Action Recognition with Visual Attention on Skeleton Images

Zhengyuan Yang[1], Yuncheng Li[2], Jianchao Yang[2], and Jiebo Luo[1]
[1]Department of Computer Science, University of Rochester, Rochester NY 14627, USA
[2]Snapchat Inc., Venice, CA 90291, USA
[1]Email: {zyang39, jluo}@cs.rochester.edu
[2]Email: {yuncheng.li, jianchao.yang}@snapchat.com

***Abstract*—Action recognition with 3D skeleton sequences is becoming popular due to its speed and robustness. The recently proposed Convolutional Neural Networks (CNN) based methods have shown good performance in learning spatio-temporal representations for skeleton sequences. Despite the good recognition accuracy achieved by previous CNN based methods, there exist two problems that potentially limit the performance. First, previous skeleton representations are generated by chaining joints with a fixed order. The corresponding semantic meaning is unclear and the structural information among the joints is lost. Second, previous models do not have an ability to focus on informative joints. The attention mechanism is important for skeleton based action recognition because there exist spatio-temporal key stages and the joint predictions can be inaccurate. To solve the two problems, we propose a novel CNN based method for skeleton based action recognition. We first redesign the skeleton representations with a depth-first tree traversal order, which enhances the semantic meaning of skeleton images and better preserves the structural information. We then propose the idea of a two-branch attention architecture that focuses on spatio-temporal key stages and filters out unreliable joint predictions. A base attention model with the simplest structure is first introduced to illustrate the two-branch attention architecture. By improving the structures in both branches, we further propose a Global Long-sequence Attention Network (GLAN). Experiment results on the NTU RGB+D dataset and the SBU Kinetic Interaction dataset show that our proposed approach outperforms the state-of-the-art, as well as the effectiveness of each component.**

I. Introduction

The frequently used modalities for action recognition include RGB videos [1], [2], [3], optic flow [4], [5], [6] and skeleton sequences. Comparing to RGB videos and optic flow, skeleton sequences require less computation. Furthermore, skeleton sequences have a better ability to represent dataset-invariant action information since no background information is included. One limitation is that labeling skeleton sequences manually is too expensive, while the automatic annotation methods may yield inaccurate predictions. With the above advantages and the fact that skeletons can now be more reliably predicted [7], [8], skeleton based human action recognition is becoming increasingly popular. The major goal for skeleton based recognition is to learn a representation that best preserves the spatio-temporal relations among the joints.

With a strong ability in modeling sequential data, Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) neurons outperform the previous hand-crafted feature based methods [9], [10]. Each skeleton frame is converted into a feature vector and the whole sequence is fed into the RNN. Despite the strong ability in modeling temporal sequences, RNN structures lack the ability to efficiently learn the spatial relations between the joints. To better use spatial information, a hierarchical structure is proposed in [11], [12] that feds the joints into the network as several pre-defined body part groups. However, the pre-defined body regions still limit the effectiveness of representing spatial relations. A spatio-temporal 2D LSTM (ST-LSTM) network [13] is proposed to learn the spatial and temporal relations simultaneously. Furthermore, a two-stream RNN structure [14] is proposed to learn the spatio-temporal relations with two RNN branches.

CNN has a natural ability to learn representations from 2D arrays. [15], [16] first propose to represent the skeleton sequences as 2D gray scale images and use CNN to jointly learn a spatio-temporal representation. Each gray scale image corresponds to one axis in the joint coordinates. For example, the coordinates in the x-axis throughout a skeleton sequence generate one single-channel image. Each row is a spatial distribution of coordinates at a certain time-stamp, and each column is the temporal evolution of a certain joint. The generated 2D arrays are then scaled and resized into a fixed size. Gray scale images generated from the same skeleton sequence are concatenated together and processed as a multi-channel image, which is called the skeleton image.

Despite the large boost in recognition accuracy achieved by previous CNN based methods, there exist two problems. First, previous skeleton image representations lose spatial information. In previous methods, each row represents skeleton's spatial information by chaining all joints with a fixed order. This concatenation process lacks semantic meaning and leads to a loss in skeleton's structural information. Although a good chain order can perverse more spatial information, it is impossible to find a perfect chain order that maintains all spatial relations in the original skeleton structure. We propose a Tree Structure Skeleton Image (TSSI) to preserve spatial relations. TSSI is generated by traversing a skeleton tree with a depth-first order. We assume the spatial relations between joints are represented by the edges that connect them in the original skeleton structure, as shown in Figure 1 (a). The fewer edges there are, the more relevant the joint pair is. Thus we prove that TSSI best preserves the spatial relation.

