Fusion-GCN: Multimodal Action Recognition using Graph Convolutional Networks

Michael Duhme, Raphael Memmesheimer[0000−0003−3602−754X], and Dietrich Paulus[0000−0002−2967−5277]

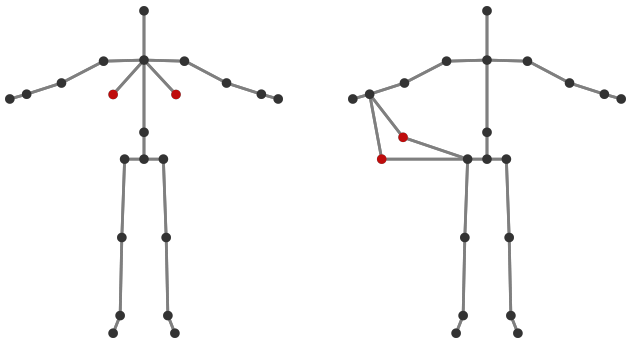
Arbeitsgruppe Aktives, University of Koblenz-Landau, Koblenz, Germany
{mduhme,raphael,paulus}@uni-koblenz.de.de

Abstract. In this paper we present Fusion-GCN, an approach for multimodal action recognition using Graph Convolutional Networks (GCNs). Action recognition methods based around GCNs recently yielded state-of-the-art performance for skeleton-based action recognition. With Fusion-GCN, we propose to integrate various sensor data modalities into a graph that is trained using a GCN model for multi-modal action recognition. Additional sensor measurements are incorporated into the graph representation either on a channel dimension (introducing additional node attributes) or spatial dimension (introducing new nodes). Fusion-GCN was evaluated on two publicly available datasets, the UTD-MHAD- and MMACT datasets, and demonstrates flexible fusion of RGB sequences, inertial measurements and skeleton sequences. Our approach gets comparable results on the UTD-MHAD dataset and improves the baseline on the large-scale MMACT dataset by a significant margin of up to 12.37% (F1-Measure) with the fusion of skeleton estimates and accelerometer measurements.

1 Introduction

Automatic Human Action Recognition (HAR) is a research area that is utilized in various fields of application where human monitoring is infeasible due to the amount of data and scenarios where quick reaction times are vital, such as surveillance and real-time monitoring of suspicious and abnormal behavior in public areas [12, 33, 34, 49] or intelligent hospitals and healthcare sectors [8, 9] with scenarios such as fall detection [36, 45], monitoring of medication intake [13] and detection of other potentially life-threatening situations [8]. Additional areas of applications include video retrieval [40], robotics [41], smart home automation [21], autonomous vehicles [52]. In recent years, approaches based on neural networks, especially GCNs, like ST-GCN [51] or 2s-AGCN [43], have achieved state-of-the-art results in classifying human actions from time series data.

GCNs can be seen as an extension to Convolutional Neural Networks (CNNs) that work on graph-structured data [19]. Its network layers operate by including a binary or weighted adjacency matrix, that describes the connections between each of the individual graph nodes. As of now, due to their graph-structured representation in the form of joints (graph nodes) and bones (graph edges), research for HAR using GCNs is mostly limited to skeleton-based recognition. However, the fusion of additional modalities into GCNs models are currently neglected. For that reason, taking skeleton-based action



(a) Appended to center joint (b) Appended to right wrist/hip

Fig. 1: Showing the skeleton as included in UTD-MHAD. IMU nodes are either appended to the central node (neck) or to both the right wrist and right hip. Two additional representations arise when all newly added nodes are themselves connected by edges.

recognition as the foundation, our objective is to research possibilities of incorporating other vision-based modalities and modalities from worn sensors into existing GCN models for skeleton-based action recognition through data fusion and augmentation of skeleton sequences. Figure 1 gives an example of two suggestions on how inertial measurements can be incorporated into a skeleton graph. To the best of our knowledge, Fusion-GCN is the first approach proposing to flexibly incorporate additional sensor modalities into the skeleton graph for HAR. We evaluated our approach on two multimodal datasets, UTD-MHAD [5] and MMACT [20].

The contributions of this paper can be summarized as: (1) We propose the fusion of multiple modalities by incorporating sensor measurements or extracted features into a graph representation. The proposed approach significantly lifts the state-of-the-art on the large-scale MMACT dataset. (2) We propose modality fusion for GCNs on two dimensionality levels: (a) the fusion at a channel dimension to incorporate additional modalities directly into the already existing skeleton nodes, (b) the fusion at a spatial dimension, to incorporate additional modalities as new nodes spatially connected to existing graph nodes. (3) We demonstrate applicability of the flexible fusion of various modalities like skeleton, inertial, RGB data in an early fusion approach.

The code for Fusion-GCN to reproduce and verify our results is publicly available on https://github.com/mduhme/fusion-gcn.

2 Related Work

In this section, we present related work from the skeleton-based action recognition domain that is based on GCN and further present recent work on multimodal action recognition.

Skeleton-based Action Recognition Approaches based on GCNs have recently shown great applicability on non-Euclean data [38] like naturally graph-structure represented

skeletons and have recently defined the state-of-the-art. Skeletons, as provided by large-scale datasets [42], commonly are extracted from depth cameras [44]. RGB images can be transformed into human pose feature that yield a similar skeleton-graph in 2D [4, 22, 29] and in 3D [14, 29, 31]. All of those approaches output skeleton-graphs that are suitable as input for our fusion approach as a base structure for the incorporation of additional modalities. The Spatial-Temporal Graph Convolutional Network (ST-GCN) [51] is one of the first proposed models for skeleton-based HAR that utilizes GCNs based on the propagation rule introduced by Kipf and Welling [19]. The Adaptive Graph Convolutional Network (AGCN) [43] builds on these fundamental ideas with the proposal of learning the graph topology in an end-to-end-manner. Peng et al. [38] propose a Neural Architecture Search (NAS) approach for finding neural architectures to overcome the limitations of GCN caused by fixed graph structures. Cai et al. [1] proposes to add flow patches to handle subtle movements into a GCN. Approaches based on GCN [6, 23, 37, 47] have been constantly improving the state-of-the-art on skeleton-based action recognition recently.

