A unified approach for action recognition on various data modalities

Raphael Memmesheimer, Dietrich Paulus
Active vision Group
University of Koblenz
raphael@uni-koblenz.de

Abstract

We present a unified approach for action recognition on various sensor data modalities. Motion is represented in an image, a EfficientNet 2D-CNN is used for training an action recognition model on the common image representation. Our unified approach handles different sensor modalities in a common way. That distinguishes our approach to many previous proposed approaches that propose different methods per modality. Therefore, it remains simple to evaluate which sensor fits a certain action recognition problem. We show that our approaches generalizes well across 5 different data modalities (pose sequences transformed from videos, skeleton sequences, motion capture data, inertial measurements, Wi-Fi CSI fingerprints) and achieves comparable results on 4 public available datasets and the MMAct challenge dataset.

1. Introduction

Action recognition is a well established topic in the computer vision domain. Potential applications are manifold like for surveillance of elder people or the improvement of human-robot-interaction. Research in this field has followed the advances from hand-crafted features to learned ones.

Approaches that generalize well across modalities are widely neglected. Our approach follows the idea to represent motion in an image and use a 2D-CNN to classify the represented motion [20, 10]. Our approach achieves reasonable accuracy but still generalizes well across a variety of modalities. A benefit is that the adaption to a new modality reduces to defining which data should be represented in the image without adoptions to the network architecture and the overall training process. This benefit makes it more flexible than many action recognition approaches that focus solely on action recognition on a single modality. In Table 1 we compare different recent approaches in terms of modality support. A unified action recognition approach for multiple modalities allows fast integration of different sensors to overcome sensor-specific drawbacks like occlusion for depth cameras or missing context for wearable sensors. In contrast to approaches like [5] our approach does not require designing sub-models per modality. Our approach remains also flexible for integration into early- or late-fusion approaches for multi-modal experiments. In this paper we experiment with an early fusion based on a representation level omitting the overhead of individual network streams.

Table 1. Modality support from various approaches.

Name	Skl	IMU	WiFi	MoCap	RGB
Liu et al. [10]	✓	✗	✗	✗	✗
Ehatisham et al. [2]	✗	✓	✗	✗	✓
Imran et al. [5]	✓	✓	✗	✗	✓
Liu et al. [11]	✓	✗	✗	✗	✗
Memmesheimer et al. [13]	✓	✓	✓	✓	(✓)
Ours	✓	✓	✓	✓	(✓)

Motion data in the form of skeleton sequences, motion capture data or inertial measurements are sampled and represented in an image. A 2D-CNN as known from image classification is used to classify the motion-images. Representing data in 2D data structures like an image has for classification has been previously proposed for i.e. speech recognition [22, 4] or skeleton-based action recognition [20, 10]. In this paper we built up on the representation as proposed for multi-modal one-shot action recognition [14] and embed it in a classification setting for action recognition in various data modalities.

A goal of this work is to motivate research not only to improve a single modality recognition accuracy, but also focus on the transferability to other sensor modalities. The overall contribution of this paper are as follows:

- A simple approach for unified action recognition for skeleton, inertial, Wi-Fi and motion capturing data is presented.
- Experiments are conducted on four publicly available datasets for various data modalities.

2. Related Work

An in-depth review for action recognition on various data-modalities is given by Sun et al. [17].

For Wi-Fi based action recognition 1D-CNN [19] and 2D-CNN [13] based approaches have been presented previously. The 52 Channel State Information (CSI) fingerprints are used to determine an action class. Our approach builds around a 2D-CNN with a dense image representation in contrast to [13] using a sparse representation. A good indicator for the progress of skeleton-based action recognition are the results on the NTU RGB+D dataset [8]. Initial approaches are based around Long Term Short Term Memory (LSTM) [9] or Recurrent Neural Networks (RNN). For skeleton-based action recognition, approaches based on Graph Convolutional Networks (GCN) are defining current state-of the art methods. With the spatio-temporal GCN approach by Yan et al. [21] steadily improved action recognition on skeleton-sequences [11, 15]. To the best of our knowledge, there is currently no GCN-based approach that is used for action recognition on various data modalities.

