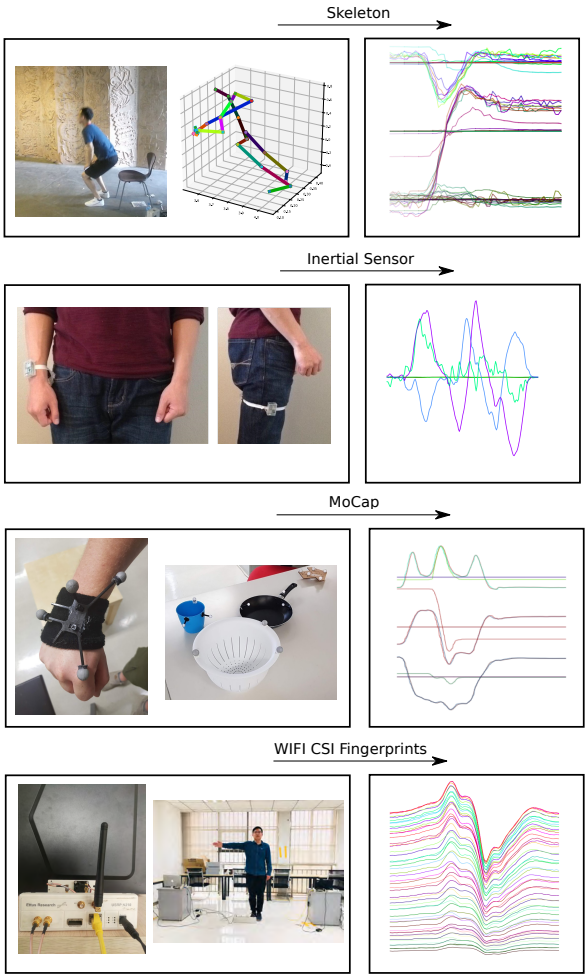

Fig. 1: We propose a representation that is suitable for multimodal action recognition. The Figure shows representations for skeletal data from the NTU [1], [2] dataset, Inertial data from the UTD-MHAD [3] dataset and WiFi Fingerprints from the ARIL [4] dataset.

All authors are with the Active Vision Group, Institute for Computational Visualistics, University of Koblenz-Landau, Germany
Corresponding email: raphael@uni-koblenz.de

constant computing complexity independent of the number of modalities used whereas multi-stream architectures raise in complexity with every modality added.

Our approach lifts the state of the art action recognition accuracy on the ARIL Wi-Fi dataset by +6.8% and the UTD-MHAD [3] (IMU +14.4) (Skeleton +0.5%). Our approach defines the current best 2D-CNN based approach on the NTU RGB+D 120 dataset (+2.9% (cross-subject), +4.6% (cross-view)%) while being outperformed by a recently proposed graph convolution approach [19] achieving remarkable results. On the Simitate dataset we achieve 96.1% accuracy on motion capturing data. In total we evaluated our approach on 4 different modalities. To the best of our knowledge, there is no approach showing a comparable high flexibility in supported sensor modalities.

The main contributions of this paper are as follows:

- We propose an action recognition approach based on the encoding of signals as images for classification with an efficient 2D-CNN.
- We propose filter methods on a signal level to remove signals with only a minor contribution to the action.
- We present an approach for information fusion on a signal level.

By considering the action recognition problem on a signal level, our approach generalizes well across different sensor modalities. The signal reduction prevents the image representation from overloading and allows flexible addition of signal streams. By fusion on a signal level, we create a flexible framework for adding additional information for instance object estimates or the fusion of different sensor modalities. The source code for the presented method is available on github[1].

II. Related Work

In this section, we present action recognition methods based on traditional feature extractors and recent advances in machine learning. Existing survey papers [20], [21], [22], [23], [24] do not include most recent publications as the action recognition field is a highly active field of research. Therefore, most recent approaches from other working groups are presented here. We put a focus on methods using skeleton sequences as input because these can be acquired on robotic systems directly from RGB-D frames or by extracting human pose features [16] from video sequences. Further, large scale benchmarks [1] are available for action recognition on skeleton sequences, thus a fair comparison of different approaches can be achieved.

