One-shot skeleton-based action recognition on strength and conditioning exercises

Michael Deyzel
Stellenbosch University
Stellenbosch, South Africa
mcdeyzel@gmail.com

Rensu P. Theart
Stellenbosch University
Stellenbosch, South Africa
rptheart@sun.ac.za

Abstract

There is a need in the sports and fitness industry for a practical system that can identify and understand human physical activity to enable intelligent workout feedback and virtual coaching. Such a system should be able to classify an athlete's actions from only limited examples since it is not feasible to collect a large quantity of human data for every action of interest. In this paper, we present SU-EMD, a novel dataset of skeleton motion sequences of seven common strength and conditioning exercises as captured by both a markerless and marker-based motion capture system. We then formulate the one-shot skeleton action recognition problem as a deep metric learning problem. We use the state-of-the-art graph convolutional network (GCN) to project dissimilar actions further away and similar actions closer together in the learned metric space. By training on NTU RGB+D 120, the metric GCN achieves a one-shot performance of 87.4% on all seven never-before-seen actions. In addition, an ablation study reveals the effect of different losses, embedding sizes and augmentations. Our results show that one-shot metric learning method can be used as a means to classify sports actions in a virtual coaching system where users cannot provide many expert examples for the enrolment of new actions.

1. Introduction

In many sports, strength and conditioning exercise is an imperative facet of competition and training [32]. Besides for sports that directly involve strength training like Olympic weightlifting, bodybuilding and CrossFit, many competitive team sports such as football [26], hockey [4] and rugby [10] require athletes to undergo plyometric and weightlifting exercises prescribed by coaches for functional fitness or injury rehabilitation [6].

Foremost, our research is predicated on the larger problem of developing a virtual fitness instructor that provides athletes and gym-goers with automatic exercise monitoring and performance feedback. By using observations from multiple cameras, such a system could automatically recognise an individual's performance of a specific strength and conditioning exercise, provide feedback and recommendations, and notify users of dangerous movements that might cause injury. Since 3D skeleton motion (the motion sequence of keypoints of interest of the human body) is a compact representation of a physical action performed by an individual and contains the motion data of important joints of the human body, it will be particularly suitable as a modality for an AI-enabled coaching system. This is especially useful in scenarios where no trainers or fitness instructors are available to correct form and advise on routines, making expert guidance more accessible. Large datasets comprised of skeleton motion sequences are available, such as *NTU RGB+D* [25] and *Kinetics 400* [11]. However, no skeleton motion dataset of strength and conditioning actions is available in the literature. To solve this shortcoming, we developed a markerless 3D pose estimation system and collected 840 skeleton motion sequences of seven exercise actions as performed by four subjects. The pose data for the dataset is specially designed to be compatible with a large-scale dataset like *NTU RGB+D*. This dataset is the first of its kind and we have released it to researchers. [1]

Using this data, we consider the utility of a skeleton-based action recognition (SBAR) system in the context of low training resources by tackling the one-shot classification problem, where a model is given only one prior example of an action. Such a system should have prior knowledge of human actions. It should be able to extract spatial-temporal features and recognize these features to classify newly enrolled actions. Along with this observation, we are also inspired by the success of ImageNet as a source for pre-training for convolutional neural networks (CNNs) on image data to test if this success can be paralleled for large-scale skeleton-based action datasets and GCNs. We trained

[1] https://github.com/michaeldeyzel/SU-EMD/

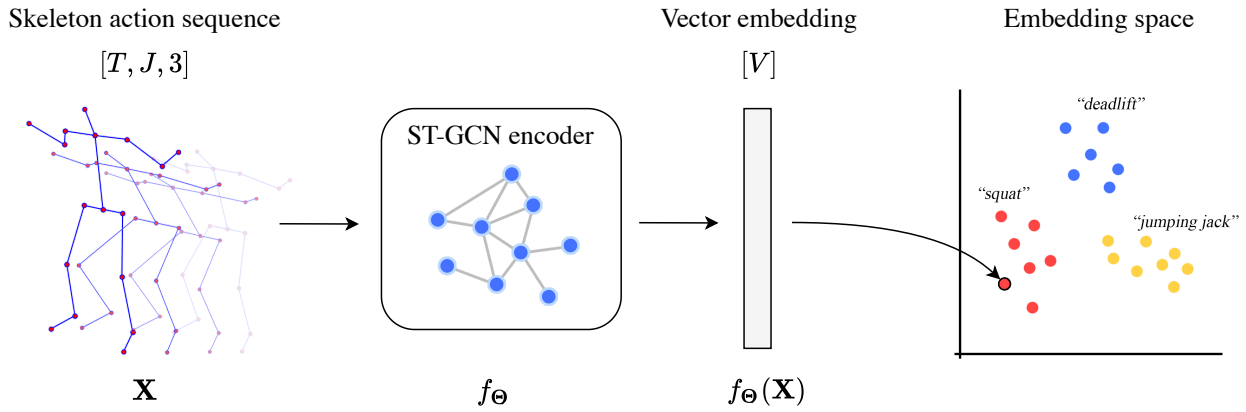


Figure 1. **The metric ST-GCN model**. Illustration of the spatial-temporal GCN feature extractor as a feature encoder in a metric learning paradigm. The model learns to distinguish between spatial-temporal features in the skeleton sequence even among never-before-seen actions.

a spatial-temporal graph convolutional neural network (ST-GCN) [31] as a feature extractor on *NTU RGB+D* with a metric learning paradigm to classify actions. Our metric GCN approach is shown in Figure 1. We then test the one-shot performance of the metric GCN on our novel dataset by providing only one example of each action class and determining the nearest class neighbour for all the samples in the dataset. The model achieves an accuracy of 87.4% on the seven never-before-seen classes.

