This WACV 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

A Multi-Task Learning Approach for Human Activity Segmentation and Ergonomics Risk Assessment

Behnoosh Parsa Ashis G. Banerjee
University of Washington
{behnoosh, ashisb}@uw.edu

Abstract

We propose a new approach to Human Activity Evaluation (HAE) in long videos using graph-based multi-task modeling. Previous works in activity evaluation either directly compute a metric using a detected skeleton or use the scene information to regress the activity score. These approaches are insufficient for accurate activity assessment since they only compute an average score over a clip, and do not consider the correlation between the joints and body dynamics. Moreover, they are highly scene-dependent which makes the generalizability of these methods questionable. We propose a novel multi-task framework for HAE that utilizes a Graph Convolutional Network backbone to embed the interconnections between human joints in the features. In this framework, we solve the Human Activity Segmentation (HAS) problem as an auxiliary task to improve activity assessment. The HAS head is powered by an Encoder-Decoder Temporal Convolutional Network to semantically segment long videos into distinct activity classes, whereas, HAE uses a Long-Short-Term-Memory-based architecture. We evaluate our method on the UW-IOM and TUM Kitchen datasets and discuss the success and failure cases in these two datasets.

1. Introduction

With the advancements in computer vision techniques, automated Human Activity Evaluation (HAE) has received significant attention. The aim of this category of problems is to design a computational model that captures the dynamic changes in human movement and measures the quality of human actions based on a predefined metric. HAE has been studied in a variety of computer vision applications such as sports activity scoring, athletes training [35, 48, 30], rehabilitation and healthcare [29, 2], interactive games [53, 25], skill assessment [20, 8], and workers activity assessment in industrial settings [33, 32]. Some of the earlier works on HAE used traditional feature extraction methods for performance analysis [38, 13]. Recently, with the popularity

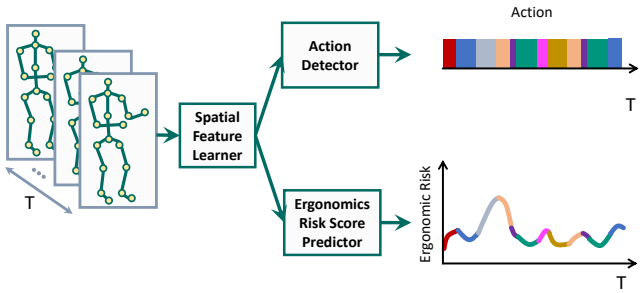

Figure 1. Multi-task activity segmentation and ergonomics risk assessment pipeline.

of deep learning methods, a multitude of creative solutions have emerged for solving HAE problems. Among the proposed methods, some directly learn a mapping from images to a quality score [50]. As the activity quality is highly task-dependent a majority of research is focused on leveraging the available activity information in the learning process [30, 32]. Another approach has been to measure the deviation of a test sequence from a template sequence for determining the activity quality [28]. This approach is valuable when the performance of humans is evaluated based on how well they followed a fixed series of activities in a certain way such as in sport competitions or manufacturing operations.

There is another aspect of HAE that has received less attention despite its importance and potential impact on the safety and health of the society. Human Postural Assessment (HPA) is studied in various fields such as biomechanics, physiotherapy, neuroscience, and more recently in computer vision [33, 32, 22]. HPA is a subcategory of HAE that focuses on determining the quality of human posture using a ergonomics-based (or biomechanics-based) criteria. There are three major challenges in solving HPA problems: (1) the type of task and the object involved in the activity highly influence the risk level. (2) The repetition of certain movements can cause accumulated pressure on specific

body parts. Therefore, it is important to analyze a video in a frame-wise fashion to be able to capture repetition. (3) Everyone does not necessarily perform a task in the same way, hence, a successful algorithm should learn the relation between human joints dynamics and the corresponding ergonomics risk score.

This work is inspired by the importance of HPA problems and their significant impact on the health and safety of industrial workers. However, our approach is not limited to this specific application and it is a novel design that can benefit other aspects of HAE research. We leverage from consistent representation of human 3D pose and propose an end-to-end multi-task framework (Figure 1) that solves Human Activity Segmentation (HAS) as an auxiliary task to improve the HPA performance. Skeleton-based methods have been shown to provide the opportunity of developing more generalizable algorithms for various applications in Human Action Recognition (HAR) and prediction problems [40]. However, they have not been leveraged enough in HAE.

Contributions: This work brings together activity segmentation and activity assessment using a novel multi-task learning framework. Our proposed framework comprises a Graph Convolutional Network (GCN) backbone and an Encoder-Decoder Temporal Convolutional Network (ED-TCN) for the activity segmentation head and a Long-Short-Term-Memory (LSTM)-based head for activity assessment. The contribution of our work is threefold. (1) We introduce a novel combination of GCN with ED-TCN for activity segmentation in long videos that outperforms state-of-the-art results on the UW-IOM dataset. (2) Our Multi-Task Learning (MTL)-emb method initiates a line of research for more informed activity assessment by fusing activity embedding with spatial features for Ergonomics Risk Assessment (ERA). (3) We present a way to use the skeletal information for activity assessment in a Multi-Task Learning (MTL) framework that may enable generalization across a variety of environments and leverage anthropometric information.

2. Related Work

HAS is the task of semantically segmenting a video into clips corresponding various activities and localizing their start and end times. HPA considers the task of finding the ergonomics risk score corresponding to the human posture at every frame of a video. To the best of our knowledge, this is the first work that combines the two separately studied problems of HAS and HPA in a multi-task setting. Moreover, the combination of the GCN backbone with a powerful ED-TCN structure for Single-Task Learning-based HAS (STL-AD) is a novel idea that can compete with methods using image-based features (if the actions are not too similar). The HPA branch also offers a new combination of GCN along with a LSTM unit to learn the relation between human joint dynamics and the corresponding ergonomics risk score. In this section, we summarize the works related to HAE, GCN, and ERA methods to provide the background for our proposed solution.