An interesting analysis from a human visual perception point of view has been presented by Johansson [25] in 1973. He found that humans are using 10-12 elements in proximal stimulus to distinguish between human motion patterns [25]. This supports the use of skeletons or pose estimation maps as underlying representations for activity recognition approaches from a visual perception perspective [26]. Recent advances in action recognition developed from hand crafted feature extractors to deep learning approaches like 2D- and 3D-CNNs, while in parallel LSTM based methods also improved results on large scale datasets. More recently, graph convolution approaches showed promising results.

Rahmani et al. [27] presented viewpoint invariant histograms of gradient descriptors for action recognition. Vemulapalli [28] represented skeleton joints as points in a Lie-group. The classification is then done by a combination of dynamic time-warping [29], Fourier temporal pyramid representation and linear SVM [28]. More recent approaches suggest representing skeleton sequences as images and 2D-CNNs for recognition. Wang et al. [30] encode joint trajectory maps into images based on three spatial perspectives. Caetano et al. [31], [32] represent a combination of reference joints and a tree-structured skeleton in images. Their approach preserves spatio-temporal relations and joint relevance. Liu et al. [33] study a pose map representation. The approach that comes closest to our approach is by Liu et al. [34]. 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. [35] presented a visual interpretable 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.

3D convolutions for video action recognition was popularized by Tran et al. [36]. They have shown good performance on direct video action classification A three-stream network has then been proposed to integrate multiple cues sequentially via a Markov chain model [37]. By the integration of additional cues from e.g. pose information, optical flow and RGB images using a Markov chain they could increase the recognition accuracy incremental with each additional cue.

CNN architectures for signal classification have also been studied previously in audio processing [38]. ResNet 1D-CNN architectures have been used for joint classification and localization of activities in Wi-Fi signals [4]. For activity classification on a set of inertial sensors Yang et al. [39] acquire time-series signals and classify the activities using a multi-layer CNN.

Liu et al. [40] presented a spatio-temporal LSTM inspired by graph-based representation of the human skeleton. They further introduced a novel trust-gating mechanism to overcome noise and occlusion. Si et al. [26] presented an Attention Enhanced Graph Convolutional LSTM Network (AGC-LSTM). They use feature augmentation and a three-layer AGC-LSTM to model discriminative spatial-temporal features and yield very good results on cross-view and cross-subject experiments on skeleton sequences. Very recently Papadopoulos et al. [19] proposed two novel modules to improve action recognition based on Spatial Graph Convo-

[1] http://github.com/airglow/gimme_signals_action_recognition

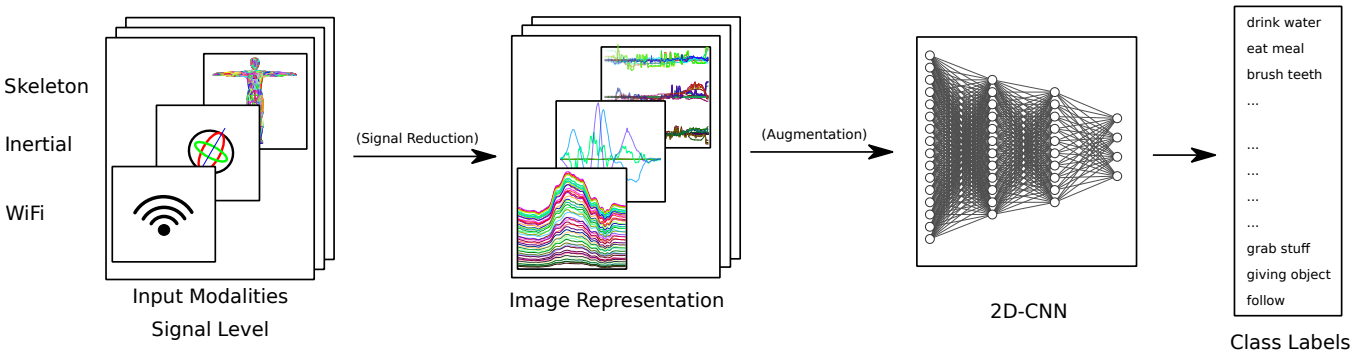

Fig. 2: Approach overview. We propose to transform individual signals of different sensor modalities and represent them as an image. The resulting images are then recognized using a 2D convolutional neural network.

lutional [41] networks. The Graph Vertex Feature Encoder learns vertex features by encoding skeleton data into a new feature space. While the Dilated Hierarchical Temporal Convolutional Network introduces new convolutional layers capturing temporal dependencies. Currently their approach is leading on NTU-RGB+D 120 [1] dataset. However, their specialization in skeletal representations does not allow direct adaption on different sensor modalities.

