Decoupled Spatial-Temporal Attention Network for Skeleton-Based Action-Gesture Recognition

Lei Shi[1,2], Yifan Zhang[1,2], Jian Cheng[1,2,3] and Hanqing Lu[1,2]

[1] NLPR & AIRIA, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
{lei.shi, yfzhang, jcheng, luhq}@nlpr.ia.ac.cn

Abstract. Dynamic skeletal data, represented as the 2D/3D coordinates of human joints, has been widely studied for human action recognition due to its high-level semantic information and environmental robustness. However, previous methods heavily rely on designing handcrafted traversal rules or graph topologies to draw dependencies between the joints, which are limited in performance and generalizability. In this work, we present a novel decoupled spatial-temporal attention network (DSTA-Net) for skeleton-based action recognition. It involves solely the attention blocks, allowing for modeling spatial-temporal dependencies between joints without the requirement of knowing their positions or mutual connections. Specifically, to meet the specific requirements of the skeletal data, three techniques are proposed for building attention blocks, namely, spatial-temporal attention decoupling, decoupled position encoding and spatial global regularization. Besides, from the data aspect, we introduce a skeletal data decoupling technique to emphasize the specific characteristics of space/time and different motion scales, resulting in a more comprehensive understanding of the human actions. To test the effectiveness of the proposed method, extensive experiments are conducted on four challenging datasets for skeleton-based gesture and action recognition, namely, SHREC, DHG, NTU-60 and NTU-120, where DSTA-Net achieves state-of-the-art performance on all of them.

1 Introduction

Human action recognition has been studied for decades since it can be widely used for many applications such as human-computer interaction and abnormal behavior monitoring [1–4]. Recently, skeletal data draws increasingly more attention because it contains higher-level semantic information in a small amount of data and has strong adaptability to the dynamic circumstance [5–7].

The raw skeletal data is a sequence of frames each contains a set of points. Each point represents a joint of human body in the form of 2D/3D coordinates. Previous data-driven methods for skeleton-based action recognition rely on manual designs of traversal rules or graph topologies to transform the raw skeletal data into a meaningful form such as a point-sequence, a pseudo-image

or a graph, so that they can be fed into the deep networks such as RNNs, CNNs and GCNs for feature extraction [5, 8, 9]. However, there is no guarantee that the hand-crafted rule is the optimal choice of modeling global dependencies of joints, which limits the performance and generalizability of previous approaches. Recently, transformer [10, 11] has achieved big success in the NLP field, whose basic block is the self-attention mechanism. It can learn the global dependencies between the input elements with less computational complexity and better parallelizability. For skeletal data, employing the self-attention mechanism has an additional advantage that there is no requirement of knowing a intrinsic relations between the elements, thus it provides more flexibility for discovering useful patterns. Besides, since the number of joints of the human body is limited, the extra cost of applying self-attention mechanism is also relatively small.

Inspired by above observations, we propose a novel decoupled spatial-temporal attention networks (DSTA-Net) for skeleton-based action recognition. It is based solely on the self-attention mechanism, without using the structure-relevant RNNs, CNNs or GCNs. However, it is not straightforward to apply a pure attention network for skeletal data as shown in following three aspects: (1) The input of original self-attention mechanism is the sequential data, while the skeletal data exists in both the spatial and temporal dimensions. A naive method is simply flattening the spatial-temporal data into a single sequence like [12]. However, it is not reasonable to treat the time and space equivalently because they contain totally different semantics [3]. Besides, simple flattening operation increases the sequence length, which greatly increases the computation cost due to the dot-product operation of the self-attention mechanism. Instead, we propose to decouple the self-attention mechanism into the spatial attention and the temporal attention sequentially. Three strategies are specially designed to balance the independence and the interaction between the space and the time. (2) There are no predefined orders or structures when feeding the skeletal joints into the attention networks. To provide unique markers for every joint, a position encoding technique is introduced. For the same reason as before, it is also decoupled into the spatial encoding and the temporal encoding. (3) It has been verified that adding proper regularization based on prior knowledge can effectively reduce the over-fitting problem and improve the model generalizability. For example, due to the translation-invariant structure of images, CNNs exploit the local-weight-sharing mechanism to force the model to learn more general filters for different regions of images. As for skeletal data, each joint of the skeletons has specific physical/semantic meaning (e.g., head or hand), which is fixed for all the frames and is consistent for all the data samples. Based on this prior knowledge, a spatial global regularization is proposed to force the model to learn more general attentions for different samples. Note the regularization is not suitable for temporal dimension because there is no such semantic alignment property.