Multimodal Action Recognition Cheron et al. [7] design CNN input features based on the positions of individual skeleton joints. Here, human poses are applied to RGB images and optical flow images. The pixel coordinates that represent skeleton joints are then grouped hierarchically starting from smaller body parts, such as arms, and upper body to full body. For each group, an RGB image and optical flow patch is cropped and passed to a 2D-CNN. The resulting feature vectors are then processed and concatenated to form a single vector, which is used to predict the corresponding action label. Similarly, Cao et al. [2] propose to fuse pose-guided features from RGB-Videos. Cao et al. [3] further, refine this method by using different aggregation techniques and an attention model. Islam and Iqbal [15] propose to fuse data of RGB, skeleton and inertial sensor modalities by using a separate encoder for each modality to create a similar shaped vector representation. The different streams are fused using either summation or vector concatenation. With Multi-GAT [16] an additional message-passing graphical attention mechanism was introduced. Li et al. [24] propose another architecture that entails skeleton-guided RGB features. For this, they employ ST-GCN to extract a skeleton feature vector and R(2+1)D [48] to encode the RGB video. Both output features are fused either by concatenation or by compact bilinear correlation.

The above-mentioned multimodal action recognition approaches follow a late-fusion method, that fuse various models for each modality. This allows a flexible per modality model-design, but comes at the computational cost of the multiple streams that need to be trained. For early fusion approaches, multiple modalities are fused on a representation level [32], reducing the training process to a single model but potentially loosing the more descriptive features from per-modality models. Kong et al. [20] presented a multi modality distillation model. Teacher models are trained separately using a 1D-CNN. The semantic embeddings from the teaching models are weighted with an attention mechanism and are ensembled with a soft target distillation loss into the student network. Similarly, Liu et al. [27] utilize distilled sensor information to improve the vision modality. Luo et al. [30] propose a graph distillation method to incorporate rich privileged information from a large-scale multi-modal dataset in the source domain, and improves the learning in the target domain More fundamentally, multimodality in neu-

ral networks is recently also tackled by the multimodal neurons that respond to photos, conceptual drawings and images of text [10]. Joze et al. [17] propose a novel intermediate fusion scheme in addition to early and late-fusion, they share intermediate layer features between different modalities in CNN streams. Perez-Rua et al. [39] presented an approach for finding neural architecture search for the fusion of multiple modalities. To the best of our knowledge, our Fusion-GCNapproach is the first that proposes to incorporate additional modalities directly into the skeleton-graphs as an early fusion scheme.

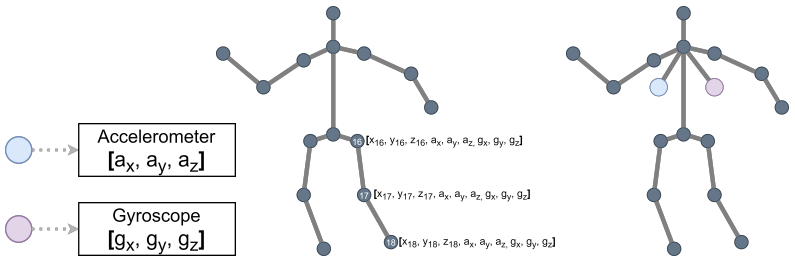
3 Approach

In the context of multimodal action recognition, early and late fusion methods have been established to either fuse on a representation or feature level. We present approaches for fusion of multiple modalities at representation level to create a single graph which is passed to a GCN.

3.1 Incorporating additional Modalities into a Graph Model

Early fusion denotes the combination of structurally equivalent streams of data before sending them to a larger (GCN) model, whereas late fusion combines resulting outputs of multiple neural network models. For early fusion, one network handles multiple data sources which are required to have near identical shape to achieve fusion. As done by Song et al. Song et al. [46], each modality may be processed by some form of an encoder to attain a common structure before being fused and passed on to further networks. Following a skeleton-based approach, for example, by employing a well established GCN model like ST-GCN or AGCN as the main component, RGB and inertial measurements are remodeled and factored into the skeleton structure. With Fusion-GCN we suggest the flexible integration of additional sensor modalities into a skeleton graph by either adding additional node attributes *(fusion on a channel dimension)* or introducing additional nodes *(fusion at a spatial dimension)*. In detail, the exact possible fusion approach is as follows.

Let $\boldsymbol{X}_{SK} \in \mathbb{R}^{(M,C_{SK},T_{SK},N_{SK})}$ be a skeleton sequence input, where M is the number of actors that are involved in an action, C_{SK} is the initial channel dimension (2D or 3D joint coordinates) and sizes T_{SK} and N_{SK} are sequence length and number of skeleton graph nodes. An input of shape $\mathbb{R}^{(M,C,T,N)}$ is required when passing data to a spatial-temporal GCN model, such as ST-GCN. Furthermore, let $\boldsymbol{X}_{RGB} \in \mathbb{R}^{(C_{RGB},T_{RGB},H_{RGB},W_{RGB})}$ be the shape of an RGB video with channels C_{RGB}, frames T_{RGB} and image size $H_{RGB} \times W_{RGB}$. For sensor data, the input is defined as $\boldsymbol{X}_{IMU} \in \mathbb{R}^{(M,C_{IMU},S_{IMU},T_{IMU})}$, where T_{IMU} is the sequence length, S_{IMU} is the number of sensors and C_{IMU} is the channel dimension. For example, given gyroscope and accelerometer with x-, y- and z-values each, the structure would be $S_{IMU} = 2$ and $C_{IMU} = 3$. Similar to skeleton data, M denotes the person wearing the sensor and its value is equivalent to that of skeleton, that is, $M_{SK} = M_{IMU}$. Considering a multimodal model using a skeleton-based GCN approach, early fusion can now be seen as a



(a) IMU Graph Nodes (b) Fusion at Channel Dimension (c) Fusion at Spatial Dimension

Fig. 2: Options for fusion of skeleton graph and IMU signal values, viewed as skeleton nodes. If both skeleton joint coordinates and wearable sensor signals share a common channel dimension, the skeleton graph can be augmented by simply appending signal nodes at some predefined location.

task of restructuring non-skeleton modalities to be similar to skeleton sequences by finding a mapping $\mathbb{R}^{(C_{RGB},T_{RGB},H_{RGB},W_{RGB})} \rightarrow \mathbb{R}^{(M,C,T,N)}$ or $\mathbb{R}^{(M,C_{IMU},S_{IMU},T_{IMU})} \rightarrow \mathbb{R}^{(M,C,T,N)}$ with some C, T and N. This problem can be reduced: If the sequence length of some modalities is different, $T_{SK} \neq T_{RGB} \neq T_{IMU}$, a common T can be achieved by resampling T_{RGB} and T_{IMU} to be of the same length as the target modality T_{SK}. Early fusion is then characterized by two variants of feature concatenation to fuse data:

