Spatial-Temporal Attention Res-TCN for Skeleton-based Dynamic Hand Gesture Recognition

Jingxuan Hou[1], Guijin Wang[1⋆], Xinghao Chen[1], Jing-Hao Xue[2], Rui Zhu[3], and Huazhong Yang[1]

[1] Tsinghua University, Beijing, China
{houjx14,chen-xh13}@mails.tsinghua.edu.cn
{wangguijin,yanghz}@tsinghua.edu.cn
[2] University College London, London, UK
jinghao.xue@ucl.ac.uk
[3] University of Kent, Kent, UK
R.Zhu@kent.ac.uk

Abstract. Dynamic hand gesture recognition is a crucial yet challenging task in computer vision. The key of this task lies in an effective extraction of discriminative spatial and temporal features to model the evolutions of different gestures. In this paper, we propose an end-to-end Spatial-Temporal Attention Residual Temporal Convolutional Network (STA-Res-TCN) for skeleton-based dynamic hand gesture recognition, which learns different levels of attention and assigns them to each spatial-temporal feature extracted by the convolution filters at each time step. The proposed attention branch assists the networks to adaptively focus on the informative time frames and features while exclude the irrelevant ones that often bring in unnecessary noise. Moreover, our proposed STA-Res-TCN is a lightweight model that can be trained and tested in an extremely short time. Experiments on DHG-14/28 Dataset and SHREC'17 Track Dataset show that STA-Res-TCN outperforms state-of-the-art methods on both the 14 gestures setting and the more complicated 28 gestures setting.

Keywords: dynamic hand gesture recognition, spatial-temporal attention, temporal convolutional networks

1 Introduction

Dynamic hand gesture recognition has attracted increasing interests due to its potential relevance to a wide range of applications, such as touchless automotive user interfaces, gaming, robotics, etc [21, 3, 28]. However, it is still challenging to develop a highly precise hand gesture recognition system, owing to high intra-class variance derived from the various possibilities to perform the same gesture [30, 5, 3].

⋆ Corresponding Author

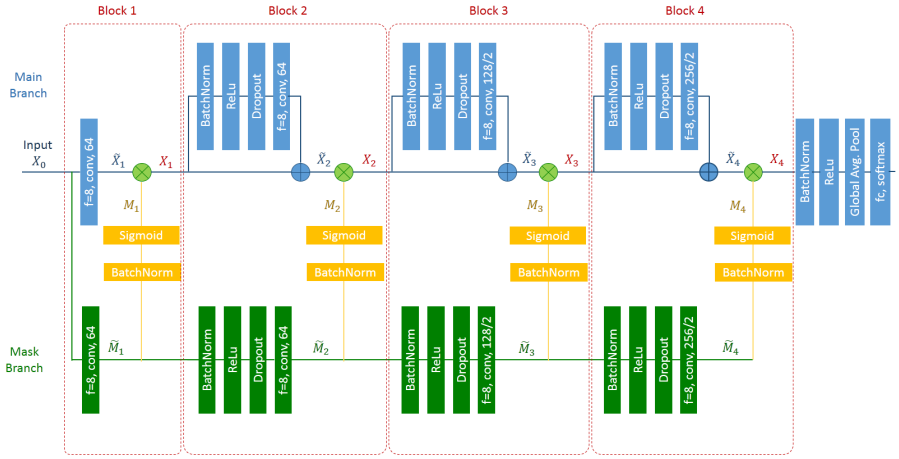

Fig. 1. Overall architecture of the proposed STA-Res-TCN, which consists of a main branch for feature processing, a mask branch for attention and an element-wise multiplication for the final generation of attention-aware features.

Early studies on dynamic hand gesture recognition mainly take 2D videos captured by RGB cameras as input, which inevitably causes the loss of valuable 3D spatial information and brings in extra challenges of occlusions and light variation [5, 28, 16, 29, 9]. In recent years, thanks to the drastic advances of cost-effective depth sensors, like Microsoft Kinect or Intel RealSense [10, 15, 33, 27], reliable joint coordinates can be easily obtained using hand pose estimation algorithms [6, 22, 34], and thus skeleton-based dynamic hand gesture recognition has become an active research field.

Traditionally, spatial-temporal hand gesture descriptors are first extracted from the input skeleton sequences, and then a classifier is employed for the final predictions [29, 30, 5, 28, 25, 9, 24, 3]. In recent years, computer vision has witnessed a great success of the introduction of deep learning methods [18, 11, 26, 13]. However, there is significantly little work in the literature using Deep Neural Networks (DNNs) to deal with skeleton-based dynamic hand gesture recognition. To the best of our knowledge, the only literature [23], which does employ DNNs, sticks to a two-stage training strategy with a Convolution Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) recurrent network, instead of an end-to-end framework. The CNN focuses on the extraction of spatial features related to the position of the skeleton joints in 3D space, and the LSTM recurrent network is then used to explore time evolutions and drawing predictions.

