LSTA-Net: Long short-term Spatio-Temporal Aggregation Network for Skeleton-based Action Recognition

Tailin Chen*
t.chen14@newcastle.ac.uk
Open Lab
Newcastle University, UK

Shidong Wang*
shidong.wang@newcastle.ac.uk
Open Lab
Newcastle University, UK

Desen Zhou
zhoudesen@baidu.com
VIS
Baidu, Inc., China

Yu Guan
yu.guan@newcastle.ac.uk
Open Lab
Newcastle University, UK

Abstract

Modelling various spatio-temporal dependencies is the key to recognising human actions in skeleton sequences. Most existing methods excessively relied on the design of traversal rules or graph topologies to draw the dependencies of the dynamic joints, which is inadequate to reflect the relationships of the distant yet important joints. Furthermore, due to the locally adopted operations, the important long-range temporal information is therefore not well explored in existing works. To address this issue, in this work we propose LSTA-Net: a novel Long short-term Spatio-Temporal Aggregation Network, which can effectively capture the long/short-range dependencies in a spatio-temporal manner. We devise our model into a pure factorised architecture which can alternately perform spatial feature aggregation and temporal feature aggregation. To improve the feature aggregation effect, a channel-wise attention mechanism is also designed and employed. Extensive experiments were conducted on three public benchmark datasets, and the results suggest that our approach can capture both long-and-short range dependencies in the space and time domain, yielding higher results than other state-of-the-art methods.[1]

1 Introduction

In recent years, skeleton-based action recognition became of popular research topic in the computer vision community[1, 27] due to the advent of cost-effective depth cameras [30] and reliable human pose estimation methods[2]. Compared with conventional RGB video based action recognition, the data structure of skeleton representation is in low-dimension and hence it can be easily stored in devices and transferred.

Skeleton-based action recognition is a challenging task due to the lack of context information compared with RGB video based action recognition. In particular, modeling the long range dependencies in both spatial and temporal dimensions is difficult due to the

[1]Code available at https://github.com/tailin1009/LSTA-Net. * Equal Contribution

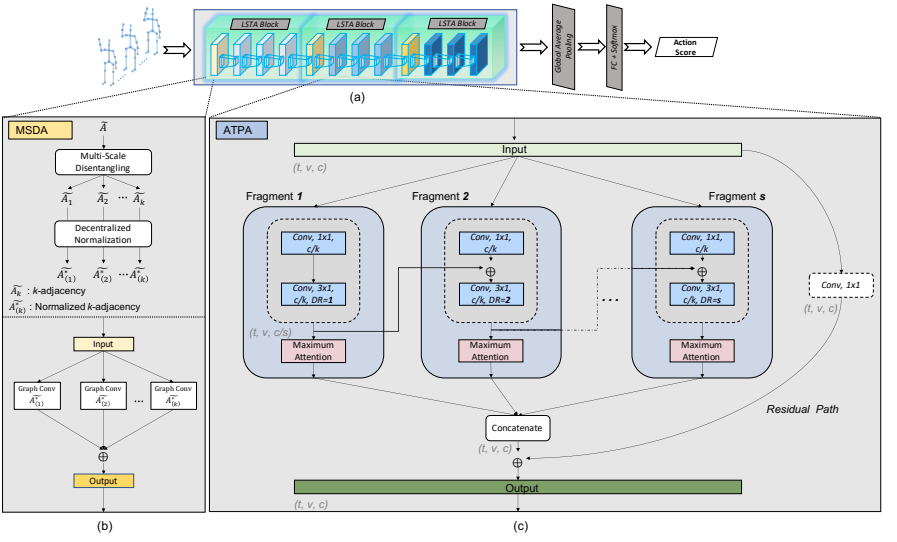

Figure 1: An illustration of the proposed LSTA-Net (a). Each LSTA block includes one MSDA module (b) and three ATPA (c) modules.