1. Given $\boldsymbol{X}_{SK}$ and an embedding $\boldsymbol{X}_E \in \mathbb{R}^{(M,C_E,T,N_E)}$ with sizes C_E and N_E where $N = N_{SK} = N_E$, fusion at the channel dimension means creating a fused feature $\boldsymbol{X} \in \mathbb{R}^{(M,C_{SK}+C_E,T,N)}$. An example is shown in Figure 2b.
2. Given an embedding where $C = C_{SK} = C_E$ instead, a second possibility is fusion at the spatial dimension, that is, creating a feature $\boldsymbol{X} \in \mathbb{R}^{(M,C,T,N_{SK}+N_E)}$. Effectively, this amounts to producing $M \cdot T \cdot N_E$ additional graph nodes and distributing them to the existing skeleton graph at each time step by resizing its adjacency matrix and including new connections. In other words, the already existing skeleton graph is extended by multiple new nodes with an identical number of channels. An example is shown in Figure 2c.

The following sections introduce multiple approaches for techniques about the early fusion of RGB video and IMU sensor modalities together with skeleton sequences by outlining the neural network architecture.

3.2 Fusion of Skeleton Sequences and RGB Video

This section explores possibilities for fusion of skeleton sequences and 2D data modalities. Descriptions and the following experiments are limited to RGB video, but all introduced approaches are in the same way applicable to depth sequences. As previously established, early fusion of RGB video and skeleton sequences in preparation for a skeleton-based GCN model is a problem of finding a mapping $\mathbb{R}^{(C_{RGB},H_{RGB},W_{RGB})} \rightarrow$

$\mathbb{R}^{(M,C,N)}$. An initial approach uses a CNN to compute vector representations of $N \cdot M \cdot T$ skeleton-guided RGB patches that are cropped around projected skeleton joint positions. Inspired by the work of Wang et al. [50] and Norcliffe-Brown et al. [35], a similar approach involves using an encoder network to extract relevant features from each image of the RGB video. This way, instead of analyzing $N \cdot M \cdot T$ cropped images, the T images of each video are utilized in their entirety. A CNN is used to extract features for every frame and fuse the resulting features with the corresponding skeleton graph, before the fused data is forwarded to a GCN. By running this procedure as part of the training process and performing fusion with skeleton sequences, the intention is to let the encoder network extract those RGB features that are relevant to the skeleton modality. For example, an action involving an object cannot be fully represented by merely the skeleton modality because an object is never part of the extracted skeleton. Objects are only visible in RGB video.

3.3 Fusion of Skeleton Sequences and IMU Signals

Fusion of skeleton and data from wearable sensors, such as IMUs, is applicable in the same way as described in the fusion scheme from the previous section. In preparation to fuse both modalities, they again need to be adjusted to have an equal sequence length first. Then, assuming both the skeleton joint coordinates and the signal values have a common channel dimension $C = 3$ and because $M_{SK} = M_{IMU}$, since all people wear a sensor, the only differing sizes between skeleton modality and IMU modality are N, the number of skeleton graph nodes, and S, the number of sensor signals. Leaving aside its structure, the skeleton graph is a collection of N nodes. A similar understanding can be applied to the S different sensors. They can be understood as a collection of S graph nodes (see Figure 2a). The fusion of sensor signals with the skeleton graph is therefore trivial because the shape is almost identical. According to channel dimension fusion as described in the previous section, the channels of all S signals can be broadcasted to the x-, y- and z-values of all N skeleton nodes to create the GCN input feature $\boldsymbol{X} \in \mathbb{R}^{(M,(1+S)\cdot C,T,N)}$, as presented in Figure 2b. The alternative is to append all S signal nodes onto the skeleton graph at some predefined location to create the GCN input feature $\boldsymbol{X} \in \mathbb{R}^{(M,C,T,N+S)}$, as illustrated in Figure 2c. Similar to the RGB fusion approaches, channel dimension fusion does not necessarily require both modalities to have the same dimension C if vector concatenation is used. In contrast, the additional nodes are required to have the same dimension as all existing nodes if spatial dimension fusion is intended.

3.4 Combining Multiple Fusion Approaches

All the introduced fusion approaches can be combined into a single model, as illustrated by Figure 3. First, the RGB modality needs to be processed using one of the variants discussed in section 3.2. Ideally, this component runs as part of the supervised training process to allow the network to adjust the RGB feature extraction process based on the interrelation of its output with the skeleton graph. Similarly, sensor signals need to be processed using one of the variants discussed previously for

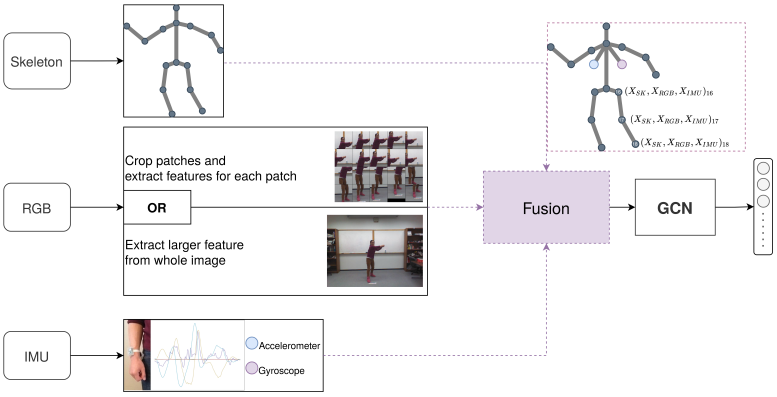


Fig. 3: All described approaches can be flexible fused together for early fusion and passed to a GCN. Fusion can be realized independent of a channel or spatial fusion dimension. Here we give an example of a mixed (channel and spatial) fusion.

that modality. Assuming all sequences are identical in length, to combine the different representations, let $\boldsymbol{X}_{SK} \in \mathbb{R}^{(M,C_{SK},T,N_{SK})}$ be the sequence of skeleton graphs. For RGB, let $\boldsymbol{X}_{RGB1} \in \mathbb{R}^{(M,C_E,T,N)}$ be the C_E-sized channel features obtained from computing individual patch features or feature extraction for the whole image or $\boldsymbol{X}_{RGB2} \in \mathbb{R}^{(M,C,T,N_E)}$ be the RGB feature representing additional graph nodes. Respectively, the two variants of generated IMU features are $\boldsymbol{X}_{IMU1} \in \mathbb{R}^{(M,S\cdot C_{IMU},T,N)}$ or $\boldsymbol{X}_{IMU2} \in \mathbb{R}^{(M,C_{IMU},T,S)}$. The following possibilities to fuse different combinations of these representations arise.