2.1. Human Activity Evaluation

Also known as Action Quality Assessment, HAE focuses on designing models that are able to learn a mapping between human body dynamics and the completion quality of the performed actions based on an accepted metric or a template sequence (refer to [17] for further literature on early methods with handcrafted features). The majority of deep learning approaches to HAE have focused on using 3D Convolutional Neural Networks (C3D) [47] and Pseudo-3D Networks (P3D) [50] to extract spatio-temporal features that are fed into a regression unit. One of the recent works in applications for physical therapy, [21], proposed a framework including performance metrics, scoring functions, and different neural network architectures for mapping joint coordinates to the activity score. Similarly, [31] used C3D to extract spatio-temporal features and conducted performance score regression using a LSTM unit for data from Olympic events. Despite the value of all these works in initiating the use of computer vision techniques for HAE in rehabilitation and sports, the proposed methods are highly dependent on the context of the video frames. Moreover, the learned mapping between the frames and the score does not incorporate the effect of human body kinematics.

Recently, there have been efforts in leveraging human body kinematics in designing deep architectures for evaluating surgical skills [9]. This work uses 75 dimensional kinematic data (3D coordinates plus velocities) of two surgical tools being manipulated by surgeons and classifies the skill level into expert, intermediate, and novice. Joint relation graph has been utilized to assess the performance of athletes in Olympic events [27]. The proposed joint relation graph is a spatial GCN with node features that are outputs of I3D [3] on image patches containing the human joints.

Parmar and Morris's work [27] is the most similar work to our paper. They propose a multi-task framework utilizing spatio-temporal features to solve action recognition, commentary generation, and HAE score estimation for Olympic events. However, the focus of their work is on short video classification, where each clip includes only one activity, namely, diving of one athlete. In contrasts, our focus is on localizing actions in a long video while simultaneously inferring the ergonomics risk of human posture at every frame.

2.2. Graph Convolution Networks

GCNs was developed to process data belonging to non-Euclidean spaces [49]. GCNs are the most intuitive choice

for human body kinematics since the commonly-used independent and identically distributed random variable assumption is not applicable. Spatio-Temporal Graph Convolutional Networks (ST-GCN) introduced a powerful tool for analyzing human motions in videos, and has been utilized in several computer vision applications [52, 14, 19, 44]. However, most of these works focus on solving Human Action Recognition (HAR) problems. Recently, [32] introduced a Spatio-Temporal Pyramid Graph Network (ST-PGN) for early action recognition. They also used the predicted activity labels to enhance ERA that was computed using 3D skeletal reconstruction. In this work, we leverage a GCN backbone to learn the joint embedding and use that to directly predict the ergonomics risks rather than solving it as a separate problem.

2.3. Ergonomics Risk Assessment

The United States alone has annually more than 150,000 workers suffering from back injuries due to repetitive lifting of heavy objects with inappropriate postures. Hence, many studies have recently looked at designing automated ERA methods [37, 46, 6, 41, 39, 22, 33, 32]. The most widely used methods in the industry are Rapid Entire Body Assessment (REBA) [12] and European Assembly Worksheet (EAWS) [42]. REBA provides a risk score between 1-15 by considering all the main body joint angles, magnitude of the applied force, and ease of grasping an object. EAWS is a similar method that focuses on the upper extremity postures in assembly tasks. In practice, the quantification of risk values is mostly based on observations.

Automated ERA research can be broadly divided into two main categories. One line of research focuses on reducing ergonomics risk in a collaborative setting, where a robot has to place the work platform in a configuration that minimizes the ergonomics risk [24, 43]. Others have used body mounted sensors to measure kinematics for real-time ERA [18, 22]. Another line of research focuses on learning ergonomics risk for various individual actions. In [33], the ERA problem is taken as an action localization problem and Temporal Convolutional Network (TCN) is used to segment the videos into tasks with different risk labels. The ergonomics risk is computed offline and the dataset is labeled with high-, medium-, and low-risk labels. In addition, a dataset on common industry-related activities is introduced in [33] that we use to evaluate the performance of our proposed method. In [32] the problem is approached as an action recognition problem on long videos, and the predicted activity class is used to modify the computed ergonomics risk through a parallel algorithm. This work, on the other hand, introduces a multi-task HPA framework that predicts ergonomics risk directly from human pose with the help of HAS as an auxiliary task.

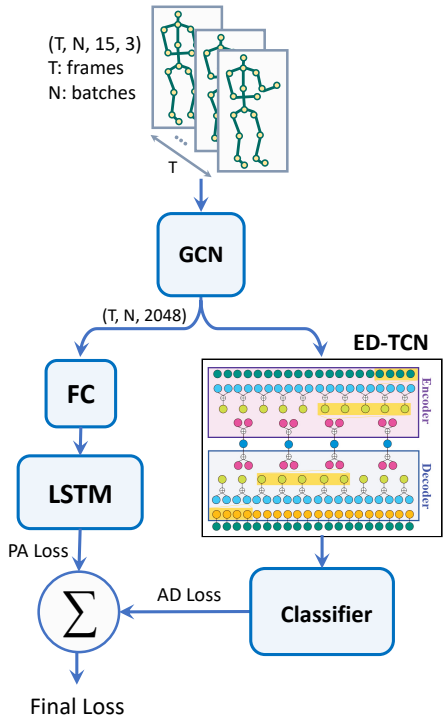

Figure 2. MTL network architecture.