Along with the UTD-MHAD dataset which contains RGB-D, skeleton and inertial data, Chen et al. [1] presented approaches to fuse depth information and inertial measurements to improve action recognition accuracy. They use separate approaches per modality, For depth sequences they use depth motion maps and for gyroscope signals they use partitioned temporal windows. Imran et al. [5] 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. Finally, the features of the submodels are fused and an action class label is predicted. The fused results are promising, and additional modality fusion improved the results. 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 and training approach for various modalities.

Most fusion methods rely on complex individual representations per modality or propose complex multi-stream architectures. We build on our previous work [13], which follows a similar approach, but we improve the sparse representation with a dense representation [14]. Various representation have been proposed for multiple data modalities [10, 20], however the focus has been mainly on single modality action recognition in contrast to a unified approach for multiple sensor modalities.

3. Approach

Our approach represents motion from various data modalities into images. We expect the motion data to be present as a multivariate signal stream. This allows direct application for skeleton-sequences, inertial measurement units, Wi-Fi CSI fingerprints and motion capturing data as we experimented with. Further sensors should like gyroscopes or positioning systems should be straight forward to integrate. Videos might be integrated by extracting featured from sampled frequencies and transform them to images. The transformed images are used to train a classifier. In our case we use a recently proposed EfficientNet-B2 [18] architecture. The overall approach is depicted in Figure 1.

3.1. Problem Formulation

A signal-level problem formulation of the action recognition problem ensures a flexible integration of various sensor modalities. We follow the signal-level formulation of [13]. The problem of action recognition with a given set of k actions $Y = \{0, \ldots, k\}$ 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. The identity of each channel is encoded in the y-axis of the image. Sampled signal states over time are encoded throughout the x-axis of the image.

3.2. Representation

Our approach builds upon a discriminable image representation. We follow a representation as proposed in [14] for one-shot action processing. Multivariate signal or higher-level feature sequences are reassembled into a 3 channel image. Each row of the resulting image corresponds to one joint and each channel corresponds to one sample in the sequence. The color channels, red, green and blue, represent respectively the signals' x-, y- and z-values. The resulting images are normalized to the range of 0 to 1. We chose to normalize over the whole image to preserve the relative magnitude of the signals. In contrast to the representations used for multimodal action classification [13] or skeleton based action recognition [20, 10] the proposed representation is invertible and more compact. Example representations are shown in Figure 2.

3.3. Architecture

In contrast to other action recognition approaches for multiple modalities [5, 1] that employ different sub-architectures per modality, we propose a common architecture for all modalities. As we represent all modalities in an image we use a 2D-CNN, instead of presenting a custom architecture we employ a EfficientNet-B2 [18] architecture which has recently proven to perform well in the image classification task. The EfficientNet model family

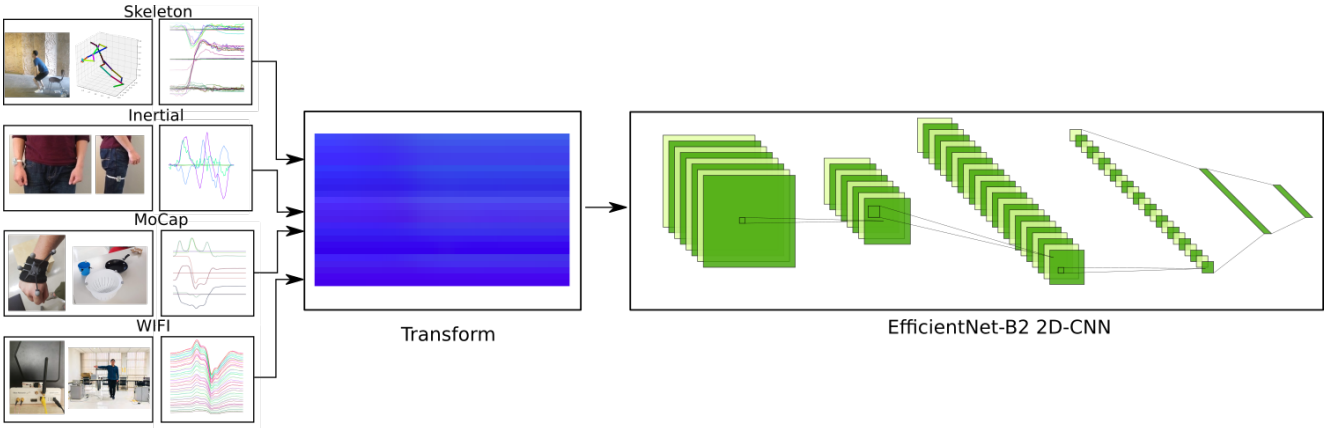


Figure 1. Approach overview. Signal data from various sensor modalities are transformed into a common dense image representation. A EfficientNet-B2 2D-CNN is used for training on basis of the image representation.