Recently a novel set of networks, Temporal Convolutional Networks (TCNs), is proved to be an effective approach to capture spatial-temporal patterns in

the context of action segmentation and human action recognition task [18, 16]. However, given the high intra-class variance nature of hand gestures, not all features extracted by TCN are necessarily informative for every specific input video at every time step. Attention mechanism needs to be introduced to assist the model to adaptively focus on the informative time frames and features.

Inspired by the work of TCN [18, 16], we propose an end-to-end Spatial-Temporal Attention Res-TCN (STA-Res-TCN). The proposed STA-Res-TCN adaptively learns different levels of attention through a mask branch, and assigns them to each spatial-temporal feature extracted by a main branch through an element-wise multiplication. Experimental results demonstrate that the STA-Res-TCN has achieved state-of-the-art performance on DHG-14/28 Dataset [29] and SHREC'17 Track Dataset [30] on both the 14 gestures setting and the more complicated 28 gestures setting.

2 Related Work

In this section, we first provide a literature review on skeleton-based dynamic hand gesture recognition. We then extend our review to works focusing on attention mechanism.

2.1 Skeleton-based Dynamic Hand Gesture Recognition

Skeleton-based dynamic hand gesture recognition has become a heated research field thanks to the advances of cost-effective depth sensors and hand pose estimation algorithms. In this section, we briefly review the existing literature on skeleton-based dynamic hand gesture recognition, which can be gathered into two main categories: approaches with **traditional feature extraction** and approaches with **DNNs**.

Approaches with traditional feature extraction Smedt et al. [29, 30] propose a new descriptor named Shape of Connected Joints (SoCJ), from which a Fisher Vector (FV) representation is computed. The FV representation is then concatenated with two other descriptors, Histogram of Hand Directions (HoHD) and Histogram of Wrist Rotations (HoWR). The temporal information is encoded using a temporal pyramid and the classification process is performed by a linear Support Vector Machine (SVM) classifier. Smedt et al. [30] also evaluate the performances of the other two depth-based descriptors, HOG^2 [24] and HON4D [25], and a skeleton-based method proposed by Devanne et al. [9] originally presented for human action recognition. Chen et al. [5] first extract finger motion features and global motion features from the input dynamic hand gesture skeleton sequence, and then feed these motion features, along with the skeleton sequence, into a recurrent neural network (RNN) to get the final predictions. Boulahia et al. [3] introduce the HIF3D feature-set [2], which is initially conceived for modeling whole body actions, to the domain of dynamic hand gesture recognition. For final classification, they also employ the SVM classifier.

Approaches with DNNs Nunez et al. [23] propose an architecture consists of a combination of a Convolution Neural Network (CNN) followed by a Long Short-Term Memory (LSTM) recurrent network. The CNN focuses on the extraction of the spatial features, and the LSTM recurrent network is then used to capture the patterns related to the time evolution. The CNN is first pre-trained independently by connecting to a Fully-connected Multilayer Perceptron (MLP). Later, the output of the CNN is connected to the LSTM for the second stage training.

All the works above fail to develop an end-to-end framework to explore spatial features and temporal features at the same time, and to include the process of final classification in the same network.

2.2 Attention Mechanism

Studies in neural science show that attention mechanism plays an important role in human visual system [35, 8]. Recently, the exploration of attention mechanism applied in deep learning has attracted increasing interests in various fields, including skeleton-based human action recognition [31, 20]. However, to the best of our knowledges, there is no work in the literature applying attention mechanism to skeleton-based dynamic hand gesture recognition. Even for the works on human action recognition, the attention modules in the existing literatures are mostly built on top of the Long Short-Term Memory (LSTM) recurrent networks. There is a lack of investigation of TCNs, which exhibit totally different characteristics from LSTM-based models.

3 Spatial-Temporal Attention Res-TCN

Our proposed STA-Res-TCN consists of a main branch for feature processing and a mask branch for attention. The overall architecture is shown in Fig. 1. In this work, we employ TCN with residual units (Res-TCN) to construct the main branch.

In order to put our proposed model into context, we first provide a brief overview of TCN and its variant Res-TCN as in the original paper [18, 16]. Then we describe our proposed Spatial-Temporal Attention Res-TCN for skeleton-based dynamic hand gesture recognition. Finally, the employed data augmentation techniques are introduced.