Second, previous CNN based methods do not have the ability to focus on spatial or temporal key stages. In skeleton based action recognition, certain joints and frames are more informative, like the joints on the arms in action 'waving hands'. Furthermore, certain joints may be inaccurately predicted and should be neglected. Therefore, it is important to

include attention mechanisms. For a 2D attention mask, each row represents the spatial importance of key joints and each column represents the temporal importance of key frames. We propose a two-branch architecture for visual attention on a single skeleton image. One branch generates an attention mask with a larger receptive field and the other branch refines the CNN feature. We first introduce the two-branch architecture with a base attention model. Furthermore, a Global Long-sequence Attention Network (GLAN) is proposed with refined branch structures. Experiments on public datasets prove the effectiveness of the two improvements. The recognition accuracy is superior to the state-of-the-art methods.

Our main contributions include the following:

- We propose a Tree Structure Skeleton Image (TSSI) that better preserves the spatial relations in skeleton sequences. TSSI is based on a depth-first tree traversal order instead of direct concatenation.
- We propose a two-branch visual attention architecture for skeleton based action recognition. A Global Long-sequence Attention Network (GLAN) is introduced based on the proposed architecture.

II. Related Work

Compared to other frequently used modalities including RGB videos [1], [2], [3] and optical flow [4], [5], [6], skeleton sequences require much less computation and are robust across views and datasets. With the advanced methods to acquire reliable skeletons from RGBD sensors [7] or even single RGB cameras [8], [17], [18], skeleton-based action recognition is becoming increasingly popular.

Many previous skeleton-based action recognition methods [19] model the temporal pattern of skeleton sequences with Recurrent Neural Networks. Hierarchical structures [11], [12] better represent the spatial relations between body parts. Other works [20], [21] adopt attention mechanisms to locate spatial key joints and temporal key stages in skeleton sequences. [13] proposes a 2D LSTM network to learn spatial and temporal relations simultaneously. [14] models spatio-temporal relations with a two-stream RNN structure. Other effective approaches include lie groups [10], [22] and nearest neighbor search [23]. Recently, graphical neural networks [24] achieve the state-of-the-art performance on the skeleton based recognition task.

Comparing to LSTM or graphical model based methods, the recently proposed CNN based approaches show a better performance in learning skeleton representations. [15], [16] propose to convert human skeleton sequences into gray scale images, where the joint coordinates are represented by the intensity of pixels. [25] proposes to generate skeleton images with 'Skepxels' to better represent the joint correlations. In this paper, we further improve the design of skeleton images with a depth-first traversal on skeleton trees.

Attention mechanisms are important for skeleton based action recognition. Previous LSTM based methods [20], [21] learn attention weights between the stacked LSTM layers. For CNN based methods, we propose that general visual attention can be directly adopted to generate 2D attention masks, where each row represents spatial importance and each column represents temporal importance. Visual attention has achieved successes in many areas, including image captioning [26], [27], RGB based action recognition [28], [29], image classification [30], [31], sentiment analysis [32] and etc. Many visual attention methods take an image sequence as input [28], [33], or use extra information from another modality like text [26], [27], [29]. Because a single skeleton image already represents a spatio-temporal sequence without the need for an extra modality, we propose a single frame based visual attention structure with a same setting in [30], [31].