Furthermore, we performed an ablation study that compares the use of triplet margin and multi-similarity loss, different mining strategies for metric learning and the effects of varying embedding sizes. We further investigated the effect of a reduction of the test set size and compare our model to the models found in the literature.

2. Related work

Skeleton motion sequences as a modality for action recognition has gained significant attention in recent years following the popularity of the Microsoft Kinect RGB-D camera, which allowed for easy 3D skeleton extraction using the Kinect SDK. This device was used to collect the state-of-the-art large-scale *NTU RGB+D* [16, 25] dataset which consists of 114,480 skeleton motion sequences of 120 different classes of actions. Since the release of this dataset, many deep learning architectures have been proposed to classify human action from pose sequences. Since skeleton motion sequence data is inherently spatial-temporal, models are tasked with the problem of fusing spatial and temporal information in the data for classification.

Many past works have proposed transforming skeleton sequences into pseudo-images in order to leverage the proven power of CNNs [1, 12, 13, 15, 19]. Notably, Liu *et al*. [15] arrange the skeleton joint indices into several 2D grids (skepxels) of different fixed orders that encode representations for the positions and velocities of joints in a frame. For the CNN input, they concatenate all the skepxel grids from a frame along the entire sequence. Caetano *et al*. [1] use a representation that encodes the magnitude of the temporal differences and orientation (angles) of joints in the sequences. Other works tackle the temporal modelling problem with recurrent network architectures [17,27], where frame-wise joint data is supplied to an RNN/LSTM unit sequentially and predictions are made on the output of the last frame given memory of previous frames.

However, current top-performing models are dominated by the graph convolutional network (GCN) architecture. In these approaches, convolution operations are performed on the skeleton directly by treating a skeleton sequence as a graph with joints and nodes. Yan *et al*. [31] proposed the first of these with the spatial-temporal graph convolutional network (ST-GCN), where the node vectors are the spatial coordinates of all joints in the sequences and the edges are both the semantic intra-frame connections and the temporal inter-frame connections. They experiment with different joint neighbour sampling strategies for convolutional kernels. Li *et al*. [14] achieved improved performance by adapting the graph edge formation to both semantic and action-based inference connections. Zuo *et al*. [33] achieved further improvements by dividing the skeleton data into several sub-graphs to learn spatial-temporal features at different part scales.

Much less research focus has been placed on training skeleton sequence action recognition models with little training resources such as a one-shot learning scenario. One-shot classification aims to classify novel, unseen examples given only a single reference example. Popular solutions to one-shot learning problems are meta-learning [5] and metric learning [8, 24].

With the release of *NTU RGB+D*, 120 [16] introduced a spatial-temporal LSTM module to extract feature embeddings from body parts and compare their similarity with Euclidean distance to make classifications. The authors also provide their one-shot test setup that researchers often copy to make fair performance comparisons. We also use this setup in our experiments.

Sabater *et al*. [23] presented a solution based on a temporal convolutional network (TCN). They calculate second-order features (bone angles and keypoint distances) and use a TCN to generate motion descriptors embeddings. They then perform a similarity evaluation between the embedding on the final frame of the anchor action and all the frames from a target motion sequence to detect when an action has occurred.

The only other work to use a metric learning approach for one-shot action recognition on skeleton data has been Memmesheimer *et al*. [20]. They follow the pseudo-image approach with a novel image representation which projects joint values for all axes into blocks over the width of the image which keeps joint spatial values grouped locally per axis. This results in a more compact representation which they use as input to a ResNet18 model trained in a triplet learning and multi-similarity learning paradigm. Metric learning has the added benefit of providing a similarity score, which is useful to detect anomalies or unknown action classes. All past works on one-shot skeleton-based action recognition have performed training and evaluation on the same dataset (*NTU RGB+D* and/or *Kinetics 400*) and none have considered using a large-scale dataset as a source for knowledge transfer of human action features representations. Since a system that can identify human action from one example can be so impactful in sport and fitness, we investigate one-shot action recognition on skeleton-based strength and conditioning actions.

3. Dataset

Our novel dataset, which we call the *Stellenbosch University Exercise Motion Dataset* (*SU-EMD*), uses the Vicon Vantage motion capture system and our markerless motion capture system to construct a labelled multimodal dataset of common exercise actions. The dataset comprises corresponding (i) multi-view RGB video clips, (ii) Vicon motion capture trajectory data, and (iii) markerless 3D pose estimation data, of five reps of seven different exercises being performed by four different subjects at three different speeds. It is designed for training and testing machine learning models for action recognition and analysis of gym exercises from human pose. This study was approved by the research ethics committee with which the authors are affiliated and participants provided their informed consent. Four male participants were recruited as subjects in the dataset.