3.1 Overview of Temporal Convolution Networks

The Temporal Convolution Network (TCN) [18] is built from stacked units of 1-dimensional convolution across the temporal domain followed by a non-linear activation function and max pooling. The input to a TCN is a temporal sequence of D-dimensional feature vectors extracted per video frame. Specifically, for a video of T frames, the input X_0 is a concatenation of all frame-wise D-dimensional feature vector across time such that $X_0 \in \mathbb{R}^{T \times D}$. Note that T is the

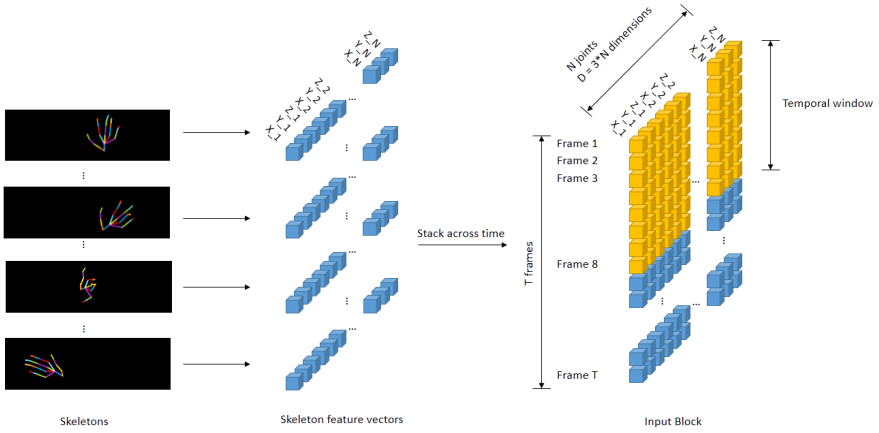

Fig. 2. The input data structure for the network. **Left:** The hand skeleton for each frame. **Middle:** The D-dimensional skeleton feature vector for each frame, constructed by concatenating the 3D coordinates of each hand joints. **Right:** The input block, constructed by stacking the skeleton feature vectors across time. The yellow-highlighted block demonstrates the first application of the $f_1 * N_0 = f_1 * D = 8 * 66$ 1-dimensional temporal convolution kernel.

length of the input and D is the number of channels of the input. In a TCN, the l-th temporal convolution layer consists of N_l filters, each with a temporal window of f_l frames, denoted as $\{W_l^{(i)}\}_{i=1}^{N_l}$ where each filter is $W_l^{(i)} \in \mathbb{R}^{f_l \times N_{l-1}}$. Given the output from the previous layer $X_{l-1} \in \mathbb{R}^{T \times N_{l-1}}$, the activations $X_l \in \mathbb{R}^{T \times N_l}$ can be computed with

$$X_l = f(W_l \otimes X_{l-1}), \tag{1}$$

where $f(\cdot)$ is non-linear activation function ReLU, and $\otimes$ denotes 1-dimensional temporal convolution.

Since the original TCN is designed for action segmentation task in RGB video, the encoder reviewed above is followed by a decoder with similar architecture, except that upsampling is used instead of pooling. Kim et al. [16] extend the original TCN to human action recognition task by adopting only the encoder portion and applying global average pooling and a softmax layer [19] for prediction. They also employ the residual connections as introduced in [11, 12].

3.2 Spatial-Temporal Attention Res-TCN

We propose an end-to-end Spatial-Temporal Attention Res-TCN (STA-Res-TCN) for skeleton-based dynamic hand gesture recognition. The overall architecture is shown in Fig. 1.

For each video frame, a D-dimensional skeleton feature vector is constructed by concatenating the 3D coordinates of each hand joints. The frame-wise skeleton feature vectors are then stacked temporally across the entire video sequence to form the input block $X_0 \in \mathbb{R}^{T \times D}$, as shown in Fig. 2, which is later fed into the STA-Res-TCN.

Given the high-intra variance nature of hand gestures, we notice that not all video frames and not all features extracted by TCN contain the most discriminative information. Irrelevant time frames and features often bring in unnecessary noises. Given this observation, along with the main branch, we introduce an extra attention branch to generate same size masks at each layer which softly weight the feature maps extracted by the main branch. Such soft attention mechanism assists the model to adaptively focus more on the informative frames and features.

To be specific, given the output of the previous block $X_{l-1} \in \mathbb{R}^{T \times N_{l-1}}$ from the main branch and $\tilde{M}_{l-1} \in \mathbb{R}^{T \times N_{l-1}}$ from the mask branch, the feature maps extracted by the main branch and the masks with the same size generated by the mask branch at the l-th block can be respectively computed with:

$$\tilde{X}_l = X_{l-1} + F(W_{l_{main}}, X_{l-1}), \tag{2}$$

$$\tilde{M}_l = G(W_{l_{mask}}, \tilde{M}_{l-1}), \tag{3}$$

where $\{W^{(i)}_{l_{main}}\}_{i=1}^{N_l}$ and $\{W^{(i)}_{l_{mask}}\}_{i=1}^{N_l}$ respectively denotes the collection of filters of the l-th block for the main branch and the mask branch; $F(\cdot)$ and $G(\cdot)$ denotes a series of operations of batch normalization [14], ReLU activation, drop out [32] and 1-dimensional temporal convolution. $\tilde{X}_l$ and $\tilde{M}_l$ both have N_l channels, and each channel has T frames. For channel $i \in \{1, 2, ..., N_l\}$, $\tilde{X}^{(i)}_l = \{\tilde{x}^{(i)}_{l,1}, ..., \tilde{x}^{(i)}_{l,T}\} \in \mathbb{R}^T$ calculates the time evolution of the response to the i-th convolution filter of the l-block. The i-th channel mask $\tilde{M}^{(i)}_l = \{\tilde{m}^{(i)}_{l,1}, ..., \tilde{m}^{(i)}_{l,T}\} \in \mathbb{R}^T$ are the scores indicating the importance of each time frame. We softly weight $\tilde{X}^{(i)}_l$ with the scores $\tilde{M}^{(i)}_l$ to achieve temporal attention for the i-th channel. Similarly, for time step $t \in \{1, 2, ..., T\}$, $\tilde{M}_{l,t} = \{\tilde{m}^{(1)}_{l,t}, ..., \tilde{m}^{(N_l)}_{l,t}\} \in \mathbb{R}^{N_l}$ are the scores indicating the importance of each channel (i.e.each feature extracted by convolution filters). We softly weight $\tilde{X}_{l,t} \in \mathbb{R}^{N_l}$ with the scores $\tilde{M}_{l,t}$ to achieve spatial attention (i.e.attention upon features). Thus, by performing an element-wise multiplication between the main branch feature maps $\tilde{X}_l$ and the masks $\tilde{M}_l$, we gain the spatial-temporal attention-aware feature maps:

$$X_l = \tilde{X}_l * M_l, \tag{4}$$

$$M_l = Sigmoid(\tilde{M}_l). \tag{5}$$

A sigmoid layer is employed to restrict the output range of the masks to $[0, 1]$.

Note that the first block of STA-Res-TCN does not involve a residual unit or any normalization and activation layer, the computation can be formulated

as follows:

$$\tilde{X}_1 = W_{1_{main}} \otimes X_0, \tag{6}$$

$$\tilde{M}_1 = W_{1_{mask}} \otimes X_0, \tag{7}$$

$$M_1 = Sigmoid(\tilde{M}_1), \tag{8}$$

$$X_1 = \tilde{X}_1 * M_1. \tag{9}$$

where $\otimes$ denotes 1-dimensional temporal convolution.

For classification, we employ global average pooling after the last block across the entire temporal sequence and followed by a softmax layer to draw final predictions.

3.3 Data Augmentation

Overfitting is a severe problem in deep neural networks. It leads to an adequate performance on the training set, but a poor performance on the test set [23, 32]. Either the DHG-14/28 Dataset or the SHREC'17 Track Dataset contains no more than 2700 hand gesture sequences for training, which are not enough to prevent overfitting. We employ the same data augmentation techniques with Nunez et al. [23] for fair comparison, including **scaling**, **shifting**, **time interpolation** and **adding noise**. We expand the original training set by 4 times.

4 Experiments

We have evaluated our proposal on two challenging datasets, DHG-14/28 Dataset [29] and SHREC'17 Track Dataset [30]. Experimental results show that STA-Res-TCN outperforms the state-of-the-art methods.

4.1 Datasets and Settings

DHG-14/28 Dataset DHG-14/28 dataset [29] is a public dynamic hand gesture dataset, which contains sequences of 14 hand gestures performed 5 times by 20 participants in 2 finger configurations, resulting in 2800 video sequences. The coordinates of 22 hand joints in the 3D world space are provided per frame, forming a full hand skeleton. The Intel RealSense short range depth camera is used to collect the dataset.

SHREC'17 Track Dataset SHREC'17 Track Dataset [30] is a public dynamic hand gesture dataset presented for the SHREC'17 Track. It contains sequences of 14 gestures performed between 1 and 10 times by 28 participants in 2 finger configurations, resulting in 2800 sequences. The coordinates of 22 hand joints in the 3D world space are provided per frame. The dataset is captured by Intel Realsense camera.

Implementation Details We perform all our experiments on a Nvidia GeForce GTX 1080 GPU with Keras 2.0 [7] using TensorFlow [1] backend. The

learning rate is initially set to be 0.01 and then is gradually reduced by a factor 10 at 0.2, 0.5, 0.6 of the total epochs for the DHG-14/28 Dataset, and at 0.3, 0.6, 0.9 for the SHREC'17 Track Dataset. We employ the Adam algorithm [17] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e^{-8}$. The batch size is set to 256, and the network is trained for up to 200 epochs. L-1 regularizer with a weight of $1e^{-4}$ is applied to all convolution layers. The dropout [32] rate is set to be 0.5 to prevent overfitting. The length of temporal window f_l is set to be 8 frames. Every skeleton sequence is subtracted by the palm position of the first frame.