Interesting fusion approaches have been presented previously. Perez et al. [42] presented an approach for multi-modal fusion architecture search using RGB, depth and skeleton fusion. Song et al. [43] extract visual features from different modalities around skeletal joints from RGB and optical flow representations. Whereas those approaches have focused on multiple modalities originating from one device (e.g. Microsoft Kinect) there are also methods for the fusion of sensor data from different devices. Imran et al. [44] propose a three-stream architecture, with different sub-architectures per modality. A 1D-CNN for gyroscopic data, a 2D-CNN for a flow-based image classification and an RNN for skeletal classification. In the end, individual features are fused and a class label is predicted. The fused results are promising and additional modalities improved the results. Additional augmentation by signal filter methods has shown to influence the result positively as well. However, the complexity of the architecture and their sub-architectures require engineering and training overhead and lead to increased run-times by each added modality. This is an issue that we overcome by using a common representation for different modalities. Chen et al. [3] fuse depth information, inertial and demonstrate positive influence. However, they also use two different approaches for each modality. Namely, they use depth motion maps for depth sequences and partitioned temporal windows for signal classification of the gyroscope signals. Most fusion methods rely on complex individual representations per modality or propose complex multi-stream architectures. In contrast, our approach allows modality fusion using matrix concatenations in a single stream. However, our approach is limited to data which can be represented as 1D signals over time. By this, our approach is directly usable for a variety of sensors used in robotics like inertial measurement units, MoCap systems or skeleton sequences and can integrate features extracted from higher dimensional image streams that result e.g. in human pose features [16].

III. Approach

The problem of action recognition with a given set of k actions $Y = \{0, \dots, k\}$ can be reformulated as a classification problem where a mapping $f : \mathbb{R}^{N \times M} \rightarrow Y$ must be found that assigns an action label to a given input. The input in our case is a Matrix $\mathbf{S} \in \mathbb{R}^{N \times M}$ 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.

After signal reduction the reduced signal matrix $\mathbf{S}_{focused}$ is transformed to an RGB image $\mathbf{I} \in \{0, \dots, 255\}^{H \times W \times 3}$ 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. An overview of our approach is given in Fig. 2.

A. Signal reduction

To avoid cluttering of the signal representation we propose a straightforward method for signal reduction which can be used across different modalities. This allows to lay focus on signals with high information content while removing the ones with low information content.

If for example sequences of skeletons are considered many of the joints are not moving significantly throughout the performance of an action. Intuitively it can be understood that when an action is performed while standing in one place the signal of the leg movement does not contribute much to help in classifying the performed action. From this intuition we developed the assumption that low variance signals do contain less information in the context of action recognition as high variance signals. Therefore we propose to set the signals to zero which are not actively contributing to the action by applying a threshold τ to the signals standard deviation σ. In our experiments τ was defined as 20% of the maximum value of all signals.

To be more concise we define the decision function $f(\vec{s}_j)$ for the j-th signal $\vec{s}_j$ in matrix $\mathbf{S}$ as

$$f(\vec{s}_j) = \begin{cases} 1, & \text{if } \sigma(\vec{s}_j) \geq \tau \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

When applying this function to each signal in matrix $\mathbf{S}$ we receive a vector $\vec{c} \in \mathbb{R}^N$ which encodes in each element if the corresponding signal contributes to the action. By element-wise multiplication of each column vector of $\mathbf{S}$ with $\vec{c}\,\mathbf{S}_{focus}$ is received where all signals that do not contribute to the action are set to zero. The signals with low contribution to actions are not removed but set to zero to prevent losing the joint identity (encoded in different colors).

Reducing the signals with low contribution to the action reduces the amount of overlapping signals in the image representation which in turn allows to increase the total number of fused signals. We suggest to apply signal reduction prior to fusion, because different scaling of sensor data can result in the elimination of all signals of a sensor with lower variance as another.