Besides, from the data aspect, the most discriminative pattern is distinct for different actions. We claim that two properties should be considered. One property is whether the action is motion relevant or motion irrelevant, which aims to choose the specific characters of space and time. For example, to classify the

gestures of "waving up" versus "waving down", the global trajectories of hand is more important than hand shape, but when recognizing the gestures like "point with one finger" versus "point with two finger", the spatial pattern is more important than hand motion. Based on this observation, we propose to decouple the data into the spatial and temporal dimensions, where the spatial stream contains only the motion-irrelevant features and temporal stream contains only the motion-relevant features. By modeling these two streams separately, the model can better focus on spatial/temporal features and identity specific patterns. Finally, by fusing the two streams, it can obtain a more comprehensive understanding of the human actions. Another property is the sensibility of the motion scales. For temporal stream, the classification of some actions may rely on the motion mode of a few consecutive frames while others may rely on the overall movement trend. For example, to classify the gestures of "clapping" versus "put two hands together", the short-term motion detail is essential. But for "waving up" versus "waving down", the long-term motion trend is more important. Thus, inspired by [3], we split the temporal information into a fast stream and a slow stream based on the sampling rate. The low-frame-rate stream can capture more about global motion information and the high-frame-rate stream can focus more on the detailed movements. Similarly, the two streams are fused to improve the recognition performance.

We conduct extensive experiments on four datasets, including two hand gesture recognition datasets, i.e., SHREC and DHG, and two human action recognition datasets, i.e., NTU-60 and NTU-120. Without the need of hand-crafted traversal rules or graph topologies, our method achieves state-of-the-art performance on all these datasets, which demonstrates the effectiveness and generalizability of the proposed method.

Overall, our contributions lie in four aspects: (1) To the best of our knowledge, we are the first to propose a decoupled spatial-temporal attention networks (DSTA-Net) for skeleton-based action recognition, which is built with pure attention modules without manual designs of traversal rules or graph topologies. (2) We propose three effective techniques in building attention networks to meet the specific requirements for skeletal data, namely, spatial-temporal attention decoupling, decoupled position encoding and spatial global regularization. (3) We propose to decouple the data into four streams, namely, spatial-temporal stream, spatial stream, slow-temporal stream and fast-temporal stream, each focuses on a specific aspect of the skeleton sequence. By fusing different types of features, the model can have a more comprehensive understanding for human actions. (4) On four challenging datasets for action recognition, our method achieves state-of-the-art performance with a significant margin. DSTA-Net outperforms SOTA 2.6%/3.2% and 1.9%/2.9% on 14-class/28-class benchmarks of SHREC and DHG, respectively. It achieves 91.5%/96.4% and 86.6%/89.0% on CS/CV benchmarks of NTU-60 and NTU-120, respectively. The code is released[4].

[4] https://github.com/lshiwjx/DSTA-Net

2 Related Work

Skeleton-based action recognition has been widely studied for decades. The main-stream methods lie in three branches: (1) the RNN-based methods that formulate the skeletal data as a sequential data with a predefined traversal rules, and feed it into the RNN-based models such as the LSTM [9, 13–15]; (2) We propose three effective techniques in building attention networks to meet the specific requirements for skeletal data, namely, spatial-temporal attention decoupling, decoupled position encoding and spatial global regularization. (3) the GCN-based methods that encode the skeletal data into a predefined spatial-temporal graph, and model it with the graph convolutional networks [5, 16, 6]. In this work, instead of formulating the skeletal data into the images or graphs, we directly model the dependencies of joints with pure attention blocks. Our model is more concise and general, without the need of designing hand-crafted transformation rules, and it outperforms the previous methods with a significant margin.

Self-attention mechanism is the basic block of transformer [10, 11], which is the mainstream method in the NLP field. Its input consists of a set of queries Q, keys K of dimension C and values V, which are packaged in the matrix form for fast computation. It first computes the dot products of the query with all keys, divides each by $\sqrt{C}$, and applies a softmax function to obtain the weights on the values [10]. In formulation:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{C}}) \tag{1}$$

The similar idea has also been used for many computer vision tasks such as relation modeling [17], detection [18] and semantic segmentation [19]. To the best of our knowledge, we are the fist to apply the pure attention networks for skeletal data and further propose several improvements to meet the specific requirements of skeletons.