3. Proposed Multi-Task Framework

In ERA, posture alone cannot accurately determine the risk level. The activity class contains information that is key to measure ergonomics risk. We, therefore, define HPA as a MTL problem consisting of an HAS and an HPA task (Figure 2). In the following sections, each component of our MTL model is described in details.

3.1. Spatial Features

The inputs to our multi-task model are 3D joints locations, which is a form of structured data. Since GCNs are known to be powerful in representing structured data [54], our model uses a sequence of stacked GCNs as the backbone for spatial feature extraction, similar to the proposed structure in [52] except for temporal convolution. Just like a 2D convolutional layer, a stacked GCN allows better feature extraction for unstructured data such as graphs.

Given the input $\mathbf{x} \in \mathbf{R}^{D \times N}$, where D is equal to 3 as the joints are represented using (x, y, z) coordinates and N is the number of joints, the adjacency matrix $\mathbf{A} \in \mathbf{R}^{N \times N}$, and the degree matrix $\hat{\mathbf{D}}$ with $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, a Graph Convolution (GC) can be written as,

$$\mathbf{f} = \hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{x}^{\top} \mathbf{W} \,. \tag{1}$$

Here, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{I}$ is the identity matrix. For a graph with human skeletal structure, $\mathbf{A}$ is designed based on the

anatomical connections among the joints. $\mathbf{W} \in \mathbf{R}^{D\times F}$ is the weight matrix that is to be learned. Hence, if the input to a GCN layer is $D \times N$, the output feature $\mathbf{f}$ is $N \times F$, where F is the chosen output feature size. In our proposed backbone, each GCN is followed by a ReLU activation. Moreover, the adjacency matrix is partitioned into three sub-matrices as described in [52] to better capture the spatial relations among the joints. Therefore, Equation (1) is written in a summation form for each GCN layer as:

$$\mathbf{f} = \sum_{a=1}^{3} \hat{\mathbf{D}}_a^{-\frac{1}{2}} \mathbf{A}_a \hat{\mathbf{D}}_a^{-\frac{1}{2}} \mathbf{x}^\top \mathbf{W}_a \ , \tag{2}$$

where a indexes each partition.

3.2. Encoder-Decoder Temporal Convolution for Human Activity Segmentation

In HAS problems, the task is to identify the activities that are happening in untrimmed videos and determine the corresponding initial and final frames [10, 16, 1, 33]. A popular approach that is inspired by works in audio generation and speech recognition [26, 51] is to use feed-forward (i.e., non-recurrent) networks for modeling long sequences. The main component of these methods is a 1D dilated causal convolution that can model long-term dependencies.

A dilated convolution is a filter that applies to an area larger than its length by skipping input values by a certain length [26]. A causal convolution is a 1D convolution which ensures the model does not violate the ordering of the input sequence. The prediction emitted by a causal convolution (that is $p(x_t|x_1, ..., x_{t-1})$) at time step t only depends on the previous data. Combining these two properties, dilated causal convolutions have large receptive fields and are faster than Recurrent Neural Networks (RNNs). Moreover, they are shallower than regular causal convolution due to dilation.

For the HAS task, inspired by [26, 16, 33] we design an ED-TCN-based on 1D dilated convolutions (Figure 2). Our design consists of a hierarchy of four temporal convolutions, pooling, and upsampling layers. The output of the ED-TCN followed by a Fully Connected (FC) layer and a `RelU` activation is fed to the classification layer.

In using ED-TCN for HAS [16, 33], the focus is on learning the temporal sequence and localizing activities. It is common to extract spatial features prior to training from an independent network like VGG16 [45] or ResNet [11]. Our proposed framework learns the spatial and temporal properties of the data in an end-to-end fashion. To our knowledge, this is the first attempt to use ED-TCN in an end-to-end architecture with a spatial feature detector. In addition, the combination of GCN with ED-TCN for solving HAS is a novel approach and it shows promising results.

3.3. Regression Module for Human Postural Assessment

We define HPA as a sub-category of HAE where the activity score is determined based on the safety of the posture. In HPA, the task is to find a mapping between the spatio-temporal features and ergonomics risk score. Our proposed regressor uses the shared spatial features coming from the GCN backbone. The GCN features go through a FC layer with `tanh` nonlinearity and are then fed into a stacked LSTM structure to predict the REBA scores.

3.4. Multi-Task Approach to ERA

MTL is a popular framework for end-to-end training of a single network for solving multiple related tasks. In these networks, a common backbone provides the data representation for branches responsible for learning a specific task. Usually in MTL, there is a main task plus multiple auxiliary tasks that complement the core task. For instance, in HAE, the main task is to determine the action quality. However, action quality is not independent of what action is carried out, which makes the HAS choice of auxiliary tasks natural for this kind of problems.

The supervision signals from the auxiliary tasks can be viewed as inductive biases [4] that limit the hypothesis search space and result in a more generalizable solution. The multi-task approach to HAE has been recently introduced by [30] for determining the quality of action in short clips from Olympic games.

In our work, the main task is to predict the REBA scores. However, the information about human action is closely related to its corresponding ergonomics risk. Therefore, the auxiliary task in this case is the HAS. The long duration videos pose an additional challenge, since, unlike most of the HAE datasets, both the activities and their risk scores vary over time. In a majority of sport HAE [30], a single activity score is predicted for a clip. Here, the HAS task consists of 17 and 20 actions for the UW-IOM and TUM datasets, respectively (see Section 4 for more information on the datasets). Therefore, in any video, activity localization and ERA task involves predicting a smooth function that shows how the risk is changing throughout the video.