B. Signal fusion

By our formulation the fusion of signals becomes a matrix concatenation:

$$\mathbf{S}_{\text{fused}} = (\mathbf{S}_1|\mathbf{S}_2), \tag{2}$$

where $\mathbf{S}_{\text{fused}}$ is the fusion of $\mathbf{S}_1$ and $\mathbf{S}_2$ under the assumption that both matrices have the same amount of columns, where columns represent the sequence length. This can be either achieved by subsampling the higher frequency signals or interpolating the lower frequency signals. An example for sensor fusion is the encoding of multiple identities i.e. from skeletal data with $\mathbf{S}_{\text{fused}} = (\mathbf{S}_{\text{id1}}|\mathbf{S}_{\text{id2}})$, where two identities are fused. Another example is fusion of two sensor modalities with i.e. $\mathbf{S}_{\text{fused}} = (\mathbf{S}_{\text{skeleton}}|\mathbf{S}_{\text{inertial}})$ or adding interaction context by $\mathbf{S}_{\text{fused}} = (\mathbf{S}_{\text{skeleton}}|\mathbf{S}_{\text{objects}})$. We therefore created a simple framework to support a wide variety of possible applications.

C. 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. We make the initial assumption that temporal relations are represented by the position in the image. However, network architectures of lower depth seem to not maintain a global overview of the input but focuses on local relations. Therefore we encode local temporal information by interpolating from white to the sampled color throughout the sequence length. Signal changes are encoded spatially and joint relation are preserved. Fig. 3 and Fig. 4 give exemplary representations for skeleton and inertial sequences (Fig. 3) and Wi-Fi CSI fingerprints (Fig. 4). 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 are to high dimensional 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.

D. Augmentation

Augmentation methods have shown to successfully influence the generalization. In our case we can create artificial training data on a signal level by interpolating, sampling, scaling, filtering, adding noise to the individual signals or augment the resulting image representation. Liu et al. [34] already proposed to synthesize view independent representations for skeletal motion. As we consider action recognition on a signal level these transformations would result in augmentations integrated as a pre-processing step for each modality separately. Therefore, we decided to focus on augmenting the resulting image representation which can be efficiently integrated into training pipelines. Augmentation applied to the image representation during training still allows interpretation of an effect on the underlying signals. Stretching the width describes the same action but executed in a different speed while perspective changes or rotations can synthesize slightly different executions during the demonstrations.

E. Architecture

Most action recognition approaches based on CNNs present custom architecture designs in their pipelines [34], [45]. A benefit is the direct control over the number of model parameters and 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. [46] suggested the use of residual layers during training resulting in more stable training. Tan et al. [18] recently proposed a novel architecture category based on compound scaling across all dimensions of a CNN. We take advantage of the recent development in architecture design and use an already established architecture for image classification. The recently proposed EfficientNet [18] architecture is especially interesting in the robotics context as it's based on architecture search conditioned on maximizing the accuracy while minimizing the floating-point operations.

F. Implementation

Our implementation is done in Pytorch Lightning [47], [48], which puts a focus on reproducible research. Hyper-parameters and optimizer states are logged directly into the model checkpoints. The source code is made publicly available. We used a re-implementation and pre-trained weights of the EfficientNet [18] architecture. For training we used a Stochastic Gradient Decent optimizer with a learning rate of 0.1 and reduction of learning rate by a factor of 0.1 every 30 epochs with a momentum of 0.9. The learning rate reduction was inspired by He et al. [46]. A batch size of 40 was used on a single Nvidia GeForce RTX 2080 TI with 11GB GDDR-6 memory. We trained for a minimum of 150 epochs and used an early stopping policy based on the accuracy after. Similar model checkpoints were created on an increased validation accuracy. For optimizing the training we