flexible configuration between different semantic parts, as well as complex movement patterns in time domain. Early works used sequential methods to model joint relations, which first extracted features of frames by treating joints as pixels and then utilised RNN-based models to model temporal relationships[6, 15]. some other works concatenated skeletons of different time steps to generate a pseudo image, and utilised CNNs to perform image-based classification[10, 11]. However, human skeletons are not in an Euclidean space and hence above methods cannot model joints' dependencies effectively. A more natural method is to model the relationship between joints through a graph neural networks (GNNs). In view of the vigorous development of convolution operations, graph convolutional networks (GCNs) have been widely used for skeleton-based action recognition due to the powerful ability for modeling non-Euclidean data. In [27] ST-GCN was proposed, which constructed a spatial graph based on the natural connections of human joints and introduced temporal edges between corresponding joints in consecutive frames. Many of the latest works [3, 13, 14, 22, 23, 26] can be regarded as variants of ST-GCN, where they typically applied different functional modules for better feature representation.

Despite the significant improvements in performance, there still exists some limitations in the above methods. Specifically, both recurrent and convolutional operations are based on local neighbours in the spatial or temporal domain. The capture of long-range dependencies only can be achieved by repeatedly performing these operations and gradually propagating signals through the data and hence it is inefficient. Directly modelling the distant joints relations and long-range temporal information is essential for distinguishing various actions. Recent work MS-G3D [17] proposed a disentangled and unified spatial-temporal graph convolution strategy to model the long range dependencies in a multi-scale manner. However, the proposed G3D module highly relies on constructing multiple pathways and hence leads to a complex model architecture with high computational costs.

In this paper, to model both long/short-range joints relations in spatial domain, and long/short-term joints dynamics in temporal domain, we propose a novel long short-term

spatio-temporal aggregation network (LSTA-Net). Specifically, each LSTA block includes a multi-scale decentralised aggregation (MSDA) module and three attention-enhanced temporal pyramid aggregation (ATPA) modules. MSDA is proposed to model the semantically related intrinsic connectivity of the disentangled/distant joints in the spatial domain, while ATPA is proposed to model long-range temporal dynamics by employing a set of sub-convolutions and formulating them with a pyramid-like/hierarchical structure. We further utilise maximum-response attention module (MAM) for further performance improvement. The main contributions of our work lie in three folds:

- We propose multi-scale spatial decentralised aggregation (MSDA) for distant/long-range dependency modelling by introducing a simple normalisation strategy.
- We propose a novel attention-enhanced long short-range temporal modelling architecture, where the temporal receptive field can be efficiently enlarged by the pyramid aggregation scheme. An attention mechanism is also devised for feature enhancement.
- Our LSTA-Net, despite smaller model size, achieves higher or comparable results when compared with other state-of-the-arts on three public action recognition datasets, suggesting its effectiveness.

2 Methodology

In Fig.1, we demonstrate the overall architecture of the proposed LSTA-Net, and in this section we introduce the basic unit LSTA block. Each LSTA block contains a multi-scale spatial decentralised aggregation module (MSDA , i.e., Fig.1(b)), and three attention-enhanced temporal pyramid aggregation (ATPA, i.e.,Fig.1(c)) modules to extract the spatial and temporal features for skeleton-based action recognition.

2.1 Multi-scale Spatial Decentralised Aggregation(MSDA)

2.1.1 Preliminaries

Notations A graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V})$ is the set of edges. The relationship of the graph is represented by the graph adjacency matrix $\mathbf{A} \in \mathbb{R}^{V \times V}$ with entry $\mathbf{A}_{ij} = 1$, when nodes i, j are connected, and 0 otherwise; V denotes the number of vertexes. $\mathbf{A}$ is a symmetric matrix while $\mathcal{G}$ is an undirected graph. Human graph sequences contain a set of node features $\mathcal{X} = \{x_t^v | 1 \leq v \leq V, 1 \leq t \leq T; v, t \in \mathbb{Z}\}$ that can also be represented as $\mathbf{X} \in \mathbb{R}^{T \times V \times C}$, where C is the feature dimension.