- $(\boldsymbol{X}_{SK}, \boldsymbol{X}_{RGB1}, \boldsymbol{X}_{IMU1}) \rightarrow \boldsymbol{X}_{FUSED} \in \mathbb{R}^{(M,C_{SK}+C_E+S\cdot C_{IMU},T,N_{SK})}$ is the feature when combining modalities at channel dimension by vector concatenation.
- $(\boldsymbol{X}_{SK}, \boldsymbol{X}_{RGB1}, \boldsymbol{X}_{IMU2}) \rightarrow \boldsymbol{X}_{FUSED} \in \mathbb{R}^{(M,C_{SK}+C_E,T,N_{SK}+S)}$ combines skeleton with computed RGB features at channel dimension and expands the skeleton graph by including additional signal nodes. Since $C_{IMU} = C_{SK}$, the newly added nodes also need to be extended to have $C_{SK} + C_E$ channels. In contrast to skeleton nodes, there exists no associated cropped patch or RGB value. Therefore, the remaining C_E values can be filled with zeros. Conversely, the same applies when replacing $\boldsymbol{X}_{RGB1}$ with $\boldsymbol{X}_{RGB2}$ and $\boldsymbol{X}_{IMU2}$ with $\boldsymbol{X}_{IMU1}$.
- $(\boldsymbol{X}_{SK}, \boldsymbol{X}_{RGB2}, \boldsymbol{X}_{IMU2}) \rightarrow \boldsymbol{X}_{FUSED} \in \mathbb{R}^{(M,C,T,N_{SK}+N_E+S)}$ introduces new nodes for both RGB and signal modalities. This is accomplished by appending them to a specific location in the graph.

4 Experiments

We conducted experiments on two public available datasets and various modality fusion experiments. If not stated otherwise we use the top-1 accuracy as reporting metric for the final epoch of the trained model.

4.1 Datasets

UTD-MHAD UTD-MHAD [5] is a relatively small dataset containing 861 samples and 27 action classes, which thereby results in shorter training durations for neural networks. Eight individuals (four females and four males) perform each action a total of four times, captured from a front-view perspective by a single Kinect camera. UTD-MHAD also includes gyroscope and accelerometer modalities by letting each subject wear the inertial sensor on either the right wrist or on the right hip, depending on whether an action is primarily performed using the hands or the legs. For the following experiments using this dataset, the protocol from the original paper [5] is used.

MMACT The MMACT dataset [20] contains more than 35k data samples and 35 available action classes. With 20 subjects and four scenes with four currently available different camera perspectives each, the dataset offers a larger variation of scenarios. RGB videos are captured with a resolution of 1920×1080 pixels at a frame rate of 30 frames per second. For inertial sensors, acceleration, gyroscope and orientation data is obtained from a smartphone carried inside the pocket of a subject's pants. Another source for acceleration data is a smartwatch, resulting in data from four sensors in total. For the following experiments using this dataset, the protocol from the original paper [20] is used which proposes a cross-subject and a cross-view split. Since skeleton sequences are missing in the dataset, we create them from RGB data using OpenPose [4].

4.2 Implementation

Models are implemented using PyTorch 1.6 and trained on a Nvidia RTX 2080 GPU with 8GB of video memory. To guarantee a deterministic and reproducible behavior, all training procedures are initialized with a fixed random seed. Unless stated otherwise, experiments regarding UTD-MHAD use a cosine annealing learning rate scheduler [28] with a total of 60 epochs, warm restarts after 20 and 40 epochs, an initial learning rate of $1\mathrm{e}{-3}$ and ADAM [18] optimization. Experiments using RGB data instead run for 50 epochs without warm restarts. Training for MMAct adopts the hyperparameters used by Shi et al. [43].For the MMACT, skeleton and RGB features were extracted for every third frame for more efficient pre-processing and training. The base GCN model is a single-stream AGCN for all experiments.

4.3 Comparison to the State-of-the-Art

UTD-MHAD Table 1a shows a ranking of all conducted experiments in comparison with other recent state-of-the-art techniques that implement multimodal HAR on UTD-MHAD with the proposed cross-subject protocol. Without GCNs and all perform better than the default skeleton-only approach using a single-stream AGCN. Additionally, another benchmark using GCNs on UTD-MHAD does not exist, thus, making a direct comparison of different approaches difficult. From the listing in Table 1a, it is clear that all fusion approaches skeleton and IMU modalities achieve the highest classification performance out of all methods introduced in this work. In comparison to the

Approach	Acc
Skeleton	92.32
RGB Patch Features R-18	27.67
RGB Encoder R-18	27.21
R(2+1)D	61.63
Skeleton + RGB Encoder R(2+1)D	91.62
Skeleton + RGB Encoder R-18	89.83
Skeleton + RGB Patch Features R-18	73.49
Skeleton + RGB Patch Features R-18 (no MLP)	44.60
Skeleton + IMU (Center)	94.42
Skeleton + IMU (Wrist/Hip)	94.07
Skeleton + IMU (Center + Add. Edges)	93.26
Skeleton + IMU (Wrist/Hip + Add. Edges)	93.26
Skeleton + IMU (Channel Fusion)	90.29
Skeleton + IMU + RGB Patch Features R-18	78.90
Skeleton + IMU + RGB Encoder R-18	92.33
Skeleton + IMU + RGB Encoder R(2+1)D	92.85
PoseMap [25]	**94.50**
Gimme Signals [32]	93.33
MCRL [26]	93.02