For the markerless system, the 2D keypoint coordinates

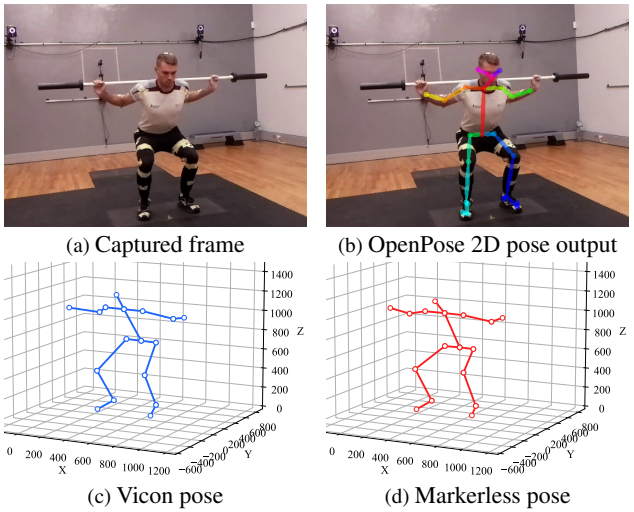


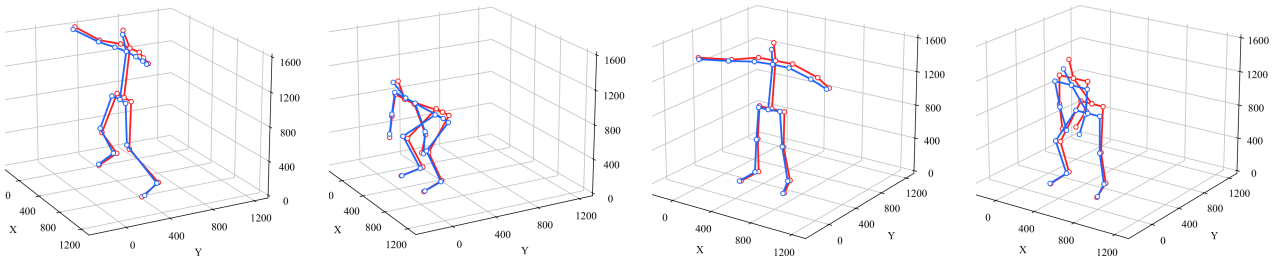
Figure 2. **Reconstructed 3D pose**. The sample S1A1D2R4 in our novel dataset at frame 70 with visualizations of the associated pose data.

of subjects are detected from video streams of four hardware synchronized cameras using OpenPose [2] and used to triangulate for their respective 3D locations in the world coordinate system using the direct linear transformation (DLT) and the projection matrices of each camera. For the marker-based motion capture, we use a gold-standard Vicon Vantage system. Retroreflective markers were placed on subjects such that the midpoints of pairs of markers denote the keypoints of interest, i.e. the OpenPose skeleton keypoints.

Table 1 outlines the matrix of the dimensions of data that were captured. We collected seven common strength and conditioning exercises: barbell back squat, barbell deadlift, dumbbell biceps curl, dumbbell lateral raise, kettlebell swing, jump rope, and jumping jacks. To add temporal variance to the data, we require subjects to vary action performances by performing actions at three different speeds: normal, fast and slow. For each of these actions and speeds, we captured five repetitions and sometimes more. This allows for at least 15 samples of an action which implies that any one subject performs 105 repetitions. Since we record actions with both our markerless motion capture system and the Vicon motion capture system, there are more than 840 data points. Inspired by *NTU RGB+D* [16], each sample dimension has a code which allows a unique and descriptive label for every sample. Sample S1A1D2R4 in the dataset is visualized in Figure 3.

3.1. Human pose data

Choosing a skeleton model similar to that of a large-scale labelled pose dataset that is already openly available is useful. If this is the case, these datasets can be inter-

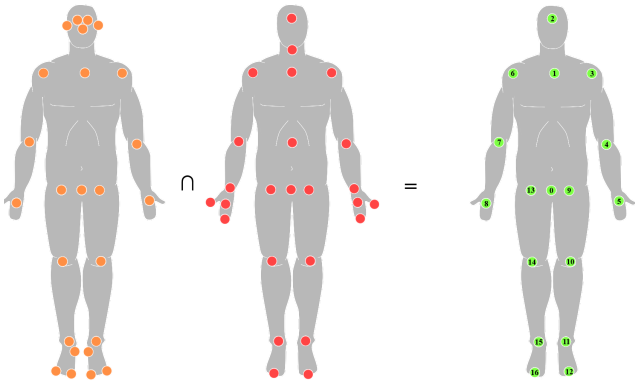


(a) Jumping jack (S4A7D1R4) (b) Deadlift (S1A2D1R1) (c) Lateral raise (S3A4D1R4) (d) Kettlebell swing (S2A5D1R1)

Figure 3. Visualizations of four of the seven strength and conditioning exercise actions available in our novel dataset. Blue shows the skeleton formed by the marker-based Vicon motion capture skeleton output and red shows the skeleton formed by the triangulated markerless pose estimation system at the same frame.

Table 1. **Captured motion samples**. The matrix of dimensions that are captured in the *SU-EMD* dataset along with their codes.