Multi-Scale Aggregation The multi-scale spatial aggregation [14] on a given graph can be implemented similar to the convolution on a regular grid graph, such as the RGB image. At timestamp t, given input graph feature $\mathbf{X}_t \in \mathbb{R}^{V \times C}$, via a multi-scale GCN operation we can get the output skeleton feature

$$\mathbf{X}_t^{\text{spat}} = \sigma\left(\sum_{k=0}^{K} \widetilde{\mathbf{A}_k} \mathbf{X}_t \mathbf{W}_k\right), \tag{1}$$

where K is the number of scales of the graph to be aggregated; $\sigma(\cdot)$ is the activation function; $\tilde{\mathbf{A}}$ is the normalised adjacency matrix [14] that can be obtained by: $\tilde{\mathbf{A}} = \hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}}$, where

$\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{I}$ and $\hat{\mathbf{D}}$ are the identity matrix and degree matrix of $\mathbf{A}$ respectively. The term $\widetilde{\mathbf{A}_k}\mathbf{X}_t$ in Eq.(1) describes a weighted sum feature that is based on the k-order neighbourhood of the selected graph nodes.

2.1.2 Decentralised Normalisation Strategy for MSDA

One drawback of the aforementioned multi-scale aggregation approaches is that it may overweight the low-order neighbours [14, 20]. To solve such biased weighting problem, in [17] Liu et al. proposed a disentangled aggregation scheme. However, such disentangled representation neglected the correlations between different scales and hence suffered from limited spatial representation capacity. To address this issue, here we employ a simple decentralised normalisation strategy based on the disentangled representation which aims to expand such representation to multiple scales.

In the decentralised normalisation strategy, the elements of k-adjacency matrix $\widetilde{\mathbf{A}_{(k)}}$ are first assigned the value 1 if node $i = j$ or with their shortest distance $d(i,j) = k$. For those elements whose shortest distance $d(i,j) < k$, a scale-adaptive value $\frac{d(i,j)}{k}$ is assigned. The rest of elements are set to 0.

It is clear when $k = 0$ or $k = 1$, we have $\widetilde{\mathbf{A}_{(0)}} = \mathbf{I}$, or $\widetilde{\mathbf{A}_{(1)}} = \widetilde{\mathbf{A}}$. When $k \geq 2$, similar to [17], the normalised k-adjacency matrix $\widetilde{\mathbf{A}^*_{(k)}}$ can be obtained by calculating the residuals of the matrix powers of current graph scale and the mean of previous graph scales, i.e.,

$$\widetilde{\mathbf{A}^*_{(k)}} = \mathbf{I} + \mathbb{1}\left(\widetilde{\mathbf{A}_k} \geq 1\right) - \mathbb{1}\left(\left(\frac{1}{k}\sum_{n=0}^{k-1}\widetilde{\mathbf{A}_n}\right) \geq 1\right). \tag{2}$$

By substituting $\widetilde{\mathbf{A}_k}$ with $\widetilde{\mathbf{A}^*_{(k)}}$ in Eq.(1), spatial feature can be extracted via:

$$\mathbf{X}_t^{\text{spat}} = \sigma\left(\sum_{k=0}^{K}\widetilde{\mathbf{A}^*_{(k)}}\mathbf{X}_t\mathbf{W}_{(k)}\right), \tag{3}$$

which is our MSDA module. Compared with [17], the application of decentralised normalisation strategy (i.e., with the scale-adaptive value assigning scheme) in MSDA can well summarise multiple scales, capturing the dependencies between short and long-range joints in the spatial domain.

2.2 Attention-enhanced Temporal Pyramid Aggregation(ATPA)

MSDA is able to capture the dependencies between both short and long-range joints in space domain, and it is desirable to investing the long short-range temporal modeling. Here the proposed Temporal Pyramid Aggregation (TPA) module divides the convolution operation of the input features into a group of subsets, which can effectively expand the equivalent receptive field of the temporal dimension without introducing additional parameters or time-consuming operations.

2.2.1 Temporal Pyramid Aggregation (TPA)

After spatial feature extraction (using MSDA, i.e., Eq.(3)), the original skeleton graph sequence can be represented as $\mathbf{X}^{\text{spat}} \in \mathbb{R}^{T\times V\times C'}$, where C' is the feature dimension. In this

subsection, we introduce temporal pyramid aggregation (TPA) for effective temporal information encoding.