(a) UTD-MHAD

Approach	F1-Measure
Skl	88.65
Skl+Acc(W+P)+Gyo+Ori	85.50
Skl+Acc(W+P)+Gyo+Ori (Add. Edges)	84.78
Skl+Acc(W)	89.55
Skl+Acc(P)	88.72
Skl+Gyo	87.41
Skl+Ori	88.64
Skl+Acc(W+P)	**89.60**
SMD [11] (Acc+RGB)	63.89
MMD [20] (Acc+Gyo+Ori+RGB)	64.33
MMAD [20] (Acc+Gyo+Ori+RGB)	66.45
Multi-GAT [16]	75.24
SAKDN [27]	77.23

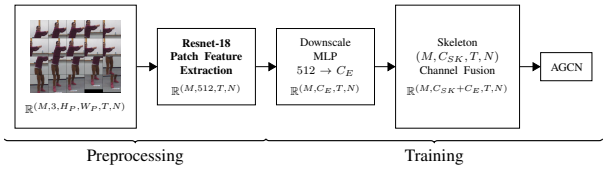
(b) MMAct

Table 1: Comparison to the State-of-the-Art

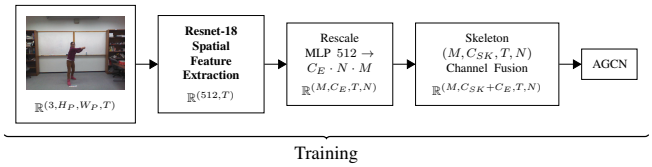
best performing fusion approach of skeleton with IMU nodes appended at its central node. MCRL [26] uses a fusion of skeleton, depth and RGB to reach 93.02% (-1.4%) validation accuracy on UTD-MHAD. Gimme Signals [32] reach 93.33% (-1.09%) using a CNN and augmented image representations of skeleton sequences. PoseMap [25] achieves 94.5% (+0.08%) accuracy using pose heatmaps generated from RGB videos. This method slightly outperforms the proposed fusion approach.

MMACT To show better generalization, we also conducted experiments on the large-scale MMACT dataset which contains more modalities, classes and samples as the UTD-MHAD dataset. Note we only use the cross-subject protocol, the signal modalities can not be seperated by view. A comparison of approaches regarding the MMAct dataset is given in Table 1b. Kong et al. Kong et al. [20] propose along with the MMACT dataset the MMAD approach, a multimodal distillation method utiliziting an attention mechanism that incorporates acceleration, gyroscope, orientation and RGB. For evaluation, they use the F1-measure and reach an average of 66.45%. Without the attention mechanism, the approach (MMD) yields 64.33%. An approach utilizing the standard distillation approach Single Modality Distillation (SMD) yields 63.89%. The current baseline is set by SAKDN [27] which distills sensor information to enhance action recognition for the vision modality. Experiments show that the skeleton-based approach can be further improved by fusion with just the acceleration data to reach a recognition F1-measure of 89.60% (+12.37%). The MMACT dataset contains two accelerometers, where only the one from the smartwatch yields a mention-able improvement. The most significant improvement of our proposed approach is yielded by introducing the skeleton graph. In contrast, while the fusion approaches of skeleton and all four sensors do not improve the purely skeleton-based approach of 88.65% (+13.41%), with 85.5% (+10.26%) without additional edges and 84.78% (+9.54%) with additional edges, both reach a higher F1-measure than the baseline but also impact the pure skeleton-based recognition negatively.

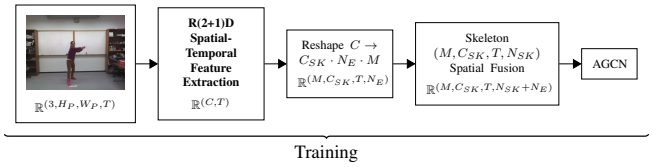
4.4 Ablation Study



(a) Skeleton + RGB Patch Feature Fusion



(b) Skeleton + Resnet-18 Generated Feature Fusion



(c) Skeleton + Modified R(2+1)D Generated Feature Fusion

Fig. 4: The three different skeleton + RGB Fusion models with reference of an image from UTD-MHAD. The first model generates a feature for each node, while the last two generate a feature for the entire image that is distributed to the nodes and adjusted as part of the supervised training.

Fusion of Skeleton and RGB Skeletons and RGB videos are combined using the three approaches depicted in Figure 4. Figure 4a shows an approach using RGB patches that are cropped around each skeleton node and passed to a Resnet-18 to compute a feature vector $\boldsymbol{X}_{RGB} \in \mathbb{R}^{(M,512,T,N)}$ as part of preprocessing. The second approach, shown in Figure 4b, uses Resnet-18 to compute a feature vector for each image. The resulting feature vector is rescaled to the size $C_E \cdot N \cdot M$ and reshaped to be able to be fused with skeleton data. Similarly, in Figure 4c, the third approach uses R(2+1)D. In terms of parameters, the basis Skeleton model has 3.454.099 parameters, only 2.532 parameters are added for incorporation of inertial measurements into the model Skeleton+IMU(Center) 3.456.631 for a 2.2% accuracy improvement. Fusion with an RGB encoder adds five times more parameters (Skeleton+RGB Encoder Resnet-18 with 17.868.514) and a massive training overhead.

Table 1a shows that the RGB approaches viewed individually (without fusion) do not reach the performance of R(2+1)D pre-trained for action recognition. Results regarding the fusion models show a low accuracy of 73.49% for RGB patch features that

have been created outside the training process and 44.6% for the same procedure without a downscaling Multilayer Perceptron (MLP). A similar conclusion can be drawn from the remaining two fusion models. Using R(2+1)D to produce features shows a slightly increased effectiveness of +1.79% (91.62%) over Resnet-18 (89.83%) but -0.7% in comparison to the solely skeleton-based approach.

Fusion of Skeleton and IMU Fusion of skeletal and inertial sensor data is done according to Figure 2. Figure 1 shows the skeleton structure of UTD-MHAD and illustrates two possibilities for fusing the red IMU graph nodes with the skeleton by connecting them to different skeleton joint nodes. In Figure 1a, nodes are appended at the central skeleton joint as it is defined in ST-GCN and AGCN papers. The configuration depicted in Figure 1b is attributed to the way sensors are worn by subjects of UTD-MHAD. This configuration is therefore not used for MMAct. Additional configurations arise when additional edges are drawn between the newly added nodes. According to Figure 2b, another experiment involves broadcasting the $\mathbb{R}^6$-sized IMU feature vector to each skeleton joint and fuse them at channel dimension. From the results in Table 1a, it is observable that all skeleton graphs with additional associated IMU nodes at each point in time improve the classification performance by at least one percent. In comparison to a skeleton-only approach, variants with additional edges between the newly added nodes perform generally worse than their not-connected counterparts and are both at 93.26% (+0.94%). The average classification accuracy of both other variants reaches 94.42% (+2.1%) and 94.07% (+1.75%). Despite having a slightly increased accuracy for appending new nodes to the existing central node, both variants almost reach equal performance and the location where nodes are appended seemingly does not matter much. While all experiments with fusion at spatial dimension show increased accuracies, the only experiment that does not surpass the skeleton-based approach is about fusion of both modalities at channel dimension, reaching 90.29% (-2.03%) accuracy.