is based on architecture search conditioned by maximizing the validation accuracy while minimizing the floating-point operations. This makes it a practical candidate for our approach and potential applications.

4. Experiments

We use four different datasets to verify that our action recognition approach is applicable to four different sensor modalities. We present the datasets along with their results in Table 2 and finally give a discussion.

4.1. Datasets

MMAct Challenge

The dataset provided for the MMAct challenge contains 35 action classes and is divided into cross-scene and cross-view splits. Video, poses gathered by OpenPifPaf [7], accelerometer, gyroscope and orientation sequences were provided during training. For testing, only video sequences were provided. For the MMAct challenge in the cross-scene protocol 12793 training samples, 5024 validation and 17154 unknown test samples were provided. For the cross-view protocol 17653 training samples, 3499 validation and 13869 test samples were provided. As during testing only video sequences were provided, we extracted human pose features using OpenPifPaf [7] with a scale of 0.2 and enforced full-body poses. Further, poses with a overall confidence under 0.2 were rejected. Frames containing no pose estimates were filled with empty poses. Similarly, we processed the training videos and used the extracted pose sequences in addition to the already provided ones. Example representations, as used for the MMAct challenge, are shown in Figure 3. Results for the challenge are given in Table 3.

NTU RGB+D 60 / NTU RGB+D 120

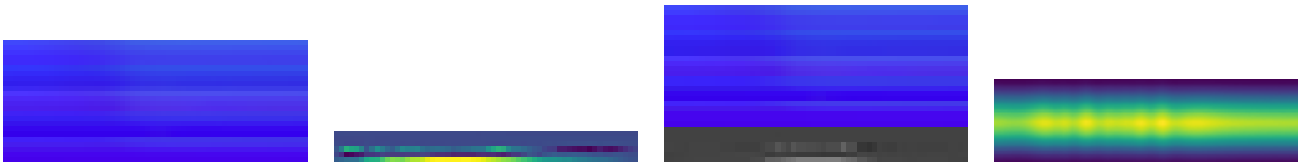
The NTU RGB+D 120 [8] dataset is a large scale action recognition dataset containing RGB+D image streams and skeleton estimates. In contrast to the first NTU RGB+D 60 version of the dataset which contained 56880 sequences with 60 classes, the extended NTU RGB+D 120 dataset consists of 114,480 sequences 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 [8]. For the 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. We report results on both versions with both cross subject and cross view splits.

UTD-MHAD

This dataset [1] contains 27 actions of 8 individuals performing 4 repetitions each. RGB-D camera, skeleton estimates and inertial measurements are included. The RGB-D camera is placed frontal to the demonstrating person. The IMU is either attached at the wrist or the leg during the movements. A cross-subject protocol is followed as proposed by the authors [1]. Half of the subjects are used for training while the other half is used for validation. This dataset is a great candidate because it contains various data modalities and also allows fusion experiments. Because of its different modalities we use it for experiments on skeleton, inertial and fused data.

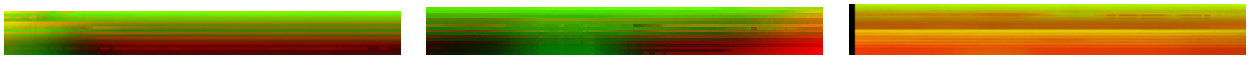
ARIL

This dataset [19] contains Wi-Fi Channel State Information (CSI) fingerprints. The CSI describes how wireless signals



(a) (b) (c) (d)

Figure 2. Example representations for skeleton sequences *(a)*, inertial measurements *(b)*, fused measurements *(c)* and Wi-Fi CSI fingerprints *(d)*. The four example representations show the range of modalities we conducted experiments on.