4.2 Comparisons with State-of-the-Arts

In the experiment on DHG-14/28 Dataset, we follow a leave-one-subject-out cross-validation strategy, i.e., we perform 20 experiments, each one using data from 19 subjects for training and data from the rest 1 subject for testing. The reported results are computed as the average over these 20 cross-validation folds.

We show performance comparisons of STA-Res-TCN with state-of-the-art methods in Table 1. The recognition rate of our proposed model achieves 89.2% for the 14 gestures setting and 85.0% for the more complicated 28 gestures setting. The experimental results in Table 1 demonstrate a significant enhancement of recognition rates in comparison with state-of-the-art methods with both settings. Note that for fair comparison, we employ the same data augmentation techniques with Nunez et al. [23], the current state-of-the-art method, to both the baseline model Res-TCN and the attention model STA-Res-TCN. By comparing the performance of the baseline model and STA-Res-TCN, we can observe that our proposed attention mechanism brings 2.3% and 1.4% accuracy raise for the 14 gestures setting and 28 gestures setting respectively.

Table 1. Comparisons of accuracy (%) on DHG-14/28 Dataset.

Method	14 gestures	28 gestures
SoCJ+HoHD+HoWR [29]	83.1	80.0
Chen et al. [5]	84.7	80.3
CNN+LSTM [23]	85.6	81.1
Res-TCN (Baseline)	86.9	83.6
STA-Res-TCN (Ours)	**89.2**	**85.0**

The confusion matrices with 14 gestures setting and 28 gestures setting are shown in Fig. 3 and Fig. 4. It can be observed that our proposed STA-Res-TCN achieves recognition rate higher than 90.0% in 11 of the 14 gestures. The accuracy comparison for each individual gesture is favorable to our proposal in 10 of the 14 gestures compared to the work of [23]. It can also be observed from the confusion matrix that the gestures *Grab* and *Pinch* are usually misclassified due to the low inter-class variance.

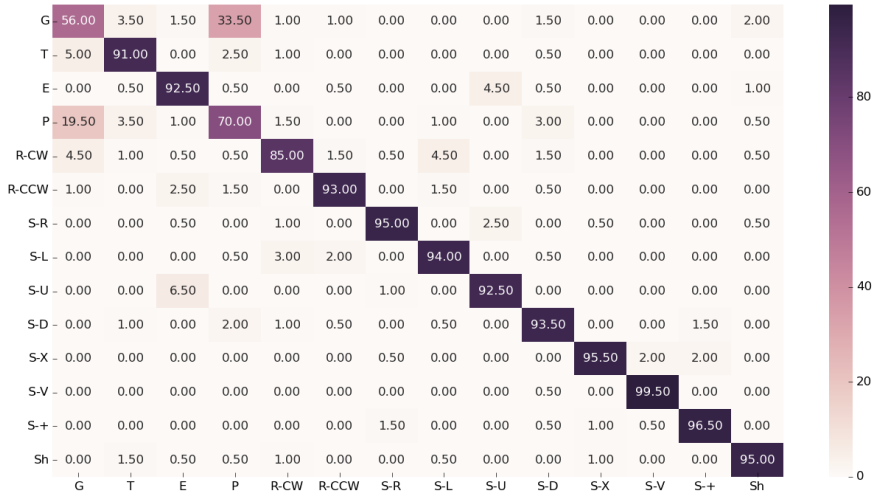

Fig. 3. Confusion matrix on DHG-14/28 Dataset with 14 gestures setting.

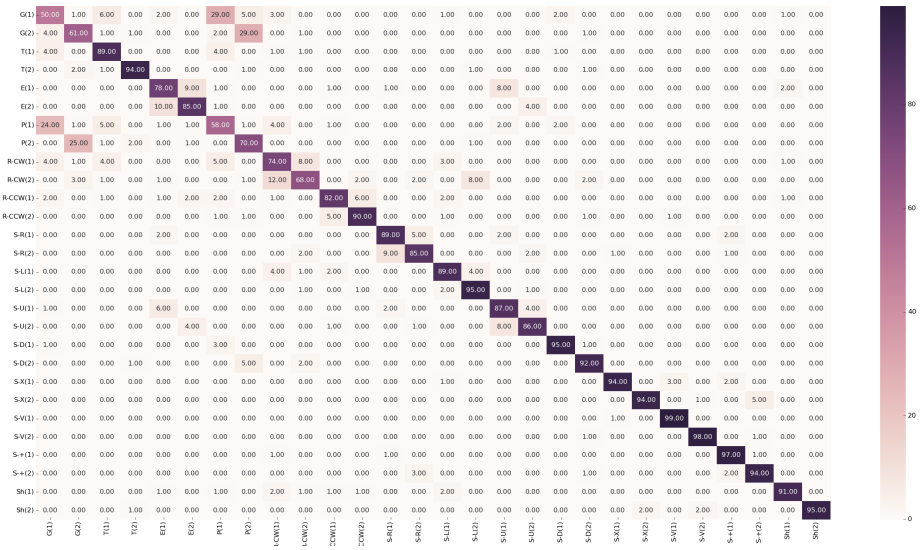

Fig. 4. Confusion matrix on DHG-14/28 Dataset with 28 gestures setting.