Action	Subject	Duration	Rep
Back squat (A1)	Male 1 (S1)	Normal (D1)	1 (R1)
Deadlift (A2)	Male 2 (S2)	Fast (D2)	2 (R2)
Biceps curl (A3)	Male 3 (S3)	Slow (D3)	3 (R3)
Lateral raise (A4)	Male 4 (S4)		4 (R4)
Kettlebell swing (A5)			5 (R5)
Jump rope (A6)			
Jumping jack (A7)			



BODY_25 (OpenPose) Microsoft Kinect (NTU RGB+D) SU-EMD (Ours)

Figure 4. ***SU-EMD* pose keypoints**. Our simple skeleton model is intended to be compatible with both the OpenPose and the *NTU RGB+D* formats.

operable. *NTU RGB+D* is such a dataset and is popular in the pose-based action recognition literature. It uses the *Microsoft Kinect* skeleton which is comparable to the *BODY_25* arrangement from OpenPose. Based on these specifications, we choose our keypoint set to be the intersection of the *BODY_25* arrangement and the *Microsoft Kinect* arrangement as illustrated in Figure 4. A skeleton sequence $\mathbf{X} \in \mathbb{R}^{T \times J \times D}$ consists of T frames of J joints in D spatial dimensions.

4. Approach

We propose to train a state-of-the-art graph convolutional architecture as an encoder model on the large-scale *NTU RGB+D* skeleton data. The GCN model learns to extract spatial-temporal features directly from skeleton sequence data and project them into an embedding space that clusters similar actions.

4.1. Spatial-Temporal GCN

Our metric GCN encoder is composed of a sequence of ST-GCN blocks as introduced by Yan *et al*. [31]. Figure 5 illustrates the model. The input to the GCN is shown, which is the skeleton motion sequence matrix, the ST-GCN blocks along with their residual skip connections, and the intermediate activation map dimensions. Similar to a CNN, the depths of the activation maps are increased as the network deepens.

The ST-GCN blocks apply batch normalisation internally. Along with the residual connections, this helps control vanishing or exploding gradients as the network becomes deeper. At the head of the model, after the feature extractor section $f_{\mathbf{\Theta}}$, a 2D global average pooling operation pools all the $T \times J$ activations down to a V-size vector $\mathbf{z}$. Different designs for V are described in Section 5. For the metric learning approach, the feature vector is used directly as a learned embedding.

4.2. Metric Learning

An overview of our metric learning approach is shown in Figure 1. We train a feature embedding $\mathbf{z} = f_{\mathbf{\Theta}}(\mathbf{X})$ which projects a skeleton sequence $\mathbf{X}$ to a vector representation $\mathbf{z} \in \mathbb{R}^V$ where V is the target embedding size. In this V-dimensional space, we wish for projections of ex-

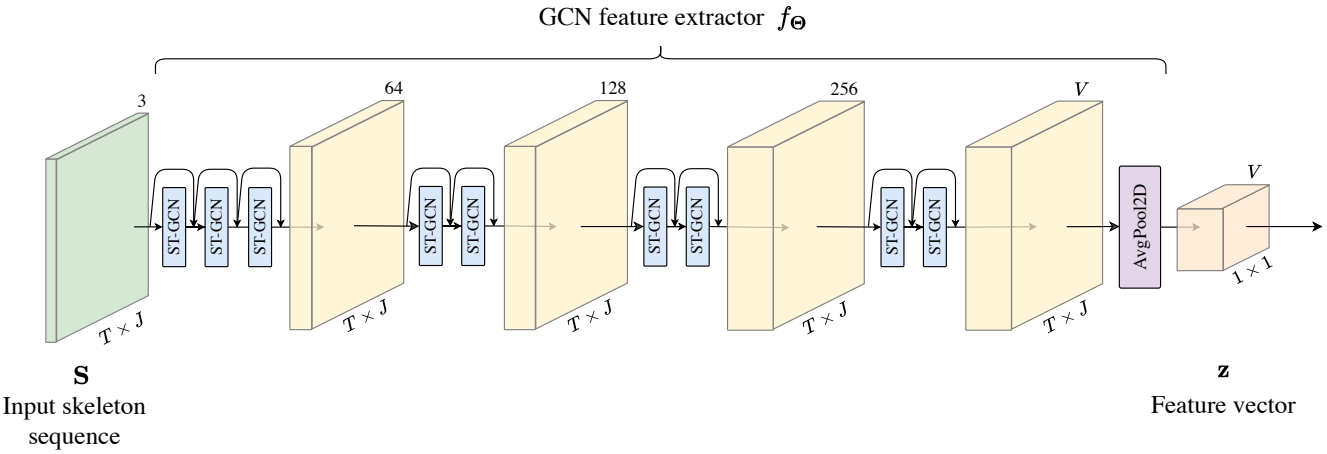


Figure 5. **The GCN feature extractor.** The ST-GCN blocks from Yan *et al*. [31] perform simultaneous spatial and temporal graph convolutions on the skeleton data.

amples from the same class to be closer together or have a higher vector similarity and for examples from different classes to be further away or have a lower similarity. The measure of distance between two projections $\mathbf{z}_i$ and $\mathbf{z}_j$ is defined as the Euclidean distance between the projections: $d(\mathbf{z}_i, \mathbf{z}_j) = \|\mathbf{z}_i, \mathbf{z}_j\|_2$. A measure for the similarity s_{ij} between two embeddings is taken as the dot product of the vectors: $s(\mathbf{z}_i, \mathbf{z}_j) = \langle \mathbf{z}_i, \mathbf{z}_j \rangle$. Two embeddings should have a low distance d_{ij} or high similarity s_{ij} for the samples ($\mathbf{X}_i$, $\mathbf{X}_j$) if their associated class labels are equal ($y_i = y_j$) and should have a high distance or low similarity if they are not equal ($y_i \neq y_j$). A model that generalizes well to the training set and is able to extract useful features from the data modality should be able to cluster these unseen classes in the metric space.