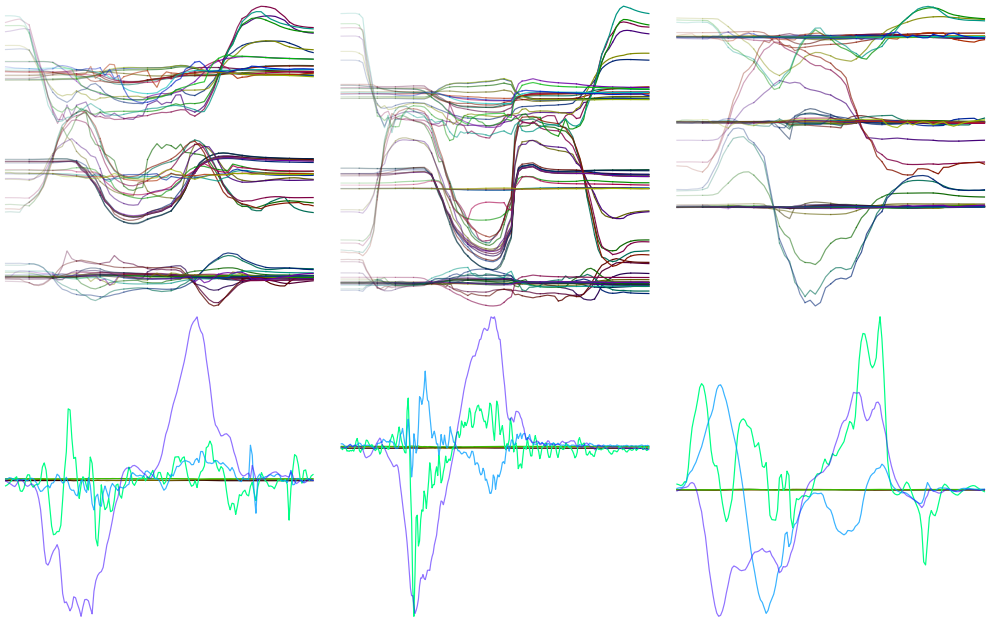

Fig. 3: Sample representations of the UTD-MHAD dataset: (a) and (b) represent the same class (a27) of different subjects. (c) is a sample of a different class (a1). The color encoded lines correspond to the joint signals. On the top the representation for skeletal data is shown and on the bottom their respective inertial data.

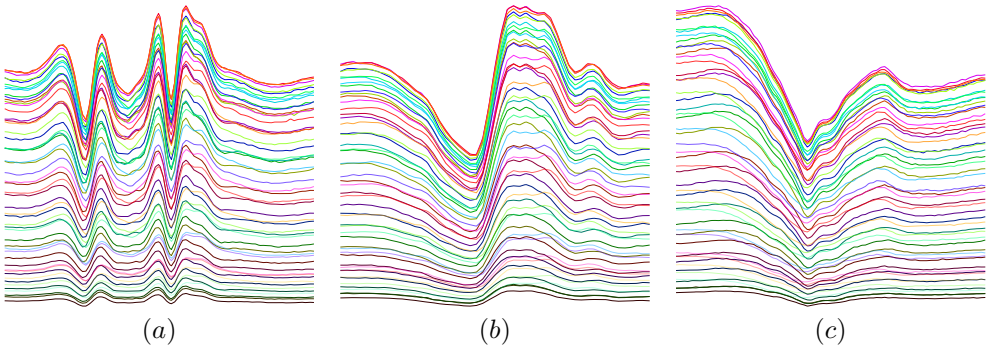

Fig. 4: Sample representations: (a) and (b) represent the same class (0) of different subjects. (c) is a sample of a different class. The color encoded lines correspond to the joint signals.

used a mixed precision approach by training using 16bit float with a 32bit float batch-norm and master weights. A gradient clipping of 0.5 prevented gradient and loss overflows.

IV. Experiments

We conducted experiments on 4 different datasets. The NTU RGB+D 120 [1], UTD-MHAD [3], ARIL [4] and the Simitate [49] dataset. These datasets contain in total 5 modalities. Skeleton sequences are evaluated on the recently released NTU RGB+D 120 [1] and the UTD-MHAD dataset [3]. The NTU RGB+D 120 dataset demonstrates the scaling capabilities of our approach as it contains 120 classes in more than 114000 sequences. The UTD-MHAD dataset [3] provides 27 classes but includes IMU data beside the skeleton estimates. Therefore it is suitable to demonstrate the cross modal capabilities of our approach. We further use it for our fusion experiments. We extend these experiments with activity recognition dataset containing Wi-Fi CSI fingerprints [4] and Motion Capturing data from the Simitate [49] dataset. For our experiments we generated the representations of the datasets prior and used an EfficientNet-B2 [18] architecture for classification. AIS in the tables denotes the additional augmentation of the training signals in image space. Results are compared to other approaches in the next section.

A. Datasets

In the following the datasets on which the experiments where performed are introduced.

1) NTU RGB+D 120: The NTU RGB+D 120 [1] 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. Cross-view and cross-subject splits are defined as protocols. For the cross-subject evaluation, the dataset is split into 53 training subjects and 53 testing subjects as reported by the dataset authors [1]. For the

TABLE I: Results on NTU RGB+D 120. Units are in %.