We studied two different architectures for solving this MTL problem. In the first architecture, the heads corresponding to each task only share the GCN-driven features. In the second architecture, the output of the *Softmax* layer of the HAS head is fused to the feature going to the LSTM regressor.

We consider a weighted average of the HAS loss and the HPA loss as the overall multi-task HPA loss function,

$$\mathcal{L}_{HPA} = \sum_{t=1}^{T} \alpha(\mathbf{x}_t - \mathbf{y}_t)^2 + \beta|\mathbf{x}_t - \mathbf{y}_t|, \tag{3}$$

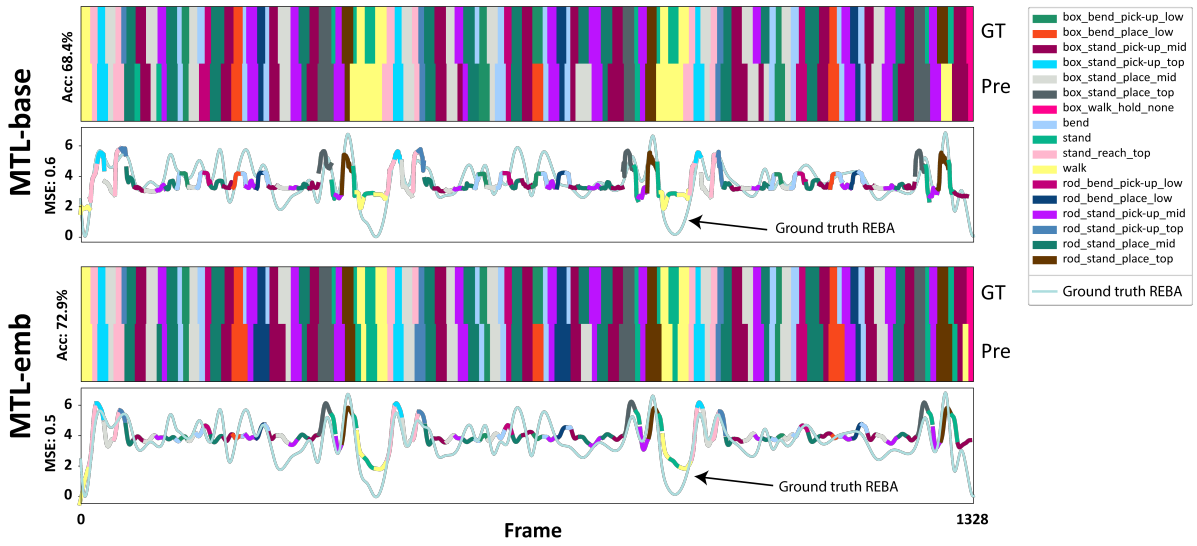

Figure 3. Visualization of HAS and REBA prediction result for a sample test video of UW-IOM dataset. The first and third plots (colored ribbons) are the segmentation results. In each ribbon the top-half is the ground truth and the bottom-half is the predictions by the network. The second and fourth plots depicting the ground truth REBA score and the network prediction. The network prediction is color-coded based on the activity class.

where $\mathbf{y}_t$ is the frame-wise ground truth REBA score and $\mathbf{x}_t$ is the model prediction. $|\cdot|$ is the $\mathcal{L}_1$ norm. α and β are weights to be learned. For HAS, we use cross-entropy loss between ground truth and model prediction,

$$\mathcal{L}_{HAS} = -\sum_{t=1}^{T}\sum_{c=1}^{Cl} \mathbf{y}_{t,c}\log(\mathbf{x}_{t,c}), \tag{4}$$

where Cl is the number of classes. The overall loss is the sum of all the losses,

$$\mathcal{L}_{MTL} = \mathcal{L}_{HPA} + \gamma\mathcal{L}_{HAS}, \tag{5}$$

where γ is to be learned.

4. Experiments

4.1. Datasets

Despite the impact of automated ERA on industry, research in this area has started gaining popularity only recently. As a result, only a few datasets are available that capture representative activities in industrial settings. In particular, two such datasets have been used in recent publications in this domain.

UW-IOM Dataset is a publicly available dataset of 20 videos by [33] that captures industry-relevant activities. This dataset has 17 action classes and labels are of four-tier hierarchy indicating the object, human motion, type of object manipulation (if applicable), and the relative height of the surface on which the activity is taking place. The longest video in this dataset has $2,384$ frames. We use the 3D poses for UW-IOM dataset from our earlier work [32].

TUM Kitchen Dataset has 19 videos consisting of daily activities in a kitchen. Learning with graph-based methods has been shown to be challenging on this datasets due to the similarity of human postures in multiple action classes [32]. We took labels provided by [33] so that we can compare our results with theirs. We used [36] to extract the 3D poses from the videos recorded by the second camera. The longest video in this dataset has $2,048$ frames.

The input features to our model are 3-dimensional keypoints (x, y, z) of $N = 15$ joints, concatenated over time T. Hence, the resulting input tensor is of dimension $3 \times 15 \times T$. The output ground truth labels are frame-wise labels that have the dimension of $1 \times T$.