In the experiment on SHREC'17 Track Dataset, we follow the division of the training set and the test set of the SHREC'17 Track [30], resulting in 1960 training sequences and 840 test sequences. We still employ the same data augmentation technique to both the baseline model and the attention model STA-Res-TCN.

Table 2. Comparisons of accuracy (%) on SHREC'17 Track Dataset.

Method	14 gestures	28 gestures
Oreifej et al.† [25]	78.5	74.0
Devanne et al.† [9]	79.6	62.0
Classify Sequence by Key Frames [30]	82.9	71.9
Ohn-Bar et al.† [24]	83.9	76.5
SoCJ+Direction+Rotation [28]	86.9	84.2
SoCJ+HoHD+HoWR [29]	88.2	81.9
Caputo et al. [4]	89.5	-
Boulahia et al.† [3]	90.5	80.5
Res-TCN (Baseline)	91.1	87.3
STA-Res-TCN (Ours)	**93.6**	**90.7**

As demonstrated in Table 2, the STA-Res-TCN achieves the accuracy of 93.6% for the 14 gestures setting and 90.7% for the more complicated 28 gestures setting. Our proposed model outperforms the state-of-the-art models, especially showing greater accuracy improvement with the more complicated 28 gestures setting, which further validates the effectiveness of our proposed model. By comparing the performance of the baseline model Res-TCN and the attention model STA-Res-TCN, we can observe that our proposed attention mechanism brings 2.5% and 3.4% accuracy raise respectively for the 14 gestures setting and 28 gestures setting.

The confusion matrices with 14 gestures setting and 28 gestures setting are shown in Fig. 5 and Fig. 6. It can be observed that our proposed STA-Res-TCN achieves recognition rate higher than 90.0% in 10 of the 14 gestures, and achieves recognition rate higher than 85.0% in 13 of the 14 gestures. The accuracy comparison for each individual gesture is favorable to our proposal in 10 of the 14 gestures compared to the work of [3].

Moreover, our proposed model can be trained in an extremely short time, no more than 30 minutes for SHREC'17 Track Dataset or one cross-validation fold of DHG-14/28 Dataset with a Nvidia GeForce GTX 1080 GPU. The inference speed of STA-Res-TCN is also considerably fast. It can process 9691 skeletons per second (i.e.161 hand gestures per second on average), which exceeds the 7615 skeletons per second performance presented by the work of [23] on the same hardware architecture. The processing speed of our proposed STA-Res-

† Implement and evaluate by Smedt et al. [30]

TCN also far exceeds the standard of real time analysis (i.e.30 skeletons per frame) in video.

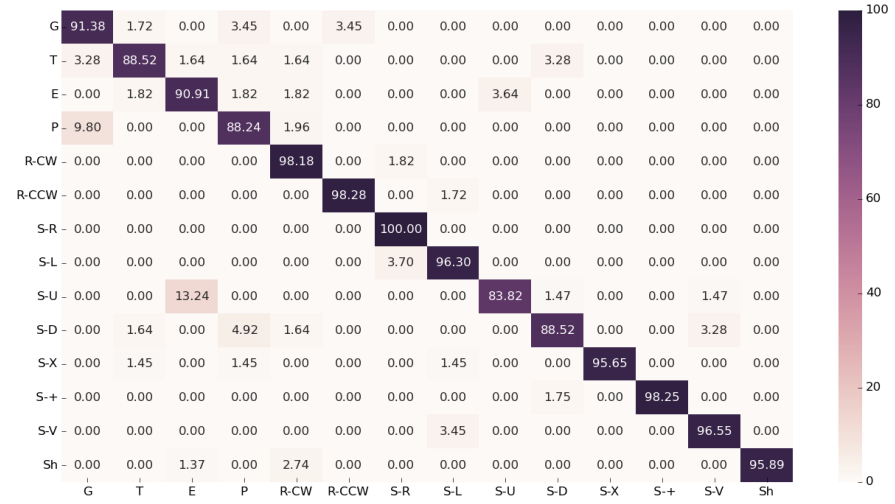

Fig. 5. Confusion matrix on SHREC-14/28 Dataset with 14 gestures setting.