3 Methods

3.1 Spatial-temporal attention module

Original transformer is fed with the sequential data, i.e., a matrix $X \in \mathbb{R}^{N \times C}$, where N denotes the number of elements and C denotes the number of channels. For dynamic skeletal data, the input is a 3-order tensor $X \in \mathbb{R}^{N \times T \times C}$, where T denotes the number of frames. It is worth to investigate how to deal with the relationship between the time and the space. Wang et al. [12] propose to ignore the difference between time and space, and regard the inputs as a sequential data $X \in \mathbb{R}^{\hat{N} \times C}$, where $\hat{N} = NT$. However, the temporal dimension and the spatial dimension are totally different as introduced in Sec. 1. It is not reasonable to treat them equivalently. Besides, the computational complexity of calculating

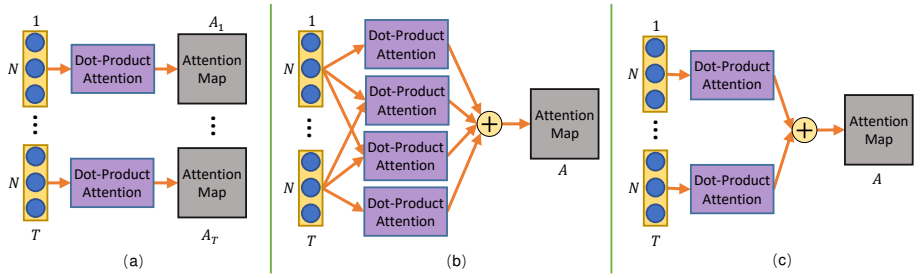

Fig. 1. Illustration of the three decoupling strategies. We use the spatial attention strategy as an example and the temporal attention strategy is an analogy. N and T denote the number of joints and frames, respectively.

the attention map in this strategy is $O(T^2N^2C)$ (using the naive matrix multiplication algorithm), which is too large. Instead, we propose to decouple the spatial and temporal dimensions, where the computational complexity is largely reduced and the performance is improved.

We design three strategies for decoupling as shown in Fig 1. Using the spatial attention as an example, the first strategy (Fig 1, a) is calculating the attention maps frame by frame, and each frame uses a unique attention map:

$$A^t = softmax(\sigma(X_t)\phi(X_t)') \tag{2}$$

where $A^t \in \mathbb{R}^{N\times N}$ is the attention map for frame t. $X_t \in \mathbb{R}^{N\times C}$. σ and ϕ are two embedding functions. $'$ denote matrix transpose. This strategy only considers the dependencies of joints in a single frame thus lacks the modeling capacity. The computational complexity of calculating spatial attention of this strategy is $O(TN^2C)$. For temporal attention, the attention map of joint n is $A^n \in \mathbb{R}^{T\times T}$ and the input data is $X_n \in \mathbb{R}^{T\times C}$. Its calculation is analogical with the spatial attention. Considering both the spatial and temporal attention, the computational complexity of the first strategy for all frames is $O(TN^2C + NT^2C)$.

The second strategy (Fig 1, b) is calculating the relations of two joints between all of the frames, which means both the intra-frame relations and the inter-frame relations of two joints are taken into account simultaneously. The attention map is shared over all frames. In formulation:

$$A^t = softmax(\sum_{t}^{T}\sum_{\tau}^{T}(\sigma(X_t)\phi(X_\tau)')) \tag{3}$$

The computational complexity of this strategy is $O(T^2N^2C + N^2T^2C)$.

The third strategy (Fig 1, c) is a compromise, where only the joints in same frame are considered to calculate the attention map, but the obtained attention maps of all frames are averaged and shared. It is equivalent to adding a time consistency restriction for attention computation, which can somewhat reduce

the overfitting problem caused by the element-wise relation modeling of the second strategy.

$$A^t = softmax(\sum_t^T (\sigma(X_t)\phi(X_t)')) \tag{4}$$

By concatenating the frames into an $N \times TC$ matrix, the summation of mat-multiplications can be efficiently implemented with one big mat-multiplication operation. The computational complexity of this strategy is $O(TN^2C + NT^2C)$. as shown in ablation study 4.3, we finally use the strategy (c) in the model.

3.2 Decoupled Position encoding

The skeletal joints are organized as a tensor to be fed into the neural networks. Because there are no predefined orders or structures for each element of the tensor to show its identity (e.g., joint index or frame index), we need a position encoding module to provide unique markers for every joint. Following [10], we use the sine and cosine functions with different frequencies as the encoding functions:

$$\begin{aligned} PE(p, 2i) &= sin(p/10000^{2i/C_{in}}) \\ PE(p, 2i+1) &= cos(p/10000^{2i/C_{in}}) \end{aligned} \tag{5}$$

where p denotes the position of element and i denotes the dimension of the position encoding vector. However, different with [10], the input of skeletal data have two dimensions, i.e., space and time. One strategy for position encoding is unifying the spatial and temporal dimensions and encoding them sequentially. For example assuming there are three joints, for the first frame the position of joints is $1, 2, 3$, and for the second frame it is $4, 5, 6$. This strategy cannot well distinguish the same joint in different frames. Another strategy is decoupling the process into spatial position encoding and temporal position encoding. Using the spatial position encoding as an example, the joints in the same frame are encoded sequentially and the same joints in different frames have the same encoding. In above examples, it means for the first frame the position is $1, 2, 3$, and for the second frame it is also $1, 2, 3$. As for the temporal position encoding, it is reversed and analogical, which means the joints in the same frame have the same encoding and the same joints in different frames are encoded sequentially. Finally, the position features are added to the input data as shown in Fig 2. In this way, each element is aligned with an unique marker to help learning the mutual relations between the joints, and the difference between space and time is also well expressed.