For their work on one-shot SBAR with metric learning, Memmesheimer *et al*. [20] experiment with both the Triplet Margin loss [24] and Multi-similarity loss [29]. Memmesheimer *et al*. is the only other work to use metric learning for one-shot SBAR. To have a more fair comparison, we also experiment with these two loss functions.

Firstly, for the **triplet learning** approach, we construct triplets from the B samples found in a mini-batch during training. Each triplet consists of a reference sample (anchor), a same-class sample (positive) and a different-class sample (negative). For a triplet consisting of embeddings $\mathbf{z}_a$, $\mathbf{z}_p$ and $\mathbf{z}_n$ for the anchor, positive and negative sample respectively, the triplet margin (TM) loss [24] is defined as:

$$\mathcal{L}_{\text{TM}} = \max\{d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + \alpha, 0\} \qquad (1)$$

where α is a margin hyperparameter greater than 0 which the model should learn as a sufficient separation between positive and negative embeddings. Therefore, the network is trained to minimize this loss function that penalises learned embeddings where the anchor-positive distance is not sufficiently smaller than the anchor-negative distance. For the mining of triplets, computing the loss from triplets containing either a hard negative or a semi-hard negative results in the best convergence [7]. A hard negative is defined as having been projected closer to the anchor than the positive and a semi-hard negative is one that has been projected further than the positive, but still within the margin α from the positive. Additionally, work by Xuan *et al*. [30] has shown the benefit of using easy positive sampling (EPS). In this strategy, only the most similar example of the same class is mined as a positive example. This loosened mining strategy is intended to avoid over-clustering and allow the model to learn more general features that associate examples for their semantic similarity instead of their class label.

Secondly, for the **multi-similarity** approach, we construct anchor-positive pairs and anchor-negative pairs whose contribution to the loss function is weighted depending on their similarities. We calculate all the similarities s between all the B samples in a mini-batch and collect them into a $B \times B$ matrix $\mathbf{S}$ that can be indexed as $\mathbf{S}_{ij} = s(\mathbf{z}_i, \mathbf{z}_j)$. The anchor-positive pairs for some anchor $\mathbf{z}_i$ comprise $\mathcal{P}_i$ and its anchor negative pairs comprise $\mathcal{N}_i$. The multi-similarity (MS) loss applies a weighting to every pair based on their similarity and is computed as:

$$\mathcal{L}_{\text{MS}} = \frac{1}{B}\sum_{i=1}^{B}\left\{\frac{1}{\beta}\log[1 + \sum_{k\in\mathcal{P}_i} e^{-\beta(\mathbf{S}_{ik}-\lambda)}] + \frac{1}{\gamma}\log[1 + \sum_{k\in\mathcal{N}_i} e^{\gamma(\mathbf{S}_{ik}-\lambda)}]\right\}$$

where the β and γ hyperparameters are the weights applied to the positive and negative pairs respectively and λ is the offset applied to the exponent which can be interpreted as

a similarity margin. For the multi-similarity approach, we also use the mining strategy proposed in original the paper [29]. To ensure that there are sufficient within-class samples to mine in a batch, we sample the training subset with the requirement that for a batch of size B samples, there must be M samples per class present in the batch. This implies that there can only be $\frac{B}{M}$ classes in a batch.

4.3. One-shot protocol

A one-shot classification setting uses a test set $\mathcal{T}$ that contains N novel classes such that none of the classes in C are in N. Additionally, one random exemplar for each of the N new classes is separated from the test set $\mathcal{T}$ to be used as a reference. There are N classes that the model has never been trained on and only one example per class is provided as a reference representation of those unseen classes. Since a labelled reference set is provided to obtain a class mean or 'centroid' for every class in N, the projected test sample is classified as belonging to the class with the nearest centroid in the embedding space. During the testing phase and for inference, we therefore use a k-nearest neighbour (k-NN) classifier implemented in the Faiss library [9]. From the k-NN, the top-k accuracy or accuracy@k is calculated as:

$$\text{Accuracy@}K = \frac{\#\text{correct}}{\|\mathcal{T}\|} \tag{2}$$

where #correct is the total number of true positives, i.e. how many test samples had their own class among their k nearest neighbours. The authors for the *NTU RGB+D* dataset document the one-shot protocol that they used in their experiments. They split the dataset into a training set $\mathcal{D}$ with $C = 100$ classes and a test set $\mathcal{T}$ with the remaining $N = 20$ unseen classes. We use this same partitioning for our experiments so that results can be compared fairly. In our problem setting, we use this as a validation set to design the hyperparameters for the model architectures and training, before performing final one-shot testing on the held-out *SU-EMD* data. Finally, once we have the top-performing model on the *NTU RGB+D* validation set $\mathcal{T}$, we perform final one-shot testing on all the *SU-EMD* data by choosing one random exemplars from each class. Since this has been held out from training and validation and contains completely unique classes from a different dataset, it will give an accurate measure of the one-shot capabilities of the metric GCN.