4.2. Ergonomics Risk Pre-processing

REBA method [12] computes a score describing the total body risk based on the joint angles and the properties of an action. The REBA scores are discrete integers from 1 (the minimum risk level) to 15 (the maximum risk level). In [33], the scores of all the subjects are averaged over the classes and a single score is reported for each activity class. We used the detected skeletons to compute the joint angles and obtained a frame-wise REBA score. However, the REBA profile then becomes a sequence of piece-wise constants, which is hard to learn by a regressor. Therefore, we smoothed the REBA sequence using the Python `UnivariateSpline` function to make it easier for the ERA regressor to learn the patterns. To help advance research in this area, the smoothed REBA scores along with the code are available on the project repository[1].

[1]https://github.com/BehnooshParsa/MTL-ERA

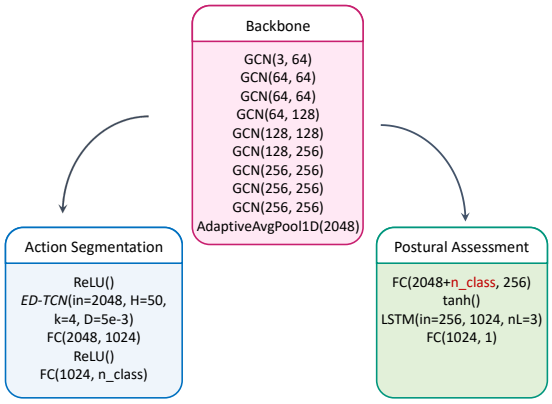

Figure 4. Detailed MTL-emb architecture. $GCN(in, out)$ is a GCN with edge-importance. ED-TCN has 4 hidden layers of size H with kernel size k and dropout of D. $FC(in, out)$ is a fully connected layer. n_{class} is the number of classes. The LSTM has nl layers.

4.3. Implementation Details

All the networks were implemented in PyTorch [34]. The initial values of the loss function weight parameters α, β, and γ were set to 1. All the networks were trained using the Adam optimizer [15]. We implemented early-stopping and trained the model with different learning rates to find the best one (the best performing learning rate is shown in Table 3). The 20 videos in the UW-IOM dataset were randomly split into 15 and 5 for the training and validation set, respectively. For the TUM dataset, the training and validation sets include 15 and 4 videos, respectively.

GCN Backbone: The details of the GCN network is displayed in Figure 4. The output of the final GCN layer is of size $(N, T, 256, 15)$ that is flattened to $(N, T, 3840)$ and passed through an adaptive pool layer. Therefore, the feature that is fed to the rest of the network is of size $(N, T, 2048)$.

Action Segmentation Head: ED-TCN requires input batches to have the same temporal length. Hence, we defined a maximum length in both the training and validation sets, and masked the rest of the inputs with a value of -1 (thus, T corresponds to the maximum sequence length). The predicted sequence was unmasked before calculating the loss. The ED-TCN output goes through two fully connected layers with `RelU` activation and is used to compute the cross-entropy loss.

Postural Assessment Head: We evaluated the performance of two architectures for HPA. In one design, we fuse the Softmax output of the HAS head to the GCN features and call this model, *multi-task-emb*. The base design does not include fusion and we refer to that as *multi-task-base*. The spatial features (from the GCN backbone) are followed by a fully connected layer with `tanh` activation and sent to three layers of LSTM. The LSTM output is followed by a fully connected layer to predict the REBA scores and is sent to the regression loss function.

4.4. Evaluation Metrics

To measure the performance of the HAS network we use *F1-overlap score*, *segmental edit score*, and *Mean Average Precision* (MAP). F1-overlap score is essentially the harmonic mean of *Precision* and *Recall* and is computed using the following well known formula:

$$F_1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \quad (6)$$

Edit score measures the closeness of the predicted sequence to the ground truth sequence. This metric penalizes if the order of the sequence and the number of action segments are not correct. The average precision is computed over all the classes and its mean is reported.

5. Results and Discussion

To evaluate the strength of our proposed multi-task approach in solving the HAS and HPA problems, we carry out two single-task experiments for the HPA task (STL-PA) and the HAS task (STL-AS). Another reason behind the STL-AS experiment is to investigate the power of our GCN model as a spatial feature extractor in solving HAS problems.

The STL-PA network has identical GCN backbone and LSTM design as the MTL network. The average MSE result is reported for the validation set in Table 1. It is clear from the results that the network cannot learn the sophisticated pattern of the REBA profile.

UW-IOM		TUM	
MSE	Sp. Corr. (%)	MSE	Sp. Corr. (%)
1.68 ±0.28	11.79 ±12.32	2.75 ±0.40	62.92±4.89

Table 1. Average MSE and Spearman's Coefficient of the activity score prediction over the validation videos using the STL-PA model.

5.1. Action Segmentation with GCN-ED-TCN

As discussed in Section 2, ED-TCN along with the input features derived from pre-trained networks, have been widely used for HAS. The idea is that given the input spatial features for every time-step of a sequence, this method can segment it into semantically similar pieces. Nonetheless, an end-to-end approach for learning both the spatial and temporal features in an HAS has not been explored with ED-TCN. While GCN models have been used both for activity classification [52, 14] and early action recognition [32], its capability has not been evaluated for HAS.

ED-TCN is used for HAS on the UW-IOM dataset in [33], where the authors compare three spatial feature extractors, namely, a pre-trained VGG16 on ImageNet [7], a fine-tuned version of VGG16 model, and a P-CNN model [5]. Our proposed GCN backbone extracts spatial features based on human pose only, but its performance is comparable with the state-of-the-art as shown in Table 2. Hence, we believe that pose-based features are more suitable for designing a generalizable algorithm. However, we should emphasize that generalizability comes with a price of the model not performing well when the pose information is poor or when the activities require similar postures, which is the case for the TUM dataset (Table 2).