3.3 Spatial global regularization

As explained in Sec. 1, each joint has a specific meaning. Based on this prior knowledge, we propose to add a spatial global regularization to force the model to learn more general attentions for different samples. In detail, a global attention map ($N \times N$ matrix) is added to the attention map ($N \times N$ matrix) learned by the

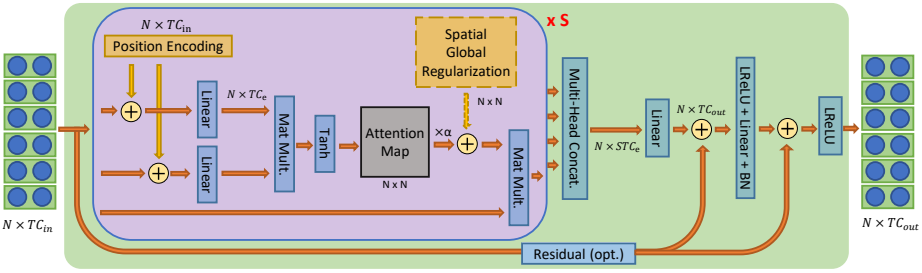

Fig. 2. Illustration of the attention module. We show the spatial attention module as an example. The temporal attention module is an analogy. The purple rounded rectangle box represents a single-head self-attention module. There are totally S self-attention modules, whose output are concatenated and fed into two linear layers to obtain the output. LReLU represents the leaky ReLU [20].

dot-product attention mechanism introduced in Sec. 3.1. The global attention map is shared for all data samples, which represents a unified intrinsic relationship pattern of the human joints. We set it as the parameter of the network and optimize it together with the model. An α is multiplied to balance the strength of the spatial global regularization. This module is simple and light-weight, but it is effective as shown in the ablation study. Note that the regularization is only added for spatial attention computing because the temporal dimension has no such semantic alignment property. Forcing a global regularization for temporal attention is not reasonable and will harm the performance.

3.4 Complete attention module

Because the spatial attention module and the temporal attention module are analogical, we select the spatial module as an example for detailed introduction. The complete attention module is showed in Fig 2. The procedures inside the purple rounded rectangle box illustrate the process of the single-head attention calculation. The input $X \in \mathbb{R}^{N \times TC_{in}}$ is first added with the spatial position encoding. Then it is embedded with two linear mapping functions to $X \in \mathbb{R}^{N \times TC_e}$. C_e is usually small than C_{out} to remove the feature redundancy and reduce the computations. The attention map is calculated by the strategy (c) of Fig. 1 and added with the spatial global regularization. Note that we found the Tanh is better than SoftMax when computing the attention map. We believe that it is because the output of Tanh is not restricted to positive values thus can generate negative relations and provide more flexibility. Finally the attention map is mat-multiplied with the original input to get the output features.

To allow the model jointly attending to information from different representation sub-spaces, there are totally S heads for attention calculations in the module. The results of all heads are concatenated and mapped to the output space $\mathbb{R}^{N \times TC_{out}}$ with a linear layer. Similar with the transformer, a point-wise feed-forward layer is added in the end to obtain the final output. We use the

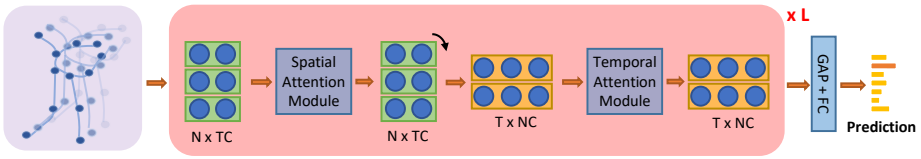

Fig. 3. Illustration of the overall architecture of the DSTA-Net. N, T, C denote the number of joints, frames and channels, respectively. The red rounded rectangle box represents one spatial-temporal attention layer. There are totally L layers. The final output features are global-average-pooled (GAP) and fed into a fully-conected layer (FC) to make the prediction.

leaky ReLU as the non-linear function. There are two residual connections in the module as shown in the Fig 2 to stabilize the network training and integrate different features. Finally, all of the procedures inside the green rounded rectangle box represent one whole attention module.

3.5 Overall architecture

Fig. 3 shows the overall architecture of our method. The input is a skeleton sequence with N joints, T frames and C channels. In each layer, we first regard the input as an $N \times TC$ matrix, i.e., N elements with TC channels, and feed it into the spatial attention module (introduced in Fig. 2) to model the spatial relations between the joints. Then, we transpose the output matrix and regard it as T elements each has NC channels, and feed it into the temporal attention module to model the temporal relations between the frames. There are totally L layers stacked to update features. The final output features are global-average-pooled and fed into a fully-connected layers to obtain the classification scores.