To exploit the temporal information in skeleton-based action recognition, previous works [13, 14, 22, 27] used temporal convolution on the neighbouring timestamps/frames, and performed repeated stacking for long-range temporal dependency modelling. However, useful features from distant frames may have been weakened after a large number of local convolution operations. In [17], Liu et al. expanded the temporal receptive field by composing a large number of local operations, which increase the model size substantially, with extremely high computational costs. In [7, 28], Res2Net-like architectures were introduced, which deformed the ordinary convolution layer into a set of sub-convolutions and constructed hierarchical residual-like connections to capture multi-scale feature representations. Motivated by [7, 28], here we employ this concept to skeleton-based action recognition for fast and efficient long-range temporal dependencies modelling.

Given $\mathbf{X}^{\text{spat}} \in \mathbb{R}^{T\times V\times C'}$, along the feature dimension, we can obtain S embedded fragments $\{\mathbf{X}_s | s = 1,2,...,S; \mathbf{X}_s \in \mathbb{R}^{T\times V\times \alpha}\}$ through multiple learnable transformations, where α is the feature dimension in the embedding space, and we set $\alpha = \lfloor C'/S \rfloor$. Then, these fragments can be formulated as a hierarchical residual architecture and thus can be hierarchically processed by temporal convolutions with gradually increasing dilation rates. Specifically, assume $S = 6$, this process can be written as:

$$\begin{aligned} \mathbf{X}_s^{\text{temp}} &= conv_{\text{temp}} * \mathbf{X}_s, && s = 1, \\ \mathbf{X}_s^{\text{temp}} &= conv_{\text{temp}} * \left(\mathbf{X}_s + \mathbf{X}_{s-1}^{temp}\right), && s = 2,3,4,5,6, \end{aligned} \tag{4}$$

where $\mathbf{X}_s^{\text{temp}}$ is the output of temporal convolution in s-th fragment; $conv_{\text{temp}}$ denotes the 3×1 temporal sub-convolution.

The above operations endow the different fragments with different receptive fields in temporal dimension, which can model the long-range temporal dependencies effectively. The final output can be easily obtained by concatenating outputs of multiple temporal convolutions as follows:

$$\mathbf{X}^{\text{temp}} = [\mathbf{X}_1^{\text{temp}}; \mathbf{X}_2^{\text{temp}}; \mathbf{X}_3^{\text{temp}}; \mathbf{X}_4^{\text{temp}}; \mathbf{X}_5^{\text{temp}}; \mathbf{X}_6^{\text{temp}}]. \tag{5}$$

Through TPA, we can obtain representation $\mathbf{X}^{\text{temp}}$, which has encoded various range of temporal information.

2.2.2 Maximum-response Attention Module (MAM)

Although the combination of MSDA and TPA modules can model such long-range and short-range spatial temporal dependencies, human skeleton sequences comprise a limited number of dynamic key joints, which supplies a favourable breeding ground for the development of attention mechanisms. Employing attention mechanism can spontaneously capture useful intrinsic correlations without knowing the content of the input sequence. These motivate us to explore an effective and efficient attention mechanism to extract the semantic dependence of skeletal data. Very few works explored the attention mechanism in skeleton-based action recognition. In [20], Shi et al. applied SE-like[8] attention modules to re-weight the feature maps in spatial, temporal and channel dimension sequentially yet it failed in capturing the joint attention. In [25], an Efficient Channel Attention (ECA) module was used to capture channel attention, yet it only included a fixed kernel and may be inadequate when facing complex data. Motivated by [25] in our multi-scale aggregation schemes, we expand the

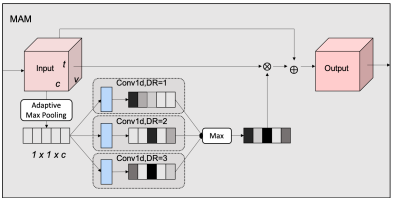

Figure 2: An illustration of the proposed MAM Module.

original ECA from single path to multiple paths and then use an adaptive maximum pooling operation for the most important features.