Approach	CS	CV
Part Aware LSTM [2]	25.5	26.3
Soft RNN [50]	36.3	44.9
Spatio-Termoral LSTM [40]	55.7	57.9
GCA-LSTM et al. [51]	58.3	59.2
Skeleton Visualization (Single Stream) [34]	60.3	63.2
Two-Stream Attention LSTM [52]	61.2	63.3
Multi-Task CNN with RotClips [53]	62.2	61.8
Body Pose Evolution Map [33]	64.6	66.9
SkeleMotion [31]	67.7	66.9
TSRJI [32]	67.9	62.8
Ours (AIS)	*70.8*	*71.6*
ST-GCN + AS-GCN w/DH-TCN [19]	**78.3**	**79.2**

TABLE II: Results on UTD-MHAD. Units are in %.

Approach	Accuracy
Zhao et al. [54]	92.8
Wang et al. [30]	85.8
Chen et al. (Kinect DMMs) [3]	66.1
Chen et al. (Inertial) [3]	67.2
Chen et al. (Fused) [3]	79.1
Ours (Skeleton)	*91.1*
Ours (Skeleton, AIS)	**93.3**
Ours (Inertial)	*72.9*
Ours (Inertial, AIS)	*81.6*
Ours (Fused)	*76.1*
Ours (Fused, AIS)	*86.5*

cross-setup evaluation, the dataset sequences with odd setup ids are reserved while the remainder is used for training. Resulting in 16 setups used during training and 16 used for testing. Results are given in Table I and are discussed in the next section.

2) UTD-MHAD: This dataset [3] 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 hand or the leg during the movements. A cross-subject protocol is followed as proposed by the authors [3]. Half of the subjects are used for training while the other half is used for validation. Results are given in Table II.

3) ARIL: This dataset [4] contains Wi-Fi Channel State Information (CSI) fingerprints. The CSI describes how wireless signals propagate from the transmitter to the receiver [55]. A standard IEEE 802.11n Wi-Fi protocol was used to collect 1398 CSI fingerprints for 6 activities. The data is varying by location. The 6 classes represent hand gestures *hand circle, hand up, hand cross, hand left, hand down, and hand right* targeting the control of smart home devices. For our experiments, we use the same train/test split as was used by the authors of the dataset (1116 train sequences / 278 test sequences). Results are given in Table III.

4) Simitate [49]: The Simitate benchmark focuses on robotic imitation learning tasks. Hand and object data are provided from a motion capturing system in 1932 sequences containing 27 classes of different complexity. 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. This dataset is interesting as we can fuse human and object measurements from the motion capturing system to add context information. Good action recognition capabilities will allow direct application to symbolic imitation approaches. We use a 80/20 train/test split for our experiments. Results are given in Table IV.

TABLE III: Results on ARIL dataset. Units are in %.

Approach	Accuracy
Wang et al. [4]	88.1
Ours (Raw)	91.2
Ours (AIS)	**94.9**

TABLE IV: Results on Simitate. Units are in %.

Approach	Accuracy
Ours (Raw)	95.7
Ours (AIS)	96.1

B. Results

We did our best to include results from the most recent approaches for comparison. We found that the proposed representation on a signal level archived good performances across different modalities. An improvement of +6.8% over the baseline has been achieved on a Wi-Fi CSI fingerprint-based dataset [4]. Augmentation has shown a positive impact on the resulting accuracy across modalities. The resulting model based on an EfficientNet-B2 performs well in interpreting spatial relations on the color encoded signals across the experiments. For the NTU RGB+D 120 dataset we give results in Table I. Related results are taken from literature [1], [31], [19]. A skeleton with 25 joints serves as input for the training of our model. In case multiple identities are contained they are fused with the presented signal fusion approach. We got a cross-subject accuracy of 70.8% and a cross-view accuracy of 71.6% without investment of dataset-specific model tuning. Intuitively, when considering sequential data, LSTM based approaches are considered. We highly outperform the LSTM based approaches [2], [40], [51], [52]. More directly comparable are CNN based approaches [34], [53], [33], [32], [31]. All of the mentioned approaches concentrate on finding representations limited to skeleton or human pose features while our approach considers action recognition on a signal level and therefore is transferable to other modalities as well. The discriminative representation we suggest comes closest to the one by Liu et al. [34]. In combination with the proposed augmentation method and the EfficientNet-B2 based architecture, we outperform the current CNN based approaches by +2.9% (cross-subject), +4.6% (cross-view). Very recently Papadopoulos et al. [19] presented an approach based on a graph convolutional network and performs 5.7% better on the cross-subject split and 8% better on a cross-view split than our approach. However, this approach is also limited to graphs constructed from skeleton sequences. Graph convolutional networks could be an interesting candidate for experiments on multiple modalities in the future.