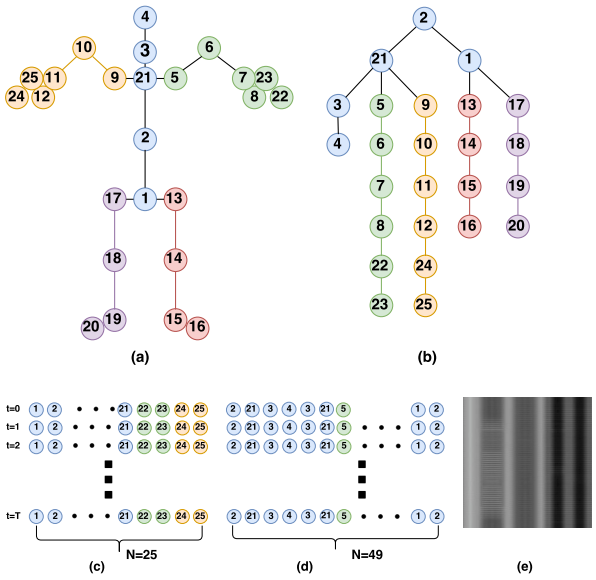

Fig. 1. Tree Structure Skeleton Image (TSSI). (a). Skeleton structure and order in NTU RGB+D. (b). Skeleton tree for TSSI generating. (c). Joint arrangements of naive skeleton images. (d). Joint arrangements of TSSI. (e). An example frame of TSSI. Different colors represent different body parts.

III. Method

In this section, we first introduce the previous design of skeleton images and the base CNN structure. Then an improved Tree Structure Skeleton Image (TSSI) is proposed. Finally, we propose the idea of two-branch visual attention architecture and introduce a Global Long-sequence Attention Network (GLAN) based on the architecture.

A. Base Model

In CNN based skeleton action recognition, joint sequences are arranged as 2D arrays that are processed as gray scale images. We call such a generated image the 'Skeleton Image'. For a channel in skeleton images, each row contains the chaining of joint coordinates at a certain time-stamp. Each column represents the coordinates of a certain joint throughout the entire video clip. The chain order of joints is pre-defined and fixed. An arrangement of the 2D array is shown in Figure 1 (c). The generated 2D arrays are then scaled into 0 to 255, and resized into a fixed size of $224 * 224$. The processed 2D arrays are processed as gray scale images, where each channel represents an axis of joint coordinates. The skeleton images

are fed into CNNs for action recognition. We use ResNet-50 [34] as the base ConvNet model. Comparing to RNN based or graph neural network based method, CNN based methods can better learn the spatio-temporal relations between joints.

B. Tree Structure Skeleton Image

A shortcoming in previous skeleton images is that each row is arranged by simply concatenating all joints. Each row contains the concatenation of all joints with a pre-defined chain order. CNN has a feature that the receptive field grows larger at higher levels. Therefore, the adjacent joints in each row or column are learned first at lower levels. This implies that the adjacent joints share more spatial relations in original skeleton structure, which often do not hold in previous skeleton images. In previous skeleton images, a generated array has 25 columns representing the joint coordinates of joint 1 to 25 with joint indexes shown in Figure 1 (a). An arrangement of the skeleton image is shown in Figure 1 (c). In this case, a convolutional kernel might cover joints [20, 21, 22, 23, 24] at a certain level since these joints are adjacent in skeleton images. However, these joints have less spatial relations in original skeleton structures and should not be learned together directly.

To solve this problem, we propose a Tree Structure Skeleton Image (TSSI) inspired by [13]. The basic assumption is that the spatially related joints in original skeletons have direct graph links between them. The less edges required to connect a pair of joints, the more related is the pair. The human structure graph is defined with semantic meanings as shown in 1 (a). In the proposed TSSI, the direct concatenation of joints is replaced by a depth-first tree traversal order. The skeleton tree is defined in Figure 1 (b) and an arrangement of TSSI is shown in Figure 1 (d). The depth-first tree traversal order for each row is [2, 21, 3, 4, 3, 21, 5, 6, 7, 8, 22, 23, 22, 8, 7, 6, 5, 21, 9, 10, 11, 12, 24, 25, 24, 12, 11, 10, 9, 21, 2, 1, 13, 14, 15, 16, 15, 14, 13, 1, 17, 18, 19, 20, 19, 18, 17, 1, 2]. With the proposed order, the neighboring columns in skeleton images are spatially related in original skeleton structures. This proves that the TSSI best preserves the spatial relations. With TSSI, the spatial relations between related joints are learned first at lower levels of CNN and the relations between less relevant joints are learned later at high levels when receptive field becomes larger. An example of the generated TSSI is shown in Figure 1 (e).

C. Attention Networks

In skeleton sequences, certain joints and frames are extra distinguishable for recognizing actions. For example in action 'waving hands', the joints in arms are more informative. These informative joints and frames are referred to as 'key stages'. Furthermore, noise exists in the captured joint data and deteriorates the recognition accuracy. The inaccurate joints should be automatically filtered out or neglected by the network.