(a) (b) (c)

Figure 3. Example representations for the pose sequences of the MMAct challenge dataset for the action classes standing up *(a)*, setting down *(b)*, transfering objects *(c)*.

propagate from the transmitter to the receiver. 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).

Simitate

The Simitate [12] 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 an 80/20 train/test split for our experiments.

4.2. Implementation

For better direct comparability, we utilize the same training procedure as in [13]. The approach is implemented in PyTorch [16, 3]. Models are trained using a EfficientNet architecture for 120 epochs on a single Nvidia GeForce RTX 2080 TI with 11 GB GDDR-6. 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 was used. For the results in Table 2, no augmentation methods were applied during the training process. For the results in Table 3 random 10 degree rotations were applied to the training set. No pre-training on larger-scale action recognition datasets was executed in advance.

4.3. Results

Table 2 gives the results for the different modalities and different datasets in relation to other related methods. Our approach achieves good accuracies across the different datasets and different splits. It not necessarily competes with recent GCN-based approaches for skeleton-based action recognition, but competes very well in comparison to CNN-based action recognition methods, while still generalizing well to various other sensor modalities. For the UTD-MHAD we got the highest accuracy on skeleton sequences, and improve by a high margin over the fused accuracy of the similar approach [13]. Individual architectures per modality potentially lead to higher recognition accuracies [5, 2]. However, we claim that our approach simplifies the action recognition training and inference by a common architecture for all modalities and relax the need for individual streams per modality. For motion capturing experiments, we compete comparably well with the augmented results of [13]. Similar to the Wi-Fi experiments, we perform better as the originally proposed approach from [19] and perform comparably well to the augmented results of sparse representation [13]. For the fusion experiments we decided to use an early fusion method like in [13] to avoid multiple network-streams to be trained individually. Fusion is done by concatenating the signal matrices after sub-sampling the higher frequent modality. As in [13] we could not improve the results for fused results over the results of only skeleton data. For the Simitate dataset, we could add object context by fusing the interacting objects to the hand pose measurements. A late fusion method might improve the fusion, however will add complexity to the overall model by introducing individual network streams. Our approach mostly benefits by the simplicity of the approach and the wide variety of supported modalities over the current available action recognition approaches. Our approach can not compete directly with the most recent approaches for skeleton-based action recognition like [11], but generalize across various

Table 2. Action recognition results on four different datasets. Accuracy in [%] is given.

Approach	Type	NTU 60		NTU 120		UTD-MHAD				Simitate	ARIL	#
		CS	CV	CS	CV	RGB	Skl	IMU	Fused	MoCap	Wi-Fi	
Gimme Signals [13]	CNN	-	-	70.8	71.6	-	93.3	81.6	86.5	**96.1**	**94.9**	**4**
Ours	CNN	83.3	81.7	76.7	80.0	-	**93.9**	80.6	93.2	95.0	93.9	**4**
Imran et al. [5]	CNN+RNN	-	-	-	-	83.5	93.5	86.5	97.9	-	-	3
Ehatisham et al. [2]	HOG	-	-	-	-	**85.2**	-	**91.6**	**98.3**	-	-	2
Liu et al. [9]	LSTM	69.2	77.7	55.7	57.9	-	-	-	-	-	-	1
Liu et al. [10]	CNN	80.0	87.2	60.3	63.2	-	-	-	-	-	-	1
Liu et al. [11]	GCN	**91.5**	**96.2**	**86.9**	**88.4**	-	-	-	-	-	-	1
Wang et al. [19]	CNN	-	-	-	-	-	-	-	-	-	89.57	1

Table 3. MMAct Challenge results

User	Entries	Date of Last Entry	mAP	x-scene AP	x-view AP
DeepBlueAI	44	06/08/21	0.9583 (1)	0.9716 (1)	0.9449 (1)
Visual_Analysis_of_Humans	32	06/08/21	0.9288 (2)	0.9468 (2)	0.9108 (2)
Ours	16	06/10/21	0.7406 (3)	0.8064 (3)	0.6748 (3)
MMAct [6]	1	05/07/21	0.4525 (4)	0.4217 (4)	0.4834 (4)

modalities. Further, our approach still achieves a quite high accuracy for both the cross-view and cross-setup accuracy, even outperforming the earlier graph convolutional neural networks [15].