3.6 Data decoupling

The action can be decoupled into two dimensions: the spatial dimension and the temporal dimension as illustrated in Fig. 4 (a, b and c). The spatial information is the difference of two different joints that are in the same frame, which mainly contains the relative position relationship between different joints. To reduce the redundant information, we only calculate the spatial information along the human bones. The temporal information is the difference of the two joints with same spatial meaning in different frames, which mainly describes the motion trajectory of one joint along the temporal dimension. When we recognize the gestures like “Point with one finger” versus “Point with two finger”, the spatial information is more important. However, when we recognize the gestures like “waving up” versus “waving down”, the temporal information will be more essential.

Besides, for temporal stream, different actions have different sensibilities of the motion scale. For some actions such as “clapping” versus “put two hands

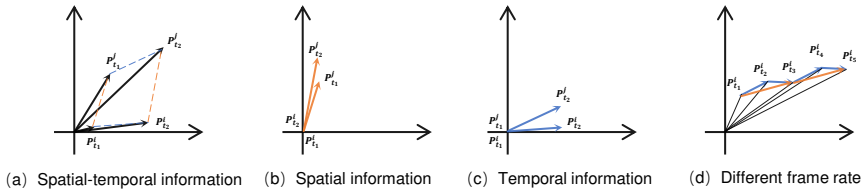

Fig. 4. For simplicity, we draw two joins in two consecutive frames in a 2D coordinate system to illustrate the data decoupling. as shown in (a), $P^i_{t_k}$ denotes the joint i in frame k. Assume that joint i and joint j are the two end joints of one bone. (a) denotes the raw data, i.e., the spatial-temporal information. The orange dotted line and blue dotted line denote the decoupled spatial information and temporal information, which are showed as (b) and (c), respectively. (d) illustrates the difference between the fast-temporal information (blue arrow) and the slow-temporal information (orange arrow).

together", the short-term motion detail is essential. But for actions like "waving up" versus "waving down", the long-term movement trend is more important. Inspired by [3], we propose to calculate the temporal motion with both the high frame-rate sampling and the low frame-rate sampling as shown in Fig. 4 (d). The generated two streams are called as the fast-temporal stream and the slow-temporal stream, respectively.

Finally, we have four streams all together, namely, spatial-temporal stream (original data), spatial stream, fast-temporal stream and slow-temporal stream. We separately train four models with the same architecture for each of the streams. The classification scores are averaged to obtain the final result.

4 Experiments

To verify the generalization of the model, we use two datasets for hand gesture recognition (DHG [21] and SHREC [22]) and two datasets for human action recognition (NTU-60 [23] and NTU-120 [24]). We first perform exhaustive ablation studies on SHREC to verify the effectiveness of the proposed model components. Then, we evaluate our model on all four datasets to compare with the state-of-the-art methods.

4.1 Datasets

DHG: DHG [21] dataset contains 2800 video sequences of 14 hand gestures performed 5 times by 20 subjects. They are performed in two ways: using one finger and the whole hand. So it has two benchmarks: 14-gestures for coarse classification and 28-gestures for fine-grained classification. The 3D coordinates

of 22 hand joints in real-world space is captured by the Intel Real-sense camera. It uses the leave-one-subject-out cross-validation strategy for evaluation.

SHREC: SHREC [22] dataset contains 2800 gesture sequences performed 1 and 10 times by 28 participants in two ways like the DHG dataset. It splits the sequences into 1960 train sequences and 840 test sequences. The length of sample gestures ranges from 20 to 50 frames. This dataset is used for the competition of SHREC'17 in conjunction with the Euro-graphics 3DOR'2017 Workshop.

NTU-60: NTU-60 [23] is a most widely used in-door-captured action recognition dataset, which contains 56,000 action clips in 60 action classes. The clips are performed by 40 volunteers and is captured by 3 KinectV2 cameras with different views. This dataset provides 25 joints for each subject in the skeleton sequences. It recommends two benchmarks: cross-subject (CS) and cross-view (CV), where the subjects and cameras used in the training/test splits are different, respectively.

NTU-120: NTU-120 [23] is similar with NTU-60 but is larger. It contains 114,480 action clips in 120 action classes. The clips are performed by 106 volunteers in 32 camera setups. It recommends two benchmarks: cross-subject (CS) and cross-setup (CE), where cross-setup means using samples with odd setup IDs for training and others for testing.