In Fig.2, we can see the structure of this Maximum-response Attention Module(MAM). Given an input $\mathbf{X} \in \mathbb{R}^{T\times V\times C}$, the channel attention can be computed by a standard $1D$ convolution with kernel size of η :

$$\omega = \tau\left(conv1D_{\eta}\left(g(\mathbf{X})\right)\right). \tag{6}$$

where $\tau(\cdot)$ is the sigmoid function; $g(\mathbf{X})$ denotes the function of 2D Adaptive Max Pooling operated on spatial and temporal dimensions. Although the $1D$ convolution only has η parameters, it provides a very limited receptive field, which leads to insufficient local information interaction. To tackle this problem, we extend the $1D$ convolution to a parallel fashion with different dilation rates. Subsequently, we apply the element-wise maximum operator to obtain the highest response of the input feature to the classifier. We propose to formulate this manipulation in the following manner:

$$\omega_{max} = \max_{\phi\in\Phi}\left(\phi(\omega)\right). \tag{7}$$

where $\phi(\omega)$ represents a set of $1D$ convolutions stacking with different dilation rates, and Φ is the total length of attending convolutions. The new features ω_{max} is constructed in such a way not only enhances the interactivity of information, but also makes their output independent of the receptive fields known in advance by the various kernels.

2.3 LSTA Block

For skeleton-based action recognition, the joint number tends to be small (e.g., 25) when compared with the frame number (e.g., 300). Therefore, we use more ATPA modules (for temporal modelling) than MSDA modules (for spatial modelling) for each LSTA block. Specifically in our LSTA-Net, each LSTA block is composed of one MSDA module followed by three ATPA modules to perform spatial temporal aggregation:

$$M_{\text{ATPA}}(M_{\text{ATPA}}(M_{\text{ATPA}}(M_{\text{MSDA}}(\mathbf{X})))) \tag{8}$$

where $\mathbf{X}$ denotes the input. In our LSTA-Net, we set multiple LSTA blocks as shown in Fig. 1, and more details can be found in Experiments Section.

3 Experiments

3.1 Datasets

NTU RGB+D 60 and NTU RGB+D 120 NTU RGB+D 60[19] and 120[16] contain 56,580 and 114,480 skeleton sequences corresponding to 60 and 120 action categories respectively.

Methods	Params	NTU RGB+D 60		NTU RGB+D 120	
		X-Sub	X-View	X-Sub	X-Set
PA-LSTM [19]	–	60.7	67.3	25.5	26.3
ST-LSTM [15]	–	69.2	77.7	55.0	57.9
VA-LSTM [29]	–	79.4	87.6	–	–
TCN [11]	–	74.3	83.1	–	–
AGC-LSTM [23]	–	89.2	95.0	–	–
ST-GCN [27]	3.1M	81.5	88.3	70.7	73.2
AS-GCN [14]	-	86.8	94.2	77.9	78.5
2s-AGCN [23]	6.9M	88.5	95.1	82.9	84.9
DGNN [21]	26.2M	89.9	96.1	–	–
NAS-GCN [18]	6.6M	89.4	95.7	–	–
4s-Shift-GCN [5]	2.8M	90.7	96.5	85.9	87.6
4s-DC-GCN+ADG [4]	-	90.8	**96.6**	86.5	88.1
PA-ResGCN [24]	3.6M	90.9	96.0	87.3	88.3
MS-G3D [17]	6.4M	**91.5**	96.2	86.9	88.2
LSTA-Net(Ours)	3.1M	**91.5**	**96.6**	**87.5**	**89.0**

Table 1: Model Comparison (in Top-1 accuracy (%)) on the NTU RGB+D 60 & 120 datasets.

Each skeleton sequence contains the 3D spatial coordinate information of 25 joints captured by Microsoft Kinect v2 cameras. We follow the same protocols in prior works[17, 23],i.e., Cross-Subject(X-Sub) and Cross-View (X-View) experiments for NTU RGB+D 60; while Cross-Subject(X-sub) and Cross-Subject(X-sub) for NTU RGB+D 120; More details of the train/test splits in different settings can be found in [17, 22].