MMAct Challenge

Results for the MMAct Challenge 2021 in the Cross-Modal Trimmed Action Recognition category are given in Table 3. Our approach ended third. The challenge focused on action analysis in video sequences during test time. During training, additional modalities were provided that could be used to guide the training process of a visual model. Our approach utilized only human pose features extracted from the video sequences and therefore remains widely applicable on only video-data. Additional sensor modalities were not integrated as they were not available during test time but could potentially further improve results, especially in occluded settings. Our approach is outperformed by a large margin by the approaches from the DeepBlueAI team and the Visual Analysis of Humans. No details about top scoring approaches were known during the challenge. Our approach however outperforms the MMAct [6] baseline approach by a large margin. The MMAct approach follows an interesting knowledge distillation process which guides a visual model with knowledge distilled from additional sensor modalities during the training. Our approach generalizes better on the cross-scene split then on the cross-view split. This might be the effect of occlusions and to high variation between the training and test samples between splits. Additional modalities, like the measurements from the accelerometer of the smart-watch or smart-phone, could potentially have a positive impact on the action recognition capabilities. A benefit of our approach is the wide applicability on various modalities, which allows simple integration of additional modalities. No additional adoptions need to be performed in case the challenge would have provided additional sensor modalities during test time.

5. Conclusion

We presented an action recognition approach that generalizes well across different sensor data modalities. Motion data is represented in an image, a well established classification CNN is used for the classification of the image representations. We showed that our approach is applicable for skeleton-sequences, inertial measurements, motion capturing data and Wi-Fi CSI fingerprints. Being generalizable across different sensor modalities is a huge practical benefit over other available approaches that often focus on improving results for a single sensor modality.

References

[1] Chen Chen, Roozbeh Jafari, and Nasser 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)*, pages 168–172. IEEE, 2015.

[2] Muhammad Ehatisham-Ul-Haq, Ali Javed, Muhammad Awais Azam, Hafiz MA Malik, Aun Irtaza, Ik Hyun Lee, and Muhammad Tariq Mahmood. Robust human activity recognition using multimodal feature-level fusion. *IEEE Access*, 7:60736–60751, 2019.

[3] W.A. et al. Falcon. Pytorch lightning. `https://github.com/PytorchLightning/pytorch-lightning`, 2019.

[4] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[5] Javed Imran and Balasubramanian Raman. Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):189–208, 2020.

[6] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8667, 2019.

[7] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11977–11986, 2019.

[8] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *CoRR*, abs/1905.04757, 2019.

[9] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016.

[10] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[11] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020.

[12] Raphael Memmesheimer, Ivanna Kramer, Viktor Seib, and Dietrich Paulus. Simitate: A hybrid imitation learning benchmark. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5243–5249. IEEE, 2019.

[13] Raphael Memmesheimer, Nick Theisen, and Dietrich Paulus. Gimme signals: Discriminative signal encoding for multimodal activity recognition. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, USA, 2020. IEEE. accepted for publication.

[14] Raphael Memmesheimer, Nick Theisen, and Dietrich Paulus. Signal level deep metric learning for multimodal one-shot action recognition. *arXiv preprint arXiv:2004.11085*, 2020.

[15] Konstantinos Papadopoulos, Enjie Ghorbel, Djamila Aouada, and Björn Ottersten. Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition. *arXiv preprint arXiv:1912.09745*, 2019.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[17] Zehua Sun, Jun Liu, Qiuhong Ke, and Hossein Rahmani. Human action recognition from various data modalities: A review. *arXiv preprint arXiv:2012.11866*, 2020.

[18] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019.

[19] Fei Wang, Jianwei Feng, Yinliang Zhao, Xiaobin Zhang, Shiyuan Zhang, and Jinsong Han. Joint activity recognition and indoor localization with wifi fingerprints. *IEEE Access*, 7:80058–80068, 2019.

[20] Pichao Wang, Wanqing Li, Chuankun Li, and Yonghong Hou. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158:43–53, 2018.

[21] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[22] Ying Zhang, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, César Laurent, Yoshua Bengio, and Aaron Courville. Towards end-to-end speech recognition with deep convolutional neural networks. *Interspeech 2016*, pages 410–414, 2016.