5. Experiments

5.1. Training

We train the ST-GCN embedder in Figure 5 as a feature extractor in a one-shot setting. We train the model on the *NTU RGB+D* training subset $\mathcal{D}$ for 100 epochs, which in our experiments have proven to be the point where the models converge on a validation accuracy.

We experiment with the multi-similarity (MS) approach and triplet margin (TM) approach. We also experiment with training with the easy positive sampling (EPS) strategy with the TM loss. Throughout training with TM, we start with a semi-hard mining approach for the first 80 epochs and use hard mining for the final 20 epochs. This applies for both positive and negative samples unless using EPS.

Like Roth *et al*. [22], we also experiment with varying embedding vector sizes. A larger embedding size will allow a greater number of dimensions to encode features, but a reduced embedding will create the information bottleneck required to force the model to learn more general features. We experiment with a small embedding size $V = 128$ and a large embedding size $V = 512$. Like Liu *et al*. [16], we investigate the effect of varying the number of training classes in $\mathcal{D}$.

We do not perform any other optimisation of hyperparameters. The learning hyperparameters in the loss functions are kept as the values used in the original papers. The dropout probabilities for the ST-GCN blocks remain at 0.5 throughout as in the original implementation [31].

5.2. Skeleton sequence data

Where the *NTU RGB+D* has $J = 25$ keypoints in their skeleton model, we have $J = 17$. For the datasets to be interoperable, it is necessary to subsample the *NTU RGB+D* keypoints to match our skeleton model by omitting the hand tip, wrist and thumb keypoints from the data (refer to Figure 4).

Furthermore, the skeleton motion sequences from these two datasets have different projections of the captured data. The distance units for *NTU RGB+D* correspond to real-world metres and the *SU-EMD* distances correspond to millimetres. Action samples are recorded in dynamic environments from different camera viewpoints and subjects may move around freely. This causes vastly different joint coordinate magnitudes and spatial relationships across data samples even for the the same action classes. These discrepancies can be factored out during skeleton normalization. We perform instance normalization on the data samples by subtracting the per-dimension means μ_d and dividing by the per-dimension standard deviations σ_d:

$$\hat{x}_{j,d}^{<t>} = \frac{x_{j,d}^{<t>} - \mu_d}{\sigma_d}$$

which can be calculated with each skeleton sequence sample as

$$\mu_d = \frac{1}{T \times J} \sum_{j=0}^{J-1} \sum_{t=0}^{T-1} x_{j,d}^{<t>}$$

and

$$\sigma_d = \frac{1}{T \times J} \sum_{j=0}^{J-1} \sum_{t=0}^{T-1} (x_{j,d}^{<t>} - \bar{\mu_d})^2.$$

Additionally, we apply augmentations to the skeletons during training. We experiment with random rotations, where the entire skeleton is rotated by random angles about the X and Y world axes. We also implement random moving, where the Gaussian noise is added per joint across time. To add temporal variation, we use random frame dropping with a 10% probability. Lastly, we introduce a novel augmentation called random pivoting. In most circumstances, slight pivoting of parts about certain joints will clearly represent the same action but in a different posture, which adds more variation to the data. For example, a lateral raise will still be considered a lateral raise even if the knees are slightly bent. These augmentations aim to guide the model to learn that motion patterns of certain parts of the body across time are more important for the action classes in question than others. To implement these, we define unit quaternions for rotation of the vectors. The use of quaternions allows for simple compounding of rotations and along arbitrary axes. Part pivoting differs from rotation in that the pivoting is done independently for every joint and the pivot is a parent joint given the natural hierarchical structure of a skeleton (e.g. hip $\rightarrow$ knee $\rightarrow$ ankle $\rightarrow$ foot tip), whereas the rotation is done for all joints about the world axes. All rotations are independently sampled from a zero-mean normal distribution with standard deviation $\frac{\pi}{7}$ radians and translation likewise with standard deviation $15\,\mathrm{mm}$.

5.3. Ablation study

Table 2 shows the validation test results under different architecture and training settings in our experiments. Shown are the accuracy@1 values on the 20 unseen *NTU RGB+D* one-shot test classes. From the results, it is clear that the multi-similarity loss approach outperforms the triplet loss approach. Similar to the findings of [20], the top-performing model was the MS loss approach with full augmentation, which classifies a sample from an unseen class correctly 46.2% of the time. Skeleton augmentation appears to be conducive to the learning of useful discriminative features and generalisation for the MS loss approach. It can also be seen that the embedding size does not have a significant impact on the learning. The easy positive sampling for the triplet margin mining showed only modest improvements in test accuracy.

5.4. Results

Table 3 shows the effect of taking the top-performing model and reducing the size of the training set. Also shown are the comparisons to other models in the literature. Note that this is not a fair comparison since we do not use all the skeletal joints in the *NTU RGB+D* skeleton model (we have sub-sampled the skeleton data to fit our *SU-EMD* skeleton model) but displaying these results provides an indication of the relative performance of the metric GCN approach.