Kinetics-Skeleton The Kinetics-Skeleton is a large-scale dataset sourced from the Kinetics 400 video dataset [9] by using the OpenPose[2] pose estimation toolbox. Effective skeleton sequences are divided into two groups, 240,436 for training and 19,796 for testing. Each skeleton sequence contains the 2D spatial coordinate information of 18 joints and the corresponding confidence scores. Following the previous works[17, 22], Top-1 and Top-5 accuracies are reported.

3.2 Implementation Details

We implemented the proposed LSTA-Net using PyTorch toolkit and ran on a server with four Tesla-V100 GPUs. The batch size was set to 64 (16 per worker). The model was trained for 100 epochs with Nesterov momentum (0.9) SGD and the cross-entropy loss. The initial learning rate was set to 0.05 and decayed with a factor of 0.1 at epoch {40,60,80,100}. The weight decay was set to 0.0005 for all experiments. The architecture is the same as the factorised path in [17] but with the different output channels (i.e., 72, 144 and 288 for each LSTA block in sequential). The input skeletal data is padded to T=300 frames by replaying the actions. All sequences were pre-processed with normalisation and translation as employed in [17, 21, 22].

3.3 Comparison with State-of-the-arts

Many state-of-the-art(SOTA) methods utilised multi-stream fusion strategies to fuse different modalities data for higher results. For fair comparison, we employed the similar multi-stream fusion strategy as [5, 21] , and devised our framework in the three-stream fashion where joint, bone, motion streams were sharing one identical architecture. The initialisation of the "bone"

Methods	Kinetics Skeleton	
	Top-1	Top-5
PA-LSTM [19]	16.4	35.3
TCN [11]	20.3	40.0
ST-GCN [27]	30.7	52.8
AS-GCN [14]	34.8	56.5
2s-AGCN [22]	36.1	58.7
DGNN [21]	36.9	59.6
NAS-GCN [18]	37.1	60.1
MS-G3D [17]	38.0	**60.9**
LSTA-Net (Ours)	**38.1**	60.7

Table 2: Model Comparison (in Top-1/5 accuracy (%)) on the Kinetics Skeleton dataset.

Method	Params	X-Sub	X-View
Shift-GCN[5]	0.7M	87.8	95.1
MS-G3D[17]	3.2M	**89.4**	95.0
LSTA-Net(Ours)	1.0M	89.1	**95.3**

Table 3: Model comparison (in parameter number and Top-1 accuracy(%)) on the NTU RGB+D 60 dataset (joint data only).

stream was set to the vector difference of adjacent joints directed away from the center of the human body. Then the "motion" stream used the temporal difference between adjacent frames of "joint" or "bone" as input. Finally, a score-level fusion strategy was applied to obtain the final prediction score.

We compare our full model with other state-of-the-art methods on NTU-RGB+D 60, NTU-RGB+D 120 and Kinetics-Skeleton datasets and the results are shown in Table 1 and Table 2. On NTU RGB+D 60 dataset, we achieve competitive performance on cross-view and cross-subject benchmarks. For NTU RGB+D 120, our method outperforms other methods on both cross-subject and cross-setup benchmarks. We additionally compare model parameter number with several SOTA GCN models in Table 1, suggesting our proposed LSTA-Net is a light-weight scheme yet with the best performance on both datasets. For Kinetics-skeleton dataset, as can be seen from Table 2, our proposed LSTA-Net achieves comparable performance with MS-G3D [17] and outperforms others. However, our model only has 50% of the parameters as in MS-G3D[17], suggesting the effectiveness of our aggregation scheme.

In Table 3, we report the experimental results using only the original skeleton data on NTU RGB+D 60 dataset, and compare with other approaches (i.e., Shift-GCN[5] and MS-G3D[17]). Although shift-GCN[5] is a light-weight model with only 0.7M parameters, our model outperforms it by a large margin with only 0.3M additional parameters, suggesting it is an alternative solution at balancing the trade-off between efficiency and effectiveness. On the other hand, when compared with MS-G3D[17], our method achieves comparable results with only 1/3 parameters. These observations suggest our approach is an effective light-weight solution.