Results on the UTD-MHAD dataset are shown in Table II. We compare our approach to the baseline of the authors as well as a more recent approach [54], [30]. While Zhao et

al. [54] perform better than our proposed approach we get slightly better results then Wang et al. [30] and further have the benefit of being applicable on other sensor modalities. It is to note that the perfect accuracy of 100.0% in [21] was falsely reported on a similar named dataset. Fused experiments are executed by fusing skeleton estimates and inertial measurements $\mathbf{S}_{\text{fused}} = (\mathbf{S}_{\text{skeleton}}|\mathbf{S}_{\text{inertial}})$. We improve the UTD-MHAD inertial baseline [3] by +14.4% and the UTD-MHAD skeleton [54] baseline by +0.5%. The proposed augmentation improved results by +2.2% for Skeletons, by +8.8% for IMU data and +10.3% for the fusion with the proposed augmentation methods. Fusion in our experiments did now have an overall positive effect. The inertial measurements seem to negatively bias the predicted action. Additional sensor confidence encoding could guide future research. The experiments we conducted on the ARIL dataset are compared to a 1D-ResNet CNN [4] architecture proposed by the datasets authors. Results are presented in Table III. Our approach performs better by +3.1% and the additional proposed augmentation methods improved the baseline by +6.8%. Wi-Fi CSI fingerprints have the benefit of being separated by their 52 bands already. Signal reduction is therefore not necessary. The additional proposed augmentation methods increase the accuracy by another 3.7%.

On the Simitate dataset a high accuracy is achieved on an 80/20 train/test split. Results are given in Table IV. Augmentation on this dataset yields only a minimal improvement. This dataset is especially interesting for adding context. In addition to the hand poses the object poses can be added by our proposed signal fusion approach. As of now, there are no comparable results published. But the results suggest applicability for symbolic imitation approaches in the future.

Most approaches focus on getting high accuracy on a single modality, whereas our approach on a signal level serves as an interesting framework for multi-modal action recognition. In total, we have shown good results across 4 modalities (Skeleton, IMU, MoCap, Wi-Fi). To the authors knowledge, no experiment with a similar extend is known. A huge benefit is the common representation that allows immediate prototyping. Run times are constant, even when additional context or sensors are added due to the representation level fusion. The EfficientNet-B2 architecture serves as a good basis for action recognition on our representation. Additional augmentation has improved the accuracy across the conducted experiments.

V. Conclusion

We propose to transform individual signals of different sensor modalities and represent them as an image. The resulting images are then classified using a EfficientNet-B2 architecture. Our approach was evaluated on action recognition datasets based on skeleton estimates, inertial measurements, motion capturing data and Wi-Fi CSI fingerprints. This is in contrast to many previously proposed approaches that often focus on action recognition on a single modality. For skeleton data we represent each joint and their respective axis as individual signals. For Wi-Fi we use each of the 52 CSI fingerprint channels as signals. For inertial measurement units we use each axis of the acceleration and angular velocity. For our motion capturing experiments we used each axis of the marker attached to the hand and the interacting objects. Additional context like subjects and object estimates or even the fusion of different modalities can be flexibly added by a matrix concatenation. As our approach is limited to sparse signals, we propose filtering methods on a signal level to reduce signals that do not contribute much to the action. By this, additional information can be added without overloading the image representation. We evaluated our approach on four different datasets. The NTU 120 dataset for skeleton data, the UTD-MHAD dataset for skeleton and inertial data, the ARIL dataset for Wi-Fi data and the Simitate dataset for motion capturing data. Experimental results show that our approach is achieving good results across the different sensor modalities.

References

[1] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[3] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE International conference on image processing (ICIP)*. IEEE, 2015, pp. 168–172.

[4] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with wifi fingerprints," *IEEE Access*, vol. 7, pp. 80 058–80 068, 2019.

[5] N. Noury, A. Fleury, P. Rumeau, A. Bourke, G. Laighin, V. Rialle, and J. Lundy, "Fall detection-principles and methods," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 1663–1666.

[6] M. D. Solbach and J. K. Tsotsos, "Vision-based fallen person detection for the elderly," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1433–1442.