To alleviate data noise and to focus on informative stages, skeleton based methods should adjust weights for different inputs automatically. We propose the idea of two-branch visual attention structure and further design a Global Long-sequence

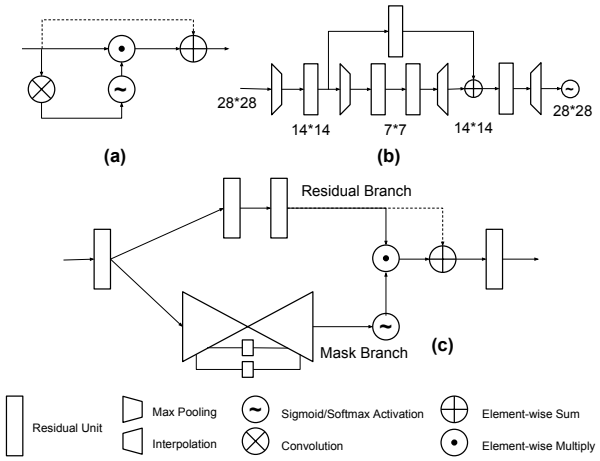

Fig. 2. A base attention module and a GLAN module. (a). A base attention block. (b). An expanded plot for the Hourglass mask branch in GLAN. (c). An attention block with GLAN structure, short for 'GLAN block'.

Attention Network (GLAN) based on the idea. In this section, we first introduce the basic idea of the two-branch attention architecture with a base attention model. Then the detailed structure of the Global Long-sequence Attention Network (GLAN) is introduced.

Base Attention Model. Skeleton images naturally represent both spatial and temporal information of skeleton sequences. Therefore a 2D attention mask can represent spatio-temporal importance simultaneously, where the weights in each row represent the spatial importance of joints and the weight in each column represent the temporal importance of frames. In order to generate the attention masks, we propose a two-branch attention architecture that learns attention masks from a single skeleton image. The two-branch structure is consist of 'mask branches' and 'residual branches'. Taking previous CNN feature blocks as inputs, the mask branch learns a 2D attention mask and the residual branch refines previous CNN feature. The two branches are then merged and output a weighted CNN feature block. To be specific, the mask branch learns an attention mask with a structure that has a larger receptive field. The residual branch is designed to maintain and refine the input CNN features with convolutional layers. The two branches are fused at the end of each attention block with element-wise multiply and sum.

We first introduce the base attention model, which is the simplest version of two-branch attention structures. As shown in Figure 2 (a), the mask branch in the base model gains a larger receptive field with a single convolutional layer. Softmax or Sigmoid functions are used for mask generating. The residual branch preserves the input CNN feature with a direct link. An 'attention block' is defined as a structure with one mask branch and one residual branch as Figure 2 (a). Attention blocks are added between the convolutional blocks in base CNN to build the whole network. In the base attention model, attention blocks are inserted between ResNet-50's residual blocks, with the structure of residual blocks unchanged.

Global Long-sequence Attention Network (GLAN) Based on the proposed two-branch structure, we improve the designs

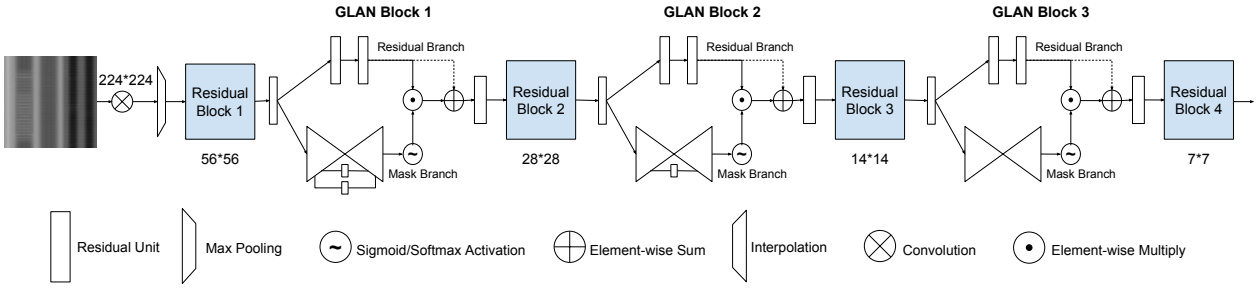

Fig. 3. The framework for Global Long-sequence Attention Network (GLAN).