The top-performing model is then tested on all the *SU-EMD* data samples, where only one exemplar was removed per class and used as reference embeddings for each. The model achieved an accuracy of 87.4% on these never-before-seen classes. Figure 7 shows the confusion matrix for the one-shot tests. There was high confusion for the *squat* and *deadlift* actions. There was also a high confusion for the jumping actions: the *jump rope* and *jumping jack* actions. In Figure 6, we provide the UMAP projection visualizations of both the validation samples and the final test samples.

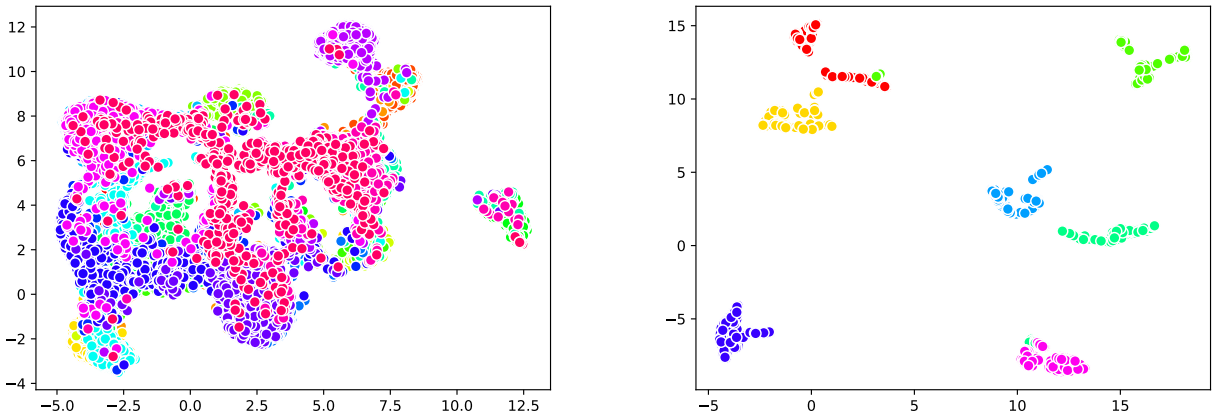
The model is able to learn features that cluster the seven never-before-seen exercise classes in the *SU-EMD* test dataset very well. The *NTU RGB+D* test set contains 20 novel classes whereas the *SU-EMD* test set contains only seven novel classes, which is an easier task. This means the test accuracies cannot be fairly compared. Future work

Table 2. **Ablation study for the metric GCN.** The validation results on the *NTU RGB+D* dataset are shown for different embedding sizes (V) under different loss approaches (TM = triplet margin, MS = multi-similarity) and data augmentation strategies. Rot = Random rotation, Pivot = Random pivoting, Drop = Frame dropping, Move = Random Moving.

			Accuracy@1 [%]	
Loss	EPS	Augmentations	128	512
TM	No	Rot, Pivot, Drop, Move	36.6	36.7
TM	Yes	Rot, Pivot, Drop, Move	37.0	37.7
TM	Yes	None	37.9	37.5
MS	—	Rot, Pivot, Drop, Move	**46.2**	45.8
MS	—	None	44.1	43.8

Table 3. Effect of training set size reduction on the *NTU RGB+D* dataset with one-shot protocol testing with comparisons to other models in the literature. All values are the one-shot test accuracy results in %.

# Training classes	20	40	60	80	100
APSR [16]	29.1	34.8	39.2	42.8	45.3
ST-LSTM [18]	—	—	—	—	42.9
TCN [23]	—	—	—	—	46.5
SL-DML [21]	36.7	42.4	49.0	46.4	50.9
Skeleton-DML [20]	28.6	37.5	48.6	48.0	54.2
JEANIE [28]	38.5	44.1	50.3	51.2	57.0
Part-aware PGN [3]	43.0	50.3	55.7	56.5	65.6
Metric GCN	28.8	34.8	39.1	43.3	46.2



(a) *NTU RGB+D* (46.2%) (b) *SU-EMD* (87.4%)

Figure 6. UMAP visualizations of projections by the metric GCN model of never-before-seen test samples. **(a)** shows projections among the 20 *NTU RGB+D* classes for one-shot validation and **(b)** shows the projections of seven *SU-EMD* classes. Projections are coloured by their true class labels.

should add more samples to *SU-EMD* so that there are 20 exercise actions.

The better performance on *SU-EMD* can be interpreted to suggest that exercise repetitions in the skeleton sequence modality have much lower intra-class variance in their spatial-temporal features as compared to the daily actions present in *NTU RGB+D*. Since an exercise (e.g. a *biceps curl*) is a series of motions with strict parameters for its performance, it should be performed similarly by all subjects. This restriction contrasts with a 'daily' action present in the training dataset (e.g. *put object into bag*). This is a welcome conclusion to our objective of creating a one-shot classifier for exercise actions and for exercise recognition systems in general. However, it could also imply that the *SU-EMD* dataset is not challenging enough in its current state and requires additional expansion; or that the reconstruction inaccuracies and noise present in the the *NTU RGB+D* dataset make the learning of discriminative features of actions particularly difficult. To support this, consider that it is challenging even for the authors to classify the action being performed in a visualization of an *NTU RGB+D* data sample.