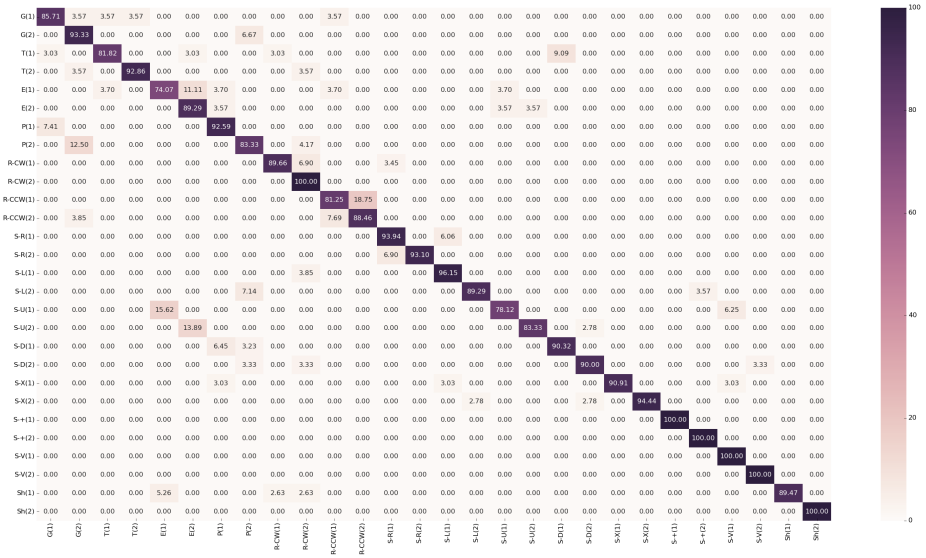

Fig. 6. Confusion matrix on SHREC-14/28 Dataset with 28 gestures setting.

4.3 Visualization of the Spatial-temporal Attention

For better understanding of our work, we analyze our proposed attention mechanism by visualizing and comparing the feature maps before/after soft attention masks.

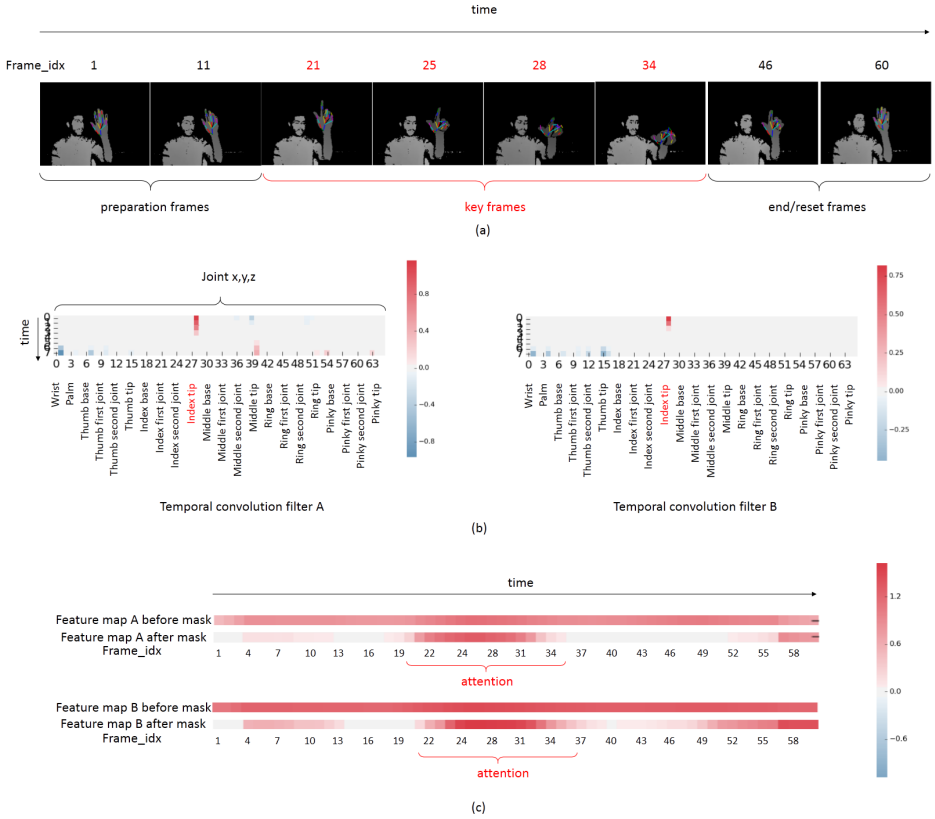

Fig. 7. Visualization of our proposed attention mechanism on a skeleton sequence of hand gesture "tap". **(a)** Input skeleton sequence of hand gesture "tap". The key frames range approximately from the 21^{st} frame to the 34^{th} frame. **(b)** Two examples of the temporal convolution filters that mainly learn the movements of the tip joint of index finger. **(c)** A comparison between the feature maps before soft attention masks and feature maps after soft attention masks corresponding to the two filters.