of both branches to learn masks and CNN features more effectively. Inspired by the hourglass structure [35], [31], we propose a Global Long-sequence Attention Network (GLAN) as shown in Figure 3. The hourglass structure is adopted in mask branches to quickly adjust the feature size and efficiently gain a larger receptive field. As shown in Figure 2 (b), the hourglass structure is consist of a series of down-sampling units followed by up-sampling units. In each hourglass mask branch, input CNN features are first down-sampled to the lowest spatial resolution of $7*7$ and recovered back to the original size. Max pooling is used for down-sampling and bilinear interpolation is used for up-sampling. Each down-sampling unit includes a max pooling layer, a followed residual unit and a link connection to the recovered feature with a same size. Each up-sampling unit contains a bilinear interpolation layer, a residual unit and a element-wise sum with the link connection. We show that the Convolution-Deconvolution structure gains a large receptive field effectively and therefore can better learn an attention mask. For residual branches, we add two residual units to further refine the learned CNN feature. All residual units are the same as ResNet-50 [34], which contains three convolutional units and a direct residual link.

As shown in Figure 3, three GLAN attention blocks are added between the four residual blocks in ResNet-50 to build the GLAN network. The depth of each GLAN blocks varies due to the different input feature sizes. Furthermore, we reduce the number of residual units in each residual block to keep a proper depth of the GLAN network, since GLAN blocks are much deeper than the base attention blocks. Only one residual unit is kept for the first three residual blocks. The final residual block keeps all three residual units as in ResNet-50.

IV. Experiments

The proposed method is evaluated on the NTU RGB+D dataset [12] and the SBU Kinect Interaction Dataset [36]. We further evaluate the effectiveness of each proposed module separately. The experiments show that both the TSSI and the two-branch attention network generates a large boost in action recognition accuracy. The performance of the proposed model outperforms the state-of-the-arts on all datasets.

A. Datasets

NTU RGB+D. The NTU RGB+D dataset [12] is so far the largest 3D skeleton action recognition dataset. NTU RGB+D has 56880 videos collected from 60 action classes, including 40 daily actions, 9 health-related actions and 11 mutual actions. The dataset is collected with Kinect and the recorded skeletons include 25 joints. The train/val/test split follows [12]. Samples with missing joints are discarded as in that paper.

SBU Kinect Interaction. The SBU Kinect Interaction dataset [36] contains 282 skeleton sequences and 6822 frames. We follow the standard experiment protocol of 5-fold cross validations with the provided splits. The dataset contains eight classes. There are two persons in each skeleton frame and 15 joints are labeled for each person. The two skeletons are processed as two data samples during training and the averaged prediction score is calculated for testing.

B. Effectiveness of the Proposed Modules

To prove the effectiveness of the TSSI and attention networks, we separately evaluate each module. Each component of the framework is evaluated on NTU RGB+D with a cross subject setting. NTU RGB+D is selected for component evaluations because it is the the largest and the most challenging dataset so far. Similar results are observed on other datasets.

Traditional Skeleton Image + ConvNet. As a baseline, we adopt the previous skeleton image representation from [15] and use ResNet-50 as a base CNN model to train spatio-temporal skeleton representations. We test the three spatial joint orders proposed by Sub-JHMDB [37], PennAction [38] and NTU RGB+D [12]. Experiments show that the NTU RGB+D's order generates a better accuracy of 1.3% than the rest two orders. Therefore, we adopt the joint order proposed by NTU RGB+D for baseline comparison. The order is shown in Figure 1 (a).

TSSI + ConvNet. The effectiveness of the proposed Tree Structure Skeleton Image (TSSI) is compared to the baseline design of skeleton images. TSSI is the skeleton image generated with a depth-first tree traversal order. The skeleton tree structure, TSSI arrangement and a TSSI example is shown in Figure 1 (b), (d), (e). A large boost in accuracy is observed from 68.0% to 73.1%, which proves the effectiveness of TSSI.