5.2. Single-task vs. Multi-task Approach

The substantial improvement in predicting the activity risk scores is evident when comparing the results in Table 1 with Table 3. We believe that the underlying reason behind this observation is that the REBA score is highly dependent on the type of activity, and learning an auxiliary HAS task can enhance the performance of the HPA head. However, the inverse dependency is not that strong. Our findings indicate that the STL-AS performs better than the MTL approach for HAS (comparing results in Table 2 and 3).

Method	UW-IOM					
	MSE	Sp. Corr. (%)	mAP (%)	Edit score (%)	F1 overlap (%)	Learned Weights
MTL-base	0.72 ±0.14	66.68 ±4.89	76.0 ±8.51	88.36 ± 4.67	89.56 ±4.45	CrE: 0.70, MSE: 0.81, L1: 0.51 lr: 0.001
MTL-emb	0.61 ±0.36	55.18 ±6.57	74.45 ±10.36	91.59 ±1.23	92.03 ±2.54	CrE: 0.72, MSE: 0.85, L1: 0.64 lr: 0.001
	TUM					
MTL-base	1.03 ±0.48	80.44 ±4.67	36.83 ±16.63	64.75 ±13.10	54.15 ±19.01	CrE: 0.95, MSE: 0.97, L1: 0.95 lr: 1e-04
MTL-emb	1.01 ±0.38	73.83 ±8.00	39.23 ±17.00	65.87± 9.13	58.24 ±11.23	CrE: 0.86, MSE: 0.89, L1: 0. 86 lr: 0.0005

Table 3. Results for the MTL network. mAP, edit, and F1-overlap scores are represented using mean and standard deviation values over the validation splits in the UW-IOM and TUM datasets for different activity segmentation methods and modalities. MSE and Spearman's coefficient show the model's performance in predicting the activity risk scores.

5.3. Fusion vs. No Fusion Approach

The main purpose of this experiment is to validate the idea that action information can improve REBA score predictions. Table 3 and Figure 3 show the MTL-base and MTL-emb results, where improvements are observed when the HPA head has access to the `Softmax` output of the HAS head. In Figure 3, we see the highly nonlinear ground truth REBA scoreline (in solid light blue-green) and the corresponding predictions for each action by both the MTL networks. The figure suggests that the network with embedding predicts the REBA scores more accurately. On the contrary, the shared embedding model does not significantly improve the performance of the HAS head. Figure 5 depicts the difference in the confusion matrices of the two models. For simplicity, the off-diagonal elements are ignored. While there are small improvements in a few classes, the overall improvement is not substantial.

5.4. Failure Cases

Although we show that our MTL-emb and STL-AS methods perform well on the UW-IOM dataset and even better than using context heavy features such as VGG16, these models are not particularly successful on the TUM dataset. We present the confusion matrices for the UW-IOM and TUM datasets in Figure 6. In the following, we describe our insights on the performance of the models in detail.

The camera view in the TUM dataset is from the top. As a result, arm pose estimation quality is poor for the activities where the person's back is facing the camera and the arm is occluded such as for *pickup-drawer* and *close-drawer*. Another source of confusion is between *Pickup-hold-**both**-hands* and *Pickup-hold-**one**-hands* due to the fact that the poses are very similar.

Since the segmentation head is not very successful on the TUM dataset, the improvement in the REBA score prediction between the MTL-emb and MTL-base models is also not significant unlike in the case of the UW-IOM dataset. For the TUM dataset, fusing image-based features with the GCN can be potentially useful in decreasing the ambiguity in the GCN spatial descriptors, thereby, improving both the STL-AS and MTL results for REBA score prediction.

6. Conclusions and Future Work

We introduce a graph-based multi-task learning approach for Human Postural Assessment and show that it outperforms the equivalent Single-Task Learning due to the importance of the activity type in the risk associated with a posture. Human Postural Assessment tasks, specifically Ergonomics Risk Assessment, are more challenging than regular Human Activity Evaluation problems since the assessment has to happen in a frame-wise manner and is highly dependent on joint kinematics. Despite the challenge of tracking the intricacies of our risk assessment (REBA) profile, the proposed method shows competence in predicting the risk scores. More importantly, our work demonstrates the effectiveness of the GCN model as a spatial feature extraction backbone, compared to context-based features that have been traditionally used with ED-TCN for Human Activity Segmentation tasks. To showcase the weaknesses of this framework, we implemented our method on a challenging dataset (TUM) and discussed the failure cases.

Although the focus of this work is on Ergonomics Risk Assessment, we believe that our Multi-Task Learning ap-

Method	UW-IOM			TUM		
	mAP (%)	Edit score (%)	F1 overlap (%)	mAP (%)	Edit score (%)	F1 overlap (%)
ED-TCN / Pre-trained VGG16 [33]	-	88.52 ± 1.17	93.24 ± 0.58	-	86.34 ± 3.15	87.92 ± 2.16
ED-TCN / Fine-tuned VGG16 [33]	-	82.96 ± 3.33	87.77 ± 2.51	-	84.96 ± 4.37	87.29 ± 2.78
ED-TCN / Simplified P-CNN [33]	-	89.90 ± 1.16	93.99 ± 0.77	-		
GCN-ED-TCN (STL-AS)	49.61 ± 0.17	92.08 ± 1.18	92.33 ± 0.78	24.17 ± 11.99	67.53 ± 5.16	52.20 ± 22.02

Table 2. mAP, edit, and F1-overlap score represented using mean and standard deviation values over the test videos in the UW-IOM and TUM datasets for different methods and modalities solving the HAS task.