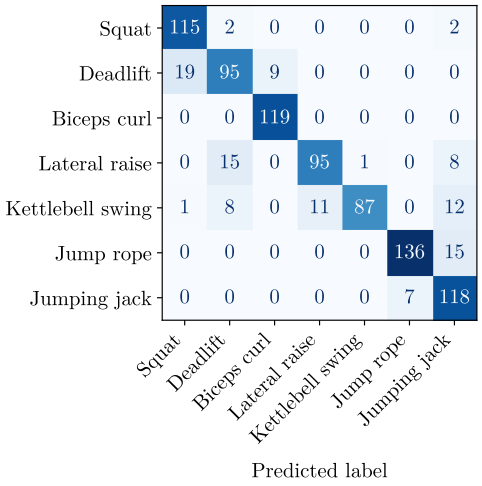


Figure 7. Confusion matrix for the final one-shot tests in the seven *SU-EMD* classes.

6. Conclusion

This work contributes to the development of a practical action recognition system for strength and conditioning movements in sports. We presented a novel dataset of skeleton-based strength and conditioning actions. Our dataset comprises of 840 samples of seven exercise action classes and we have made it available to researchers to study action recognition in sport, health and fitness.

We used our dataset as a one-shot test set and trained on a separate large-scale dataset to examine the feasibility of knowledge transfer of spatial-temporal features using the state-of-the-art ST-GCN architecture. We have shown that skeleton augmentations (random moving, rotation and frame dropping and pivoting) and the multi-similarity loss achieve the top performance on our validation set. Our finding is that a regular ST-GCN trained as a feature embedding in a metric learning paradigm competes with but does not improve on the state-of-the-art.

Yet, we have shown for the first time that spatial-temporal features can be easily learned and transferred to classify never-before-seen classes of exercise with high accuracy. This method can be used as a means to classify sports actions in a virtual coaching system where users cannot provide many expert examples for the enrolment of new actions.

References

[1] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A. Dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019. 2

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021. 3

[3] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Qian He, Chuanyang Hu, Errui Ding, Yu Guan, and Xuming He. Part-aware prototypical graph network for one-shot skeleton-based action recognition, 2022. 7

[4] William P. Ebben, Ryan M. Carrol, and Christopher J. Simens. Strength and conditioning practices of national hockey league strength and conditioning coaches. *The Journal of Strength & Conditioning Research*, 18(4), 2004. 1

[5] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1126–1135. JMLR.org, 2017. 2

[6] Jürgen Freiwald, Matthias W. Hoppe, Sasha Javanmardi, Thilo Hotfiel, Martin Engelhardt, Casper Grim, and Christian Baumgart. Current misjudgments and future trends in rehabilitation after knee injuries (part 1). *Sports Orthopaedics and Traumatology*, 36(3):250–259, 2020. Thema: Krafttrainung / Kraftsport // Strength Training / Strength Sports. 1

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[8] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco Loog, editors, *Similarity-Based Pattern Recognition*, pages 84–92, Cham, 2015. Springer International Publishing. 2

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[10] Thomas W. Jones, Andrew Smith, Lindsay S. Macnaughton, and Duncan N. French. Strength and conditioning and concurrent training practices in elite rugby union. *The Journal of Strength & Conditioning Research*, 30(12), 2016. 1

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 1

[12] Bo Li, Mingyi He, Xuelian Cheng, Yucheng Chen, and Yuchao Dai. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN, 2017. 2

[13] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using LSTM and CNN. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 585–590, 2017. 2

[14] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[15] Jian Liu, Naveed Akhtar, and Ajmal Mian. Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[16] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 2, 3, 6, 7

[17] Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):3007–3021, 2017. 2

[18] Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot, and Gang Wang. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):3007–3021, 2018. 7

[19] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 2

[20] Raphael Memmesheimer, Simon Häring, Nick Theisen, and Dietrich Paulus. Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 837–845, 2022. 3, 5, 7

[21] Raphael Memmesheimer, Nick Theisen, and Dietrich Paulus. SL-DML: Signal level deep metric learning for multimodal one-shot action recognition, 2020. 7

[22] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Björn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *ICML*, pages 8242–8252, 2020. 6

[23] Alberto Sabater, Laura Santos, Jose Santos-Victor, Alexandre Bernardino, Luis Montesano, and Ana C. Murillo. One-shot action recognition in challenging therapy scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2777–2785, June 2021. 3, 7

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 2, 5

[25] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1, 2

[26] Anthony N. Turner and Perry F. Stewart. Strength and conditioning for soccer players. *Strength & Conditioning Journal*, 36(4), 2014. 1

[27] H. Wang and L. Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3633–3642, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 2

[28] Lei Wang, Jun Liu, and Piotr Koniusz. 3D skeleton-based few-shot action recognition with JEANIE is not so naïve, 2021. 7

[29] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5017–5025, 2019. 5, 6

[30] H. Xuan, A. Stylianou, and R. Pless. Improved embeddings with easy positive triplet mining. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2463–2471, Los Alamitos, CA, USA, mar 2020. IEEE Computer Society. 5

[31] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 2, 4, 5, 6

[32] Warren B. Young. Transfer of strength and power training to sports performance. *International Journal of Sports Physiology and Performance*, 1(2):74 – 83, 2006. 1

[33] Qi Zuo, Lian Zou, Cien Fan, Dongqian Li, Hao Jiang, and Yifeng Liu. Whole and part adaptive fusion graph convolutional networks for skeleton-based action recognition. *Sensors*, 20(24), 2020. 2