3.4 Ablation Study

In this section, we report the results of our ablation study to validate the effectiveness of our proposed model components or strategies. Unless stated, performance is reported as classification accuracy on the Cross-Subject benchmark of NTU RGB+D 60 dataset using

Spatial Aggregation	Number of Scales			
	k=1	k=4	k=8	k=12
GCN	87.1	88.2	88.6	88.1
MS-GCN[17]	87.1	88.2	88.9	88.2
MSDA (Ours)	87.1	88.3	**89.1**	88.3

Table 4: Ablation study on spatial aggregation schemes, with top-1 accuracies reported (%).

Temporal Aggregation	Params	Attention	Acc
MS-TCN[17]	1.2M	-	88.2
TPA(Ours)	1.0M	-	88.5
ATPA(Ours)	1.0M	✓	**89.1**

Table 5: Comparisons between regular MS-TCN and our TPA module with or without Temporal Maximum Attention, with Top-1 accuracies reported (%).

Method	Number of Subsets	Acc
ATPA	$S=4$	89.0
	$S=6$	**89.1**
	$S=8$	88.9

Table 6: Comparisons between TPA modules with respect to S. S is the number of subsets for sub-convolution operations, with Top-1 accuracies reported (%).

MAM	Number of Dilation Rates						
	$\eta=3$	$\eta=5$				$\eta=7$	$\eta=9$
	1∼3	1	1∼2	1∼3	1∼4	1∼3	1∼3
w/ Average-Pooling	88.5	88.3	88.5	88.8	88.5	88.3	88.2
w/ Max-Pooling	88.6	88.3	88.6	**89.1**	88.4	88.4	88.1

Table 7: Parameter selection of MAM with Average-Pooling/Max-Pooling functions. η denotes the kernel size of $1D$ convolution as in Eq.(6), and 1∼3 indicates the dilation rates of 1, 2 and 3.

the joint data only.

MSDA module In Table 4, we compare the proposed MSDA with the basic adjacency powering method and disentangling [17] method in terms of scale number. We replace the spatial aggregation strategy of the LSTA blocks, referred to as "GCN", "MS-GCN"[17] and "MSDA", respectively. We observe that our decentralised aggregation strategy MSDA can outperform basic adjacency powering method on different scales.

ATPA module To validate the effectiveness of our attention-enhanced temporal aggregation method, we conducted ablation experiments on different temporal aggregation schemes, and the results are shown in Table 5. From the table we can see that our proposed TPA (w/o attention) scheme outperforms the direct aggregation MS-TCN[17]. The final result can be further improved when combining with maximum response attention (ATPA). We also conducted extensive experiments to explore the hyper-parameters S in TPA and η in MAM module, and the results are shown in Table 6 and Table 7, suggesting both hyper-parameters are quite stable. Additionally, we compare different pooling functions in our MAM module. As shown in Table 7, the max-pooling yields the best accuracy while the average-pooling achieves more stable results, indicating it is less sensitive to kernel size. Furthermore, at the end of the entire network, a global average pooling is adopted for final feature generation.

Visualisation We visualise the output feature maps of the last LSTA-block in Fig.3. For the spatial modelling (TOP), the size of green circle around each joint indicates its importance. We can see our model can focus on the parts that are most relevant to the action. Specifically,