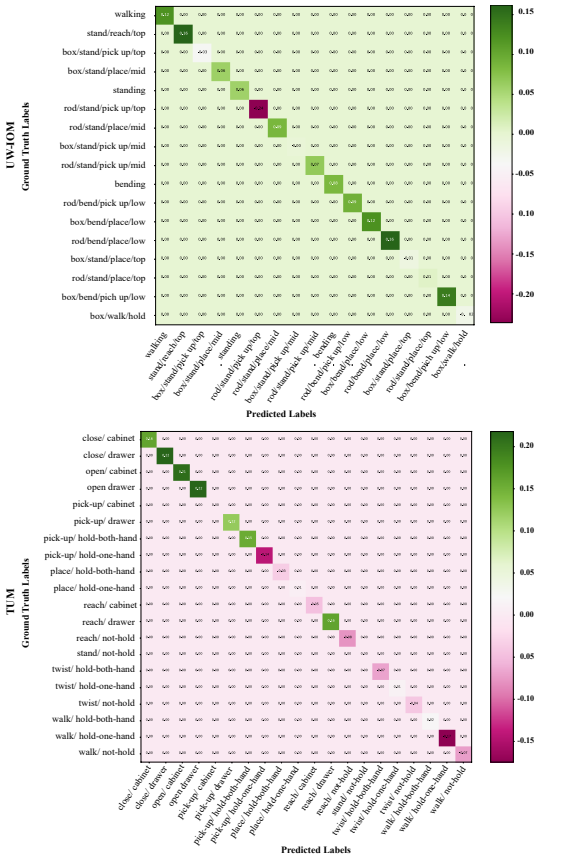

Figure 5. The difference in confusion matrices. The top and bottom matrices are for the UW-IOM and TUM dataset, respectively. The diagonal elements show the differences between the diagonal values of the MTL-emb and MTL-base confusion matrices and the off-diagonal elements are shown as "0.0" for simplicity.

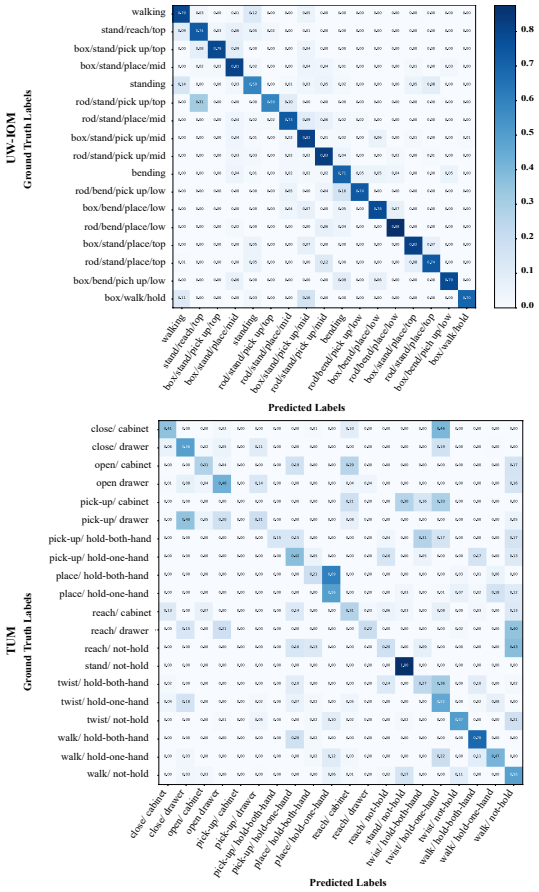

Figure 6. Confusion matrices using MTL-base. The top and bottom matrices are for the UW-IOM and TUM dataset, respectively.

proach can be applied to many other action and skill assessment problems. The mapping of skeletal representation to the activity score using GCN is a new approach for solving ERA, which can initiate a new direction by exploiting the natural connection between posture and activity risk/quality.

Although we outperform state-of-the-art on HAS and ERA on the UW-IOM dataset, some open issues remain. First, generalization concerning other activities has not been addressed. Our method learns the ergonomics risk scores in a supervised learning framework, which makes the performance of the model limited to the labeled activities that have been observed. Second, only joint positions are considered in the spatial representation, while other kinematic information such as velocity and acceleration have been shown to be important for many types of injury such as back injuries [23]. In the future, we hope to address these issues by developing a biomechanics-based human pose representation model that learns the causal relation between joint kinematics and the resultant ergonomics risk.

References

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[2] Renato Baptista, Michel Goncalves Almeida Antunes, Djamila Aouada, and Björn Ottersten. Video-based feedback for assisting physical activity. In *12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2017.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[5] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-CNN: Pose-based CNN features for action recognition. In *IEEE Int. Conf. Comput. Vis.*, pages 3218–3226, 2015.

[6] Ana Colim, Paula Carneiro, Nélson Costa, Pedro M Arezes, and Nuno Sousa. Ergonomic assessment and workstation design in a furniture manufacturing industry—a case study. In *Occupational and Environmental Safety and Health*, pages 409–417. Springer, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 248–255, 2009.

[8] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6057–6066, 2018.

[9] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Evaluating surgical skills from kinematic data using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 214–221. Springer, 2018.

[10] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Fine-grained action segmentation using the semi-supervised action gan. *Pattern Recognition*, 98:107039, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Sue Hignett and Lynn McAtamney. Rapid entire body assessment. In *Handbook of Human Factors and Ergonomics Methods*, pages 97–108. CRC Press, 2004.

[13] Winfried Ilg, Johannes Mezger, and Martin Giese. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In *Joint Pattern Recognition Symposium*, pages 523–531. Springer, 2003.