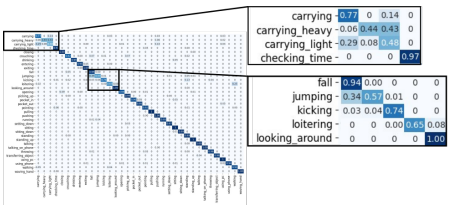


Fig. 5: Confusion matrix for the results on MMACT with the fusion of skeleton and accelerometer measurements from the smartwatch with highlighted high-confused actions.

For MMAct, all experiments are conducted using only the configuration in Figure 1a and its variation with interconnected nodes. Table 1b shows that the skeleton-based approach reaches 87.85% accuracy for a cross-subject split, fusion approaches including all four sensors perform worse and reach only 84.85% (-3%) and 84.4% (-3.45%). Mixed results are achieved when individual sensors are not part of the fusion model. Fusion using only one of the phone's individual sensors, acceleration, gyroscope or orientation, reaches comparable results with 87.70% (-0.15%), 86.35% (-1.5%) and 87.65% (-0.2%) accuracy, respectively. On the contrary, performing a fusion of skeleton and acceleration data obtained by the smartwatch or with the fusion of both acceleration sensors shows an improved accuracy of 89.32% (1.47%) and 89.30% (1.45%).

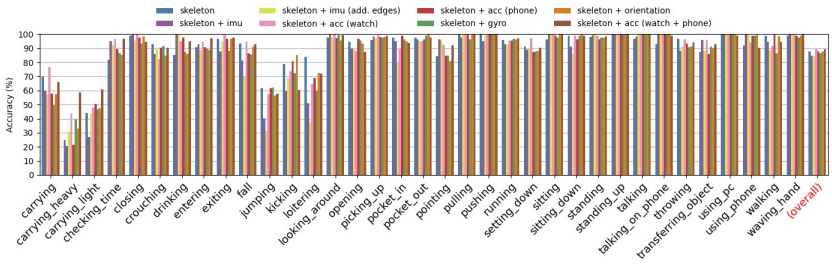


Fig. 6: Class specific accuracy for all MMACT classes for the fusion of various data modalities with Fusion-GCN.

Class	Skl	Skl + Acc
carrying_heavy	24.69	43.83
checking_time	81.93	96.58
drinking	85.00	95.00
transferring_object	87.23	96.10
pointing	84.52	92.34

Table 2: Top-5 most improved classes by the fusion of skeleton (Skl) and additional accelerometer (Acc) data from the smartwatch.

Table 2 shows the top-5 improved classes by the fusion with the accelerometer measurements of a smartwatch. All the top-5 improved actions have a high arm movement in common. In Figure 5 we give a confusion matrix for the Skeleton + Accelerometer (Watch) and highlight the most confused classes. Especially the variations of the ”carrying” actions are hard to distinguish by their obvious similarity. Also, actions that contain sudden movements with high acceleration peaks are often confused (”jumping” is often considered as ”falling”). In general, most of the activities can be recognized quite well. Figure 6 gives a general comparison of all class-specific results on different fusion experiments. Especially the fusion from skeleton-sequences with the accelerometer measurements (skeleton + acc (watch)) suggest a high improvement on many classes, especially the similar ”carrying” classes.

Fusion of Skeleton, RGB and IMU One experiment is conducted using skeleton, RGB and IMU with IMU nodes appended to the skeleton central node without additional edges in combination with and all three RGB early fusion approaches. The results in Table 1a show that, like previously except for the RGB patch feature model, all models achieve an accuracy over 90%, albeit not reaching the same values as the skeleton and IMU fusion approach.

4.5 Limitations and Discussion

Comparing skeleton and skeleton + IMU, the fused approach generally has less misclassifications in all areas. Especially similar actions, such as ”throw”, ”catch”, ”knock” or ”tennis swing”, are able to be classified more confidently. The only action with decreased recognition accuracy using the fused approach is ”jog” which is misclassified more often as ”walk”, two similar actions and some of the few with sparse involvement of arm movement. Common problems for all RGB approaches regarding UTD-MHAD are a small number of training samples, resulting in overfitting in some cases that can not be lifted by either weight decay or dropout. Another fact is the absence of object interactions in UTD-MHAD. With the exception of ”sit2stand” and ”stand2sit”, actions

such as ”throwing”, ”catching”, ”pickup_throw” or sports activities never include any objects. As pointed out previously, skeleton is focused purely on human movements and, by that, omits all other objects inside of a scene. RGB still contains such visual information, making it supposedly more efficient in recognizing object interactions. In contrast, many of MMACT's actions, like ”transferring_object”, ”using_pc”, ”using_phone” or ”carrying”, make use of real objects. While fusion with RGB modality achieves similar accuracies as other approaches, incorporating the data into the network increases the training time by up to a magnitude of ten; hence, the RGB fusion models do not provide a viable alternative to skeleton and IMU regarding the current preprocessing and training configurations. Therefore, due to timely constraints, experiments for fusion of skeleton and RGB modalities on the larger dataset MMACT are omitted.

5 Conclusion

With Fusion-GCN, we presented an approach for multimodal action recognition using GCNs. To incorporate additional modalities we suggest two different fusion dimensions, either on a channel- or spatial dimension. Further integration into early- and late fusion approaches have been presented. In our experiments we considered the flexible fusion of skeleton sequences, with inertial measurements, accelerometer-, gyro-, orientation- measurements separately, as well as RGB features. Our presented fusion approach successfully improved the previous baselines on the large-scale MMACT dataset by a significant margin. Further, it was showcased that additional modalities can further improve recognition from skeleton-sequences. However, the addition of too many modalities decreased the performance. We believe that Fusion-GCN demonstrated successfully that GCNs serve as good basis for multimodal action recognition and could potentially guide future research in this domain.