4.2 Training details

To show the generalization of our methods, we use the same configuration for all experiments. The network is stacked using 8 DSTA blocks with 3 heads. The output channels are 64, 64, 128, 128, 256, 256, 256 and 256, respectively. The input video is randomly/uniformly sampled to 150 frames and then randomly/centrally cropped to 128 frames for training/test splits. For fast-temporal features, the sampling interval is 2. When training, the initial learning rate is 0.1 and is divided by 10 in 60 and 90 epochs. The training is ended in 120 epochs. Batch size is 32. Weight decay is 0.0005. We use the stochastic gradient descent (SGD) with Nesterov momentum (0.9) as the optimizer and the cross-entropy as the loss function.

4.3 Ablation studies

In this section, we investigate the effectiveness of the proposed components of the network and different data modalities. We conduct experiments on SHREC dataset. Except for the explored object, other details are set the same for fair comparison.

Network architectures We first investigate the effect of the position embedding. as shown in Tab. 1, removing the position encoding will seriously harm the performance. Decoupling the spatial and temporal dimension (DPE) is better than not (UPE). This is because the spatial and temporal dimensions actually have different properties and treat them equivalently will confuse the model.

Then we investigate the effect of the proposed spatial global regularization (SGR). By adding the SGR, the performance is improved from 94.3% to 96.3%, but if we meanwhile regularize the temporal dimension, the performance drops. This is reasonable since there are no specific meanings for temporal dimension and forced learning of a unified pattern will cause the gap between the training set an testing set.

Finally, we compare the three strategies introduced in Fig. 1. It shows that the strategy (a) obtains the lowest performance. We conjecture that it dues to the fact that it only considers the intra-frame relations and ignore the inter-frame relations. Modeling the inter-frame relations exhaustively (strategy b) will improve the performance and a compromise (c) obtains the best performance. It may because that the compromise strategy can somewhat reduce the overfitting problem.

Table 1. Ablation studies for architectures of the model on the SHREC dataset. ST-ATT-c denotes the **s**patial **t**emporal **att**ention networks with attention type **c** introduced in Fig 1. PE denotes **p**osition **e**ncoding. UPE/DPE denote using **u**nified/**d**ecoupled encoding for spatial and temporal dimensions. STGR denotes **s**patial-**t**emporal **g**lobal **r**egularizations for computing attention maps.

Method	Accuracy
ST-Att-c w/o PE	89.4
ST-Att-c + UPE	93.2
ST-Att-c + DPE	94.5
ST-Att-c + DPE + SGR	**96.3**
ST-Att-c + DPE + STGR	94.6
ST-Att-a + DPE + SGR	94.6
ST-Att-b + DPE + SGR	95.1

We show the learned attention maps of different layers (layer #1 and layer #8) in Fig. 5. Other layers are showed in supplement materials. It shows that the attention maps learned in different layers are not the same because the information contained in different layers has distinct semantics. Besides, it seems the model focuses more on the relations between the tips of the fingers (T4, I4, M4, R4) and wrist, especially in the lower layers. This is intuitive since these joints are more discriminative for human to recognize gestures. On the higher layers, the information are highly aggregated and the difference between each of the joints becomes unapparent, thus the phenomenon also becomes unapparent.

Data decoupling To show the necessity of decoupling the raw data into four streams as introduced in Sec. 3.6, we show the results of using four streams separately and the result of fusion in Tab. 2. It shows that the accuracies of decoupled streams are not as good as the raw data because some of the information is lost. However, since the four streams focus on different aspects and are

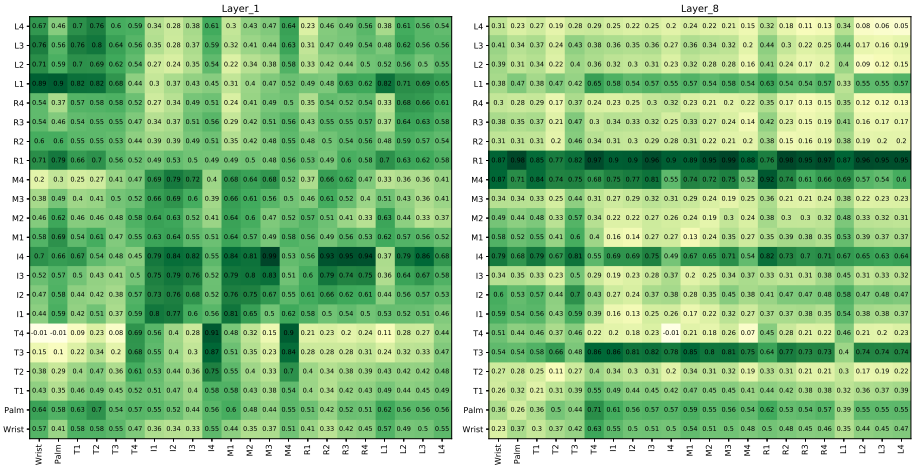

Fig. 5. Examples of the learned attention maps for different layers. T, I, M, R and L denote thumb, index finger, middle finger, ring finger and little finger, respectively. As for articulation, T1 denotes the base of the thumb and T4 denote the tip of the thumb.

complementary with each other, when fusing them together, the performance is improved significantly.