For a skeleton sequence of hand gesture "tap", as shown in Fig. 7(a), the key movements mainly relate to the tip joint of index finger, and the key frames

that contain the most discriminative information range approximately from the 21^{st} frame to the 34^{th} frame. Fig. 7(b) shows the two temporal convolution filters which our proposed attention mechanism has the greatest impact on. These two filters mainly learn a downward translation movement of the tip joint of index finger along the y axis, which is in accord with which human considers as key movement. Fig. 7(c) shows a comparison between the feature maps before/after soft attention masks corresponding to the two convolution filters mentioned above. It can be observed that the time frames which our proposed attention mechanism stressed more attention on are consistent with which human perceives as discriminative.

5 Conclusion

We present an end-to-end Spatial-Temporal Attention Res-TCN for skeleton-based dynamic hand gesture recognition, which learns to adaptively assign different levels of attention to each spatial-temporal features at each time step as layers going deeper. Experimental results demonstrate the effectiveness of the proposed STA-Res-TCN, which achieves significant accuracy enhancement in comparison with other state-of-the-art methods. Moreover, our proposed STA-Res-TCN is a lightweight model, which can be trained and tested in an extremely short time.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Boulahia, S., Anquetil, E., Kulpa, R., Multon, F.: Hif3d: Handwriting-inspired features for 3d skeleton-based action recognition. In: ICPR (2017)
3. Boulahia, S.Y., Anquetil, E., Multon, F., Kulpa, R.: Dynamic hand gesture recognition based on 3d pattern assembled trajectories. In: IPTA (2017)
4. Caputo, F., Prebianca, P., Carcangiu, A., Spano, L.D., Giachetti, A.: Comparing 3d trajectories for simple mid-air gesture recognition. Computers & Graphics **73**, 17–25 (2018)
5. Chen, X., Guo, H., Wang, G., Zhang, L.: Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In: ICIP (2017)
6. Chen, X., Wang, G., Guo, H., Zhang, C.: Pose guided structured region ensemble network for cascaded hand pose estimation. Neurocomputing (2018)
7. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
8. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nature reviews neuroscience **3**, 201–215 (2002)

9. Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Bimbo, A.D.: 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. IEEE transactions on cybernetics **45**(7), 1340–1352 (2015)
10. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: A review. IEEE Trans. Cybern. **43**(5), 1318–1334 (2013)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV (2016)
13. Huang, G., Liu, Z., Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
15. Keselman, L., Woodfill, J.I., G.-Jepsen, A., Bhowmik, A.: Intel realsense stereoscopic depth cameras. In: CVPRW (2017)
16. Kim, T., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. In: CVPR BNMW Workshop (2017)
17. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: CVPR (2017)
19. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
20. Liu, J., Wang, G., Hu, P., Duan, L.Y., Kot, A.C.: Global context-aware attention lstm networks for 3d action recognition. In: CVPR (2017)
21. Molchanov, P., Gupta, S., Kim, K., Kautz, J.: Hand gesture recognition with 3d convolutional neural networks. In: CVPRW (2015)
22. Moon, G., Chang, J.Y., Lee, K.M.: V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. arXiv preprint arXiv:1711.07399 (2018)
23. Nunez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., Velez, J.F.: Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition **76**, 80–94 (2018)
24. Ohn-Bar, E., Trivedi, M.: Joint angles similarities and hog2 for action recognition. In: CVPRW (2013)
25. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: CVPR (2013)
26. Ross, Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
27. Shi, C., Wang, G., Yin, X., Pei, X., He, B., Lin, X.: High-accuracy stereo matching based on adaptive ground control points. IEEE Transactions on Image Processing **24**(4), 1412–1423 (2015)
28. Smedt, Q.D.: Dynamic hand gesture recognition - from traditional handcrafted to recent deep learning approaches, computer Vision and Pattern Recognition [cs.CV]. Universite de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, 2017. English.
29. Smedt, Q.D., Wannous, H., Vandeborre, J.P.: Skeleton-based dynamic hand gesture recognition. In: CVPRW (2016)
30. Smedt, Q.D., Wannous, H., Vandeborre, J.P., Guerry, J., Saux, B.L., Filliat, D.: Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset. In: Eurographics Workshop on 3D Object Retrieval (2017)

31. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI (2017)
32. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014)
33. Wang, G., Yin, X., Pei, X., Shi, C.: Depth estimation for speckle projection system using progressive reliable points growing matching. Applied Optics **52**, 516–524 (2013)
34. Wang, G., Chen, X., Guo, H., Zhang, C.: Region ensemble network: Towards good practices for deep 3d hand pose estimation. Journal of Visual Communication and Image Representation **55**, 404–414 (2018)
35. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)