References

1. Cai, J., Jiang, N., Han, X., Jia, K., Lu, J.: JOLO-GCN: mining joint-centered light-weight information for skeleton-based action recognition. In: IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021. pp. 2734–2743. IEEE (2021). https://doi.org/10.1109/WACV48630.2021.00278, https://doi.org/10.1109/WACV48630.2021.00278
2. Cao, C., Zhang, Y., Zhang, C., Lu, H.: Action recognition with joints-pooled 3d deep convolutional descriptors. In: Kambhampati, S. (ed.) Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. pp. 3324–3330. IJCAI/AAAI Press (2016), http://www.ijcai.org/Abstract/16/470
3. Cao, C., Zhang, Y., Zhang, C., Lu, H.: Body joint guided 3-d deep convolutional descriptors for action recognition. IEEE Trans. Cybern. **48**(3), 1095–1108 (2018). https://doi.org/10.1109/TCYB.2017.2756840, https://doi.org/10.1109/TCYB.2017.2756840
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2021). https://doi.org/10.1109/TPAMI.2019.2929257, https://doi.org/10.1109/TPAMI.2019.2929257

5. Chen, C., Jafari, R., Kehtarnavaz, N.: 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 2015, Quebec City, QC, Canada, September 27-30, 2015. pp. 168–172. IEEE (2015). https://doi.org/10.1109/ICIP.2015.7350781, https://doi.org/10.1109/ICIP.2015.7350781
6. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 180–189. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00026, https://doi.org/10.1109/CVPR42600.2020.00026
7. Chéron, G., Laptev, I., Schmid, C.: P-CNN: pose-based CNN features for action recognition. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 3218–3226. IEEE Computer Society (2015). https://doi.org/10.1109/ICCV.2015.368, https://doi.org/10.1109/ICCV.2015.368
8. Duong, T.V., Bui, H.H., Phung, D.Q., Venkatesh, S.: Activity recognition and abnormality detection with the switching hidden semi-markov model. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA. pp. 838–845. IEEE Computer Society (2005). https://doi.org/10.1109/CVPR.2005.61, https://doi.org/10.1109/CVPR.2005.61
9. Gao, Y., Xiang, X., Xiong, N., Huang, B., Lee, H.J., Alrifai, R., Jiang, X., Fang, Z.: Human action monitoring for healthcare based on deep learning. IEEE Access **6**, 52277–52285 (2018). https://doi.org/10.1109/ACCESS.2018.2869790, https://doi.org/10.1109/ACCESS.2018.2869790
10. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal neurons in artificial neural networks. Distill **6**(3), e30 (2021)
11. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. CoRR **abs/1503.02531** (2015), http://arxiv.org/abs/1503.02531
12. Hu, W., Xie, D., Fu, Z., Zeng, W., Maybank, S.J.: Semantic-based surveillance video retrieval. IEEE Trans. Image Process. **16**(4), 1168–1181 (2007). https://doi.org/10.1109/TIP.2006.891352, https://doi.org/10.1109/TIP.2006.891352
13. Huynh, H.H., Meunier, J., J.Sequeira, M.Daniel: Real time detection, tracking and recognition of medication intake. International Journal of Computer and Information Engineering **3**(12), 2801 – 2808 (2009), https://publications.waset.org/vol/36
14. Iqbal, U., Doering, A., Yasin, H., Krüger, B., Weber, A., Gall, J.: A dual-source approach for 3d human pose estimation from single images. Comput. Vis. Image Underst. **172**, 37–49 (2018). https://doi.org/10.1016/j.cviu.2018.03.007, https://doi.org/10.1016/j.cviu.2018.03.007
15. Islam, M.M., Iqbal, T.: HAMLET: A hierarchical multimodal attention-based human activity recognition algorithm. CoRR **abs/2008.01148** (2020), https://arxiv.org/abs/2008.01148
16. Islam, M.M., Iqbal, T.: Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition. IEEE Robotics and Automation Letters (2021)
17. Joze, H.R.V., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: multimodal transfer module for CNN fusion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 13286–13296. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01330, https://doi.org/10.1109/CVPR42600.2020.01330
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980

19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=SJU4ayYgl
20. Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., Murakami, T.: Mmact: A large-scale dataset for cross modal human action understanding. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 8657–8666. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00875, https://doi.org/10.1109/ICCV.2019.00875
21. Kotyan, S., Kumar, N., Sahu, P.K., Udutalapally, V.: HAUAR: home automation using action recognition. CoRR **abs/1904.10354** (2019), http://arxiv.org/abs/1904.10354
22. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 11977–11986. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01225, http://openaccess.thecvf.com/content_CVPR_2019/html/Kreiss_PifPaf_Composite_Fields_for_Human_Pose_Estimation_CVPR_2019_paper.html
23. Li, B., Li, X., Zhang, Z., Wu, F.: Spatio-temporal graph routing for skeleton-based action recognition. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. pp. 8561–8568. AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33018561, https://doi.org/10.1609/aaai.v33i01.33018561
24. Li, J., Xie, X., Pan, Q., Cao, Y., Zhao, Z., Shi, G.: Sgm-net: Skeleton-guided multimodal network for action recognition. Pattern Recognit. **104**, 107356 (2020). https://doi.org/10.1016/j.patcog.2020.107356, https://doi.org/10.1016/j.patcog.2020.107356
25. Liu, M., Yuan, J.: Recognizing human actions as the evolution of pose estimation maps. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1159–1168. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00127, http://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Recognizing_Human_Actions_CVPR_2018_paper.html
26. Liu, T., Kong, J., Jiang, M.: RGB-D Action Recognition Using Multimodal Correlative Representation Learning Model. IEEE Sensors Journal **19**(5), 1862–1872 (Mar 2019). https://doi.org/10.1109/JSEN.2018.2884443
27. Liu, Y., Wang, K., Li, G., Lin, L.: Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. IEEE Trans. Image Process. **30**, 5573–5588 (2021). https://doi.org/10.1109/TIP.2021.3086590, https://doi.org/10.1109/TIP.2021.3086590
28. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=Skq89Scxx
29. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C., Yong, M.G., Lee, J., Chang, W., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines. CoRR **abs/1906.08172** (2019), http://arxiv.org/abs/1906.08172
30. Luo, Z., Hsieh, J., Jiang, L., Niebles, J.C., Fei-Fei, L.: Graph distillation for action detection with privileged modalities. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV. Lecture Notes in Computer Science, vol.