Table 2. Ablation studies for feature fusion on the SHREC dataset. Spatial-temporal denotes the raw data, i.e., the joint coordinates. Other types of features are introduced in Sec. 3.6.

Method	Accuracy
spatial-temporal	96.3
spatial	95.1
fast-temporal	94.5
slow-temporal	93.7
Fusion	**97.0**

As shown in Fig. 6, We plot the per-class accuracies of the four streams to show the complementarity clearly. We also plot the difference of accuracies between different streams, which are represented as the dotted lines. For spatial information versus temporal information, it (orange dotted lines) shows that the network with spatial information obtains higher accuracies mainly in classes that are closely related with the shape changes such as “grab”, “expand” and “pinch”, and the network with temporal information obtains higher accuracies mainly in classes that are closely related with the positional changes such as “swipe”, “rot” and “shake”. As for different frame-rate sampling, it (red dotted lines) shows that the slow-temporal performs better for classes of “expand”,

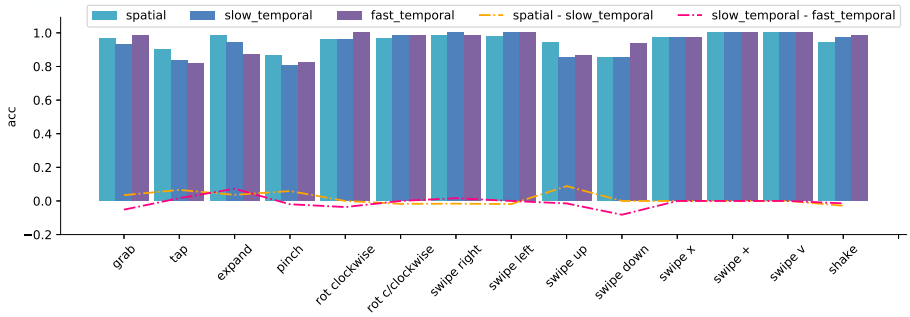

Fig. 6. Per-class accuracies for different modalities on SHREC-14 dataset. The dotted lines shows the difference between two modalities.

"tap", etc, and the fast-temporal performs better for classes of "swipe", "rot", etc. These phenomenons verify the complementarity of the four modalities.

Table 3. Recognition accuracy comparison of our method and state-of-the-art methods on SHREC dataset and DHG dataset.

Method	Year	SHREC		DHG	
		14 gestures	28 gestures	14 gestures	28 gestures
ST-GCN [5]	2018	92.7	87.7	91.2	87.1
STA-Res-TCN [25]	2018	93.6	90.7	89.2	85.0
ST-TS-HGR-NET [26]	2019	94.3	89.4	87.3	83.4
DG-STA. [27]	2019	94.4	90.7	91.9	88.0
DSTA-Net(ours)	-	**97.0**	**93.9**	**93.8**	**90.9**

4.4 Comparison with previous methods

We evaluate our model with state-of-the-art methods for skeleton-based action recognition on all four datasets, where our model significantly outperforms the other methods. Due to the space restriction, we only show some representative works, where more comparisons are showed in supplement materials. On SHREC/DHG datasets for skeleton-based hand gestures recognition (Tab. 3), our model brings 2.6%/1.9% and 3.2%/2.9% improvements for 14-gestures and 28-gestures benchmarks compared with the state-of-the-arts. Note that the state-of-the-art accuracies are already very high (94.4%/91.9% and 90.7%/88.0% for 14-gestures and 28-gestures, respectively), but our model still obtains remarkable performance. On NTU-60 dataset (Tab. 4), our model obtains 1.6% and 0.3% improvements. The performance of CV benchmark is nearly saturated. For both CS and CV benchmarks, we visualize the wrong examples and find that it is even impossible for human to recognize many examples using only the skeletal

Table 4. Recognition accuracy comparison of our method and state-of-the-art methods on NTU-60 dataset. CS and CV denote the cross-subject and cross-view benchmarks, respectively.

Methods	Year	CS (%)	CV (%)
ST-GCN [5]	2018	81.5	88.3
SRN+TSL [14]	2018	84.8	92.4
2s-AGCN [6]	2019	88.5	95.1
DGNN [7]	2019	89.9	96.1
NAS [28]	2020	89.4	95.7
DSTA-Net(ours)	-	**91.5**	**96.4**

Table 5. Recognition accuracy comparison of our method and state-of-the-art methods on NTU-120 dataset. CS and CE denote the cross-subject and cross-setup benchmarks, respectively.