[7] Q. Ni, A. B. Garcia Hernando, D. la Cruz, and I. Pau, "The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development," *Sensors*, vol. 15, no. 5, pp. 11 312–11 362, 2015.

[8] P. Lago, C. Roncancio, and C. Jiménez-Guarín, "Learning and managing context enriched behavior patterns in smart homes," *Future Generation Computer Systems*, vol. 91, pp. 191–205, 2019.

[9] W. Niu, J. Long, D. Han, and Y.-F. Wang, "Human activity detection and recognition for video surveillance," in *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*, vol. 1. IEEE, 2004, pp. 719–722.

[10] A. Wiliem, V. Madasu, W. Boles, and P. Yarlagadda, "A suspicious behaviour detection using a context space model for smart surveillance systems," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 194–209, 2012.

[11] V. Krüger, D. Kragic, A. Ude, and C. Geib, "The meaning of action: A review on action recognition and mapping," *Advanced robotics*, vol. 21, no. 13, pp. 1473–1501, 2007.

[12] K. Charalampous, I. Kostavelis, and A. Gasteratos, "Robot navigation in large-scale social maps: An action recognition approach," *Expert Systems with Applications*, vol. 66, pp. 261–273, 2016.

[13] S. Choi, J. Kim, D. Kwak, P. Angkititrakul, and J. H. Hansen, "Analysis and classification of driver behavior using in-vehicle can-bus information," in *Biennial workshop on DSP for in-vehicle and mobile systems*, 2007, pp. 17–19.

[14] F. Martinelli, F. Mercaldo, A. Orlando, V. Nardone, A. Santone, and A. K. Sangaiah, "Human behavior characterization for driving style recognition in vehicle system," *Computers & Electrical Engineering*, 2018.

[15] M. Rigolli, Q. Williams, M. J. Gooding, and M. Brady, "Driver behavioural classification from trajectory data," in *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.* IEEE, 2005, pp. 889–894.

[16] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *European Conference on Computer Vision (ECCV)*, 2018.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[18] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[19] K. Papadopoulos, E. Ghorbel, D. Aouada, and B. Ottersten, "Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition," *arXiv preprint arXiv:1912.09745*, 2019.

[20] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[21] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, 2019.

[22] Z. Wang, Z. Yang, and T. Dong, "A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time," *Sensors*, vol. 17, no. 2, p. 341, 2017.

[23] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[24] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognition Letters*, vol. 42, pp. 11–24, 2014.

[25] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[26] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

[27] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 12, pp. 2430–2443, 2016.

[28] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[29] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[30] P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," *Knowledge-Based Systems*, vol. 158, pp. 43–53, 2018.

[31] C. Caetano, J. Sena, F. Brémond, J. A. d. Santos, and W. R. Schwartz, "Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition," *arXiv preprint arXiv:1907.13025*, 2019.

[32] C. Caetano, F. Brémond, and W. R. Schwartz, "Skeleton image representation for 3d action recognition based on tree structure and reference joints," *arXiv preprint arXiv:1909.05704*, 2019.

[33] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1159–1168.

[34] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[35] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2017, pp. 1623–1631.

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[37] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2904–2913.

[38] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[39] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[40] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[41] 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.

[42] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6966–6975.

[43] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Skeleton-indexed deep multi-modal feature learning for high performance human action recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[44] J. Imran and B. Raman, "Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–20, 2019.

[45] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3288–3297.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[48] W. e. a. Falcon, "Pytorch lightning," https://github.com/PytorchLightning/pytorch-lightning, 2019.

[49] R. Memmesheimer, I. Mykhalchyshyna, V. Seib, and D. Paulus, "Simitate: A hybrid imitation learning benchmark," *arXiv preprint arXiv:1905.06002*, 2019.

[50] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J.-H. Lai, and J. Zhang, "Early action prediction by soft regression," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[51] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1647–1656.

[52] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2017.

[53] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "Learning clip representations for skeleton-based 3d action recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2842–2855, 2018.

[54] R. Zhao, W. Xu, H. Su, and Q. Ji, "Bayesian hierarchical dynamic model for human action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7733–7742.

[55] M. A. Al-qaness, M. Abd Elaziz, S. Kim, A. A. Ewees, A. A. Abbasi, Y. A. Alhaj, and A. Hawbani, "Channel state information from pure communication to sense and track human motion: A survey," *Sensors*, vol. 19, no. 15, p. 3329, 2019.