11218, pp. 174–192. Springer (2018). https://doi.org/10.1007/978-3-030-01264-9_11, https://doi.org/10.1007/978-3-030-01264-9_11
31. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: real-time multi-person 3d motion capture with a single RGB camera. ACM Trans. Graph. **39**(4), 82 (2020). https://doi.org/10.1145/3386569.3392410, https://doi.org/10.1145/3386569.3392410
32. Memmesheimer, R., Theisen, N., Paulus, D.: Gimme signals: Discriminative signal encoding for multimodal activity recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021. pp. 10394–10401. IEEE (2020). https://doi.org/10.1109/IROS45743.2020.9341699, https://doi.org/10.1109/IROS45743.2020.9341699
33. Ni, B., Yan, S., Kassim, A.A.: Recognizing human group activities with localized causalities. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 1470–1477. IEEE Computer Society (2009). https://doi.org/10.1109/CVPR.2009.5206853, https://doi.org/10.1109/CVPR.2009.5206853
34. Niu, W., Long, J., Han, D., Wang, Y.F.: Human activity detection and recognition for video surveillance. In: Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, 27-30 June 2004, Taipei, Taiwan. pp. 719–722. IEEE Computer Society (2004)
35. Norcliffe-Brown, W., Vafeias, S., Parisot, S.: Learning conditioned graph structures for interpretable visual question answering. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 8344–8353 (2018), https://proceedings.neurips.cc/paper/2018/hash/4aeae10ea1c6433c926cdfa558d31134-Abstract.html
36. Noury, N., Fleury, A., Rumeau, P., Bourke, A.K., Laighin, G.O., Rialle, V., Lundy, J.E.: Fall detection-principles and methods. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 1663–1666. IEEE (2007)
37. Papadopoulos, K., Ghorbel, E., Aouada, D., Ottersten, B.E.: Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition. CoRR **abs/1912.09745** (2019), http://arxiv.org/abs/1912.09745
38. Peng, W., Hong, X., Chen, H., Zhao, G.: Learning graph convolutional network for skeleton-based human action recognition by neural searching. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 2669–2676. AAAI Press (2020), https://aaai.org/ojs/index.php/AAAI/article/view/5652
39. Perez-Rua, J., Vielzeuf, V., Pateux, S., Baccouche, M., Jurie, F.: MFAS: multimodal fusion architecture search. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6966–6975. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00713, http://openaccess.thecvf.com/content_CVPR_2019/html/Perez-Rua_MFAS_Multimodal_Fusion_Architecture_Search_CVPR_2019_paper.html
40. Ramezani, M., Yaghmaee, F.: A review on human action analysis in videos for retrieval applications. Artif. Intell. Rev. **46**(4), 485–514 (2016). https://doi.org/10.1007/s10462-016-9473-y, https://doi.org/10.1007/s10462-016-9473-y
41. Ryoo, M.S., Fuchs, T.J., Xia, L., Aggarwal, J.K., Matthies, L.H.: Robot-centric activity prediction from first-person videos: What will they do to me'. In: Adams, J.A., Smart, W.D., Mutlu, B., Takayama, L. (eds.) Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2015, Portland, OR, USA, March

2-5, 2015. pp. 295–302. ACM (2015). https://doi.org/10.1145/2696454.2696462, https://doi.org/10.1145/2696454.2696462
42. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A large scale dataset for 3d human activity analysis. CoRR **abs/1604.02808** (2016), http://arxiv.org/abs/1604.02808
43. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 12026–12035. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.01230, http://openaccess.thecvf.com/content_CVPR_2019/html/Shi_Two-Stream_Adaptive_Graph_Convolutional_Networks_for_Skeleton-Based_Action_Recognition_CVPR_2019_paper.html
44. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: Cipolla, R., Battiato, S., Farinella, G.M. (eds.) Machine Learning for Computer Vision, Studies in Computational Intelligence, vol. 411, pp. 119–135. Springer (2013). https://doi.org/10.1007/978-3-642-28661-2_5, https://doi.org/10.1007/978-3-642-28661-2_5
45. Solbach, M.D., Tsotsos, J.K.: Vision-based fallen person detection for the elderly. In: 2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017. pp. 1433–1442. IEEE Computer Society (2017). https://doi.org/10.1109/ICCVW.2017.170, https://doi.org/10.1109/ICCVW.2017.170
46. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: Skeleton-indexed deep multi-modal feature learning for high performance human action recognition. In: 2018 IEEE International Conference on Multimedia and Expo, ICME 2018, San Diego, CA, USA, July 23-27, 2018. pp. 1–6. IEEE Computer Society (2018). https://doi.org/10.1109/ICME.2018.8486486, https://doi.org/10.1109/ICME.2018.8486486
47. Song, Y., Zhang, Z., Shan, C., Wang, L.: Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In: Chen, C.W., Cucchiara, R., Hua, X., Qi, G., Ricci, E., Zhang, Z., Zimmermann, R. (eds.) MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. pp. 1625–1633. ACM (2020). https://doi.org/10.1145/3394171.3413802, https://doi.org/10.1145/3394171.3413802
48. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 6450–6459. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00675, http://openaccess.thecvf.com/content_cvpr_2018/html/Tran_A_Closer_Look_CVPR_2018_paper.html
49. Tripathi, R.K., Jalal, A.S., Agrawal, S.C.: Suspicious human activity recognition: a review. Artif. Intell. Rev. **50**(2), 283–339 (2018). https://doi.org/10.1007/s10462-017-9545-7, https://doi.org/10.1007/s10462-017-9545-7
50. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V. Lecture Notes in Computer Science, vol. 11209, pp. 413–431. Springer (2018). https://doi.org/10.1007/978-3-030-01228-1_25, https://doi.org/10.1007/978-3-030-01228-1_25
51. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA,

February 2-7, 2018. pp. 7444–7452. AAAI Press (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17135
52. Zheng, Y., Bao, H., Xu, C.: A method for improved pedestrian gesture recognition in self-driving cars. Australian Journal of Mechanical Engineering **16**(sup1), 78–85 (2018)