Methods	Year	CS (%)	CE (%)
Body Pose Evolution Map [29]	2018	64.6	66.9
SkeletonMotion [30]	2019	67.7	66.9
DSTA-Net(ours)	-	**86.6**	**89.0**

data. For example, for the two classes of reading and writing, the humans are both in a same posture (standing or sitting) and holding a book. The only difference is whether there is a pen in the hand, which cannot be captured through the skeletal data. On NTU-120 dataset (Tab. 5), our model also achieves state-of-the-art performance. Since this dataset is released recently, our method can provide a new baseline on it.

5 Conclusion

In this paper, we propose a novel decoupled spatial-temporal attention network (DSTA-Net) for skeleton-based action recognition. It is a unified framework based solely on attention mechanism, with no needs of designing hand-crafted traversal rules or graph topologies. We propose three techniques in building DSTA-Net to meet the specific requirements for skeletal data, including spatial-temporal attention decoupling, decoupled position encoding and spatial global regularization. Besides, we introduce a skeleton-decoupling method to emphasize the spatial/temporal variations and motion scales of the skeletal data, resulting in a more comprehensive understanding for human actions and gestures.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 61872364 and 61876182 , in part by the Jiangsu Leading Technology Basic Research Project BK20192004.

References

1. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 6299–6308
2. Shi, L., Zhang, Y., Jian, C., Hanqing, L.: Gesture Recognition using Spatiotemporal Deformable Convolutional Representation. In: IEEE International Conference on Image Processing (ICIP). (2019)
3. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 6202–6211
4. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Action Recognition via Pose-Based Graph Convolutional Networks with Intermediate Dense Supervision. arXiv:1911.12509 (2019)
5. Yan, S., Xiong, Y., Lin, D.: Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In: AAAI. (2018)
6. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 12026–12035
7. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-Based Action Recognition With Directed Graph Neural Networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 7912–7921
8. Qiu, Z., Yao, T., Mei, T.: Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In: The IEEE International Conference on Computer Vision (ICCV). (2017) 5533–5541
9. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition From Skeleton Data. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 2117–2126
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems. (2017) 6000–6010
11. Dai, Z., Yang, Z., Yang, Y., Cohen, W.W., Carbonell, J., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. arXiv:1901.02860 (2019)
12. Wang, X., Girshick, R., Gupta, A., He, K.: Non-Local Neural Networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
13. Li, S., Li, W., Cook, C., Zhu, C., Gao, Y.: Independently recurrent neural network (indrnn): Building A longer and deeper RNN. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 5457–5466
14. Si, C., Jing, Y., Wang, W., Wang, L., Tan, T.: Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning. In: The European Conference on Computer Vision (ECCV). (2018) 103–118
15. Si, C., Chen, W., Wang, W., Wang, L., Tan, T.: An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 1227–1236
16. Tang, Y., Tian, Y., Lu, J., Li, P., Zhou, J.: Deep Progressive Reinforcement Learning for Skeleton-Based Action Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 5323–5332

17. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. In Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., eds.: Advances in Neural Information Processing Systems. (2017) 4974–4983
18. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation Networks for Object Detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
19. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3146–3154
20. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: ICML. Volume 30. (2013) 3
21. De Smedt, Q., Wannous, H., Vandeborre, J.P.: Skeleton-Based Dynamic Hand Gesture Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (2016) 1206–1214
22. De Smedt, Q., Wannous, H., Vandeborre, J.P., Guerry, J., Le Saux, B., Filliat, D.: SHREC'17 Track: 3D Hand Gesture Recognition Using a Depth and Skeletal Dataset. In Pratikakis, I., Dupont, F., Ovsjanikov, M., eds.: Eurographics Workshop on 3D Object Retrieval. (2017) 1–6
23. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 1010–1019
24. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 1–1
25. Hou, J., Wang, G., Chen, X., Xue, J.H., Zhu, R., Yang, H.: Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition. In: The European Conference on Computer Vision (ECCV). (2018) 0–0
26. Nguyen, X.S., Brun, L., Lézoray, O., Bougleux, S.: A neural network based on SPD manifold learning for skeleton-based hand gesture recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019)
27. Chen, Y., Zhao, L., Peng, X., Yuan, J., Metaxas, D.N.: Construct Dynamic Graphs for Hand Gesture Recognition via Spatial-Temporal Attention. In: BMVC. (2019)
28. Peng, W., Hong, X., Chen, H., Zhao, G.: Learning Graph Convolutional Network for Skeleton-based Human Action Recognition by Neural Searching. In: AAAI. (2020)
29. Liu, M., Yuan, J.: Recognizing Human Actions as the Evolution of Pose Estimation Maps. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 1159–1168
30. Caetano, C., Sena, J., Brémond, F., Dos Santos, J.A., Schwartz, W.R.: 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). (2019) 1–8