[14] Sunoh Kim, Kimin Yun, Jongyoul Park, and Jin Young Choi. Skeleton-based action recognition of people handling objects. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 61–70. IEEE, 2019.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[17] Qing Lei, Ji-Xiang Du, Hong-Bo Zhang, Shuang Ye, and Duan-Sheng Chen. A survey of vision-based human action evaluation methods. *Sensors*, 19(19):4129, 2019.

[18] Chunxia Li and SangHyun Lee. Computer vision techniques for worker motion analysis to reduce musculoskeletal disorders in construction. In *Comput. Civil Eng.*, pages 380–387. 2011.

[19] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3595–3603, 2019.

[20] Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato. Manipulation-skill assessment from videos with spatial attention network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[21] Yalin Liao, Aleksandar Vakanski, and Min Xian. A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2):468–477, 2020.

[22] Adrien Malaisé, Pauline Maurice, Francis Colas, and Serena Ivaldi. Activity recognition for ergonomics assessment of industrial tasks with automatic feature selection. *IEEE Robotics and Automation Letters*, 4(2):1132–1139, 2019.

[23] William S Marras, Gregory G Knapik, and Sue Ferguson. Loading along the lumbar spine as influence by speed, control, load magnitude, and handle height during pushing. *Clinical biomechanics*, 24(2):155–163, 2009.

[24] Pauline Maurice, Adrien Malaisé, Clélie Amiot, Nicolas Paris, Guy-Junior Richard, Olivier Rochel, and Serena Ivaldi. Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *The International Journal of Robotics Research*, 38(14):1529–1537, 2019.

[25] Meng Meng, Hassen Drira, and Jacques Boonaert. Distances evolution analysis for online and off-line human object interaction recognition. *Image and Vision Computing*, 70:32–45, 2018.

[26] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[27] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6331–6340, 2019.

[28] German I Parisi, Sven Magg, and Stefan Wermter. Human motion assessment in real time using recurrent self-organization. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 71–76. IEEE, 2016.

[29] Paritosh Parmar and Brendan Tran Morris. Measuring the quality of exercises. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2241–2244. IEEE, 2016.

[30] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019.

[31] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2017.

[32] Behnoosh Parsa, Behzad Dariush, et al. Spatio-temporal pyramid graph convolutions for human action recognition and postural assessment. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1080–1090, 2020.

[33] Behnoosh Parsa, Ekta U Samani, Rose Hendrix, Cameron Devine, Shashi M Singh, Santosh Devasia, and Ashis G Banerjee. Toward ergonomic risk prediction via segmentation of indoor object manipulation actions using spatiotemporal convolutional networks. *IEEE Robotics and Automation Letters*, 4(4):3153–3160, 2019.

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[35] Fotini Patrona, Anargyros Chatzitofis, Dimitrios Zarpalas, and Petros Daras. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, 76:612–622, 2018.

[36] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[37] L Peppoloni, A Filippeschi, E Ruffaldi, and CA Avizzano. A novel wearable system for the online assessment of risk for biomechanical load in repetitive efforts. *International Journal of Industrial Ergonomics*, 52:1–11, 2016.

[38] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *European Conference on Computer Vision*, pages 556–571. Springer, 2014.

[39] Gessieli Possebom, Airton dos Santos Alonço, Sabrina Dalla Corte Bellochio, Tiago Gonçalves Lopes, Dauto Pivetta Carpes, Rafael Sobroza Becker, Antonio Robson Moreira, Tiago Rodrigo Francetto, Fernando Pissetti Rossato, and Bruno Christiano Corrêa Ruiz Zart. Comparison of methods for postural assessment in the operation of agricultural machinery. *Journal of Agricultural Science*, 10(9), 2018.

[40] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017.

[41] Akram Sadat Jafari Roodbandi, Forough Ekhlaspour, Maryam Naseri Takaloo, and Samira Farokhipour. Prevalence of musculoskeletal disorders and posture assessment by qec and inter-rater agreement in this method in an automobile assembly factory: Iran-2016. In *Congress of the International Ergonomics Association*, pages 333–339. Springer, 2018.

[42] Karlheinz Schaub, Gabriele Caragnano, Bernd Britzke, and Ralph Bruder. The european assembly worksheet. *Theoretical Issues in Ergonomics Science*, 14(6):616–639, 2013.

[43] Ali Shafti, Ahmad Ataka, Beatriz Urbistondo Lazpita, Ali Shiva, Helge A. Wurdemann, and Kaspar Althoefer. Real-time robot-assisted ergonomics. In *IEEE Int. Conf. Robot. Autom.*, 2019.

[44] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu 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*, pages 1227–1236, 2019.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Ashish Kumar Singh, ML Meena, Himanshu Chaudhary, and GS Dangayach. Ergonomic assessment and prevalence of musculoskeletal disorders among washer-men during carpet washing: guidelines to an effective sustainability in workstation design. *International Journal of Human Factors and Ergonomics*, 5(1):22–43, 2017.

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[48] Kokum Weeratunga, Anuja Dharmaratne, and Khoo Boon How. Application of computer vision and vector space model for tactical movement classification in badminton. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 76–82, 2017.

[49] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596*, 2019.

[50] Xiang Xiang, Ye Tian, Austin Reiter, Gregory D Hager, and Trac D Tran. S3d: Stacking segmental p3d for action quality assessment. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 928–932. IEEE, 2018.

[51] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. The microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5934–5938. IEEE, 2018.

[52] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[53] Weichen Zhang, Zhiguang Liu, Liuyang Zhou, Howard Leung, and Antoni B Chan. Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation. *Image and Vision Computing*, 61:22–39, 2017.

[54] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.