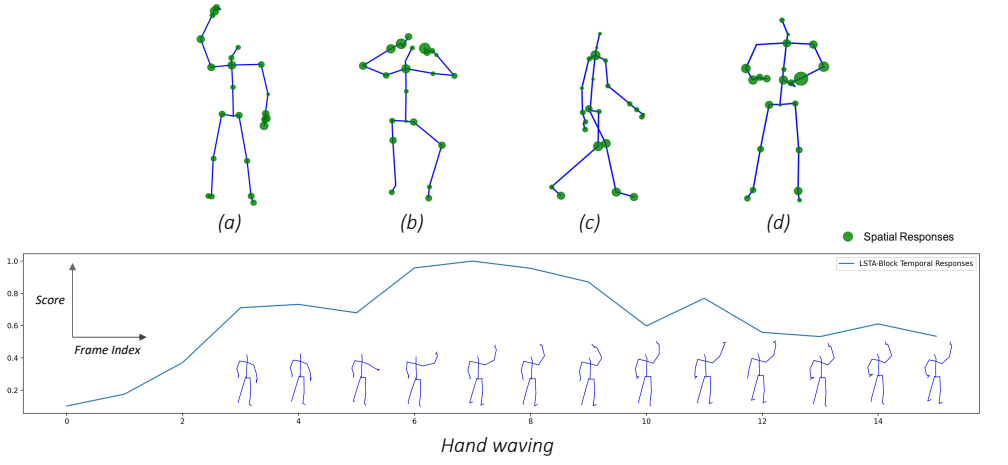

Figure 3: TOP: Examples of the joint feature responses for four actions (a) "walking" (b) "put on a hat" (c)"hand waving" (d) "type on a keyboard". The size of green circles indicates the importance of the joint. BOTTOM: Visualisation of the temporal feature responses for each of the frame for action "hand waving". X-axis denotes the input skeleton frame index, and Y-axis indicates the importance of each frame (scaled to range $[0,1]$). .

both hands are well focused for actions "type on a keyboard", "put on a hat"; the model focuses more on arm parts for action "hand waving"; for "walking" action, the model focuses on the lower body, especially the feet and knees.

For the temporal modelling, we show an example of the learned feature responses for several frames and the corresponding skeleton sketches are in Fig.3 (BOTTOM). For action "hand waving", we can see the model focuses more on the process of raising hand in temporal domain and also pays attention on localised motion patterns in the spatial domain (i.e., "the raised hand"), suggesting the capability of our model in capturing the spatio-temporal dependencies in skeleton-based action recognition.

4 Conclusion

In this work, we propose the light-weight LSTA-Net, which can alternately performs long short-term spatio-temporal feature aggregation for improved skeleton-based action recognition. Specifically, MSDA for capturing distant spatial information, and ATPA for capturing long-range temporal information are proposed, with MAM employed for further performance gain. On three large-scale public datasets, despite smaller model size, our LSTA-Net achieves higher accuracies than most of other state-of-the-arts, suggesting it is a practical solution for skeleton-based action recognition.

5 Acknowledgment

This research is jointly funded by EPSRC Centre for Doctoral Training in Digital Civics (EP/L016176/1) and EPSRC DERC: Digital Economy Research Centre (EP/M023001/1).

References

[1] Jake K Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 2014.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[3] Tailin Chen, Desen Zhou, Jian Wang, Shidong Wang, Yu Guan, Xuming He, and Errui Ding. Learning multi-granular spatio-temporal graph network for skeleton-based action recognition. In *ACMMM*, 2021.

[4] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *ECCV*, 2020.

[5] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*, 2020.

[6] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015.

[7] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *TPAMI*, 2019.

[8] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[10] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017.

[11] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *CVPRW*, 2017.

[12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *AAAI*, 2019.

[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 *CVPR*, 2019.

[15] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, 2016.

[16] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 2019.

[17] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020.

[18] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *in AAAI*, 2020.

[19] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[20] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. *TIP*, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01230.

[21] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *CVPR*, 2019.

[22] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.

[23] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *CVPR*, 2019.

[24] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, Faster and More Explainable: A Graph Convolutional Baseline for Skeleton-based Action Recognition. 2020. doi: 10.1145/3394171.3413802.

[25] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, 2020.

[26] Yu-Hui Wen, Lin Gao, Hongbo Fu, Fang-Lue Zhang, and Shihong Xia. Graph cnns with motif and variable temporal block for skeleton-based action recognition. In *AAAI*, 2019.

[27] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.

[28] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

[29] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *ICCV*, 2017.

[30] Zhengyou Zhang. Microsoft kinect sensor and its effect. *ACMMM*, 2012.