TSSI + Base Attention. The base attention model provides a baseline for two-branch attention networks. The base attention blocks with and without residual links are inserted at three different locations in ResNet-50, that is at the front after the first convolutional layer, in the middle after the second residual block and in the end after the final residual block. The input feature blocks to the three attention blocks have the shapes of $112 * 112 * 64$, $28 * 28 * 512$ and $7 * 7 * 2048$. The recognition accuracy boosts from 73.1% to 74.9%. This

TABLE I
THE ACTION RECOGNITION ACCURACY COMPARING TO THE STATE-OF-THE-ART METHODS ON THE NTU RGB+D DATASET.

State-of-the-art	Cross Subject	Cross View
Lie Group [10]	51.0	52.8
HBRNN [11]	59.1	64.0
Part-aware LSTM [12]	62.9	70.3
Trust Gate LSTM [13]	69.2	77.7
Two-stream RNN [14]	71.3	79.5
TCN [16]	74.3	83.1
Global Attention LSTM [20]	74.4	82.8
A^2GNN [24]	72.7	82.8
Clips+CNN+MTLN [15]	79.6	84.8
Ensemble TS-LSTM [19]	76.0	82.6
Proposed Model	Cross Subject	Cross View
Base Model	68.0	75.5
With TSSI	73.1	76.5
TSSI + Base Attention	74.9	79.1
TSSI + GLAN	**80.1**	**85.2**

experiment shows that even the simplest two-branch attention network can improve the recognition accuracy.

TSSI + GLAN. Finally, we evaluate the proposed Global Long-sequence Attention Network (GLAN). The number of link connections and the depth of the hourglass mask branch can be manually adjusted. In experiments, we first down-sample the feature blocks to a lowest resolution of $7 * 7$ and then up-sample them back to the input size. Each max pooling layer goes with one residual unit, one link connection and one up-sampling unit. With a GLAN structure shown in Figure 3, the recognition accuracy increases from 74.9% to 80.1%.

C. Comparisons to Other State-of-the-Art

NTU RGB+D. As shown in Table I, the base model with naive skeleton images already outperforms a number of previous LSTM based method, without adopting attention mechanism. This shows that CNN based methods are promising for skeleton based action recognition. With the improved TSSI, the cross subject accuracy achieves 73.1%, which is comparable to the state-of-the-art LSTM methods. Finally, the proposed two-branch attention architecture achieves a good performance and the GLAN outperforms the state-of-the-arts. Experiments prove the effectiveness of the proposed CNN based action recognition method.

SBU Kinect Interaction. Similar to the performance on the NTU RGB+D dataset, the proposed TSSI and GLAN generates a large boost in recognition accuracy and outperforms the state-of-the-arts. The performances are shown in Table II.

Furthermore, the proposed TSSI and two-branch attention networks can be adopted as components in future work, e.g., by fusing features generated by TSSI + GLAN with other LSTM based features, for further improvements.

D. Error Case Analysis

To better understand the successful and failure cases, experiments are conducted to analyze the performances of each class. As shown in Table III, two parts of analysis are conducted. First, eight classes that constantly perform the best or worst are selected on the left side of Table III. Results show that the

TABLE II
THE RECOGNITION ACCURACY COMPARING TO THE STATE-OF-THE-ART METHODS ON THE SBU KINETIC INTERACTION DATASET.

State-of-the-art	Accuracy
Raw Skeleton [36]	49.7
HBRNN [11]	80.4
Trust Gate LSTM [13]	93.3
Two-stream RNN [14]	94.8
Global Attention LSTM [20]	94.1
Clips+CNN+MTLN [15]	93.6
Proposed Model	Accuracy
Base Model	82.0
With TSSI	89.2
TSSI + Base Attention	93.6
TSSI + GLAN	**95.6**

actions with dynamic body movements like standing, sitting and walking can be well classified with skeletons, while the classes with less motions like reading, writing and clapping usually have a poor result. This follows human intuition that skeletons are more useful for distinguishing dynamic actions, while additional background context information is necessary for recognizing the actions with less motions. The results also show that the proposed TSSI and GLAN both generates a large boost in performance in all the listed classes. On the right side of the table, statistics of the best and worst classes are listed. Results show that TSSI + GLAN greatly improve the accuracy in challenging classes. The top 1 worst class in TSSI + GLAN has an accuracy of 39.7%, which is even better than the averaged accuracy of the worst 10 in base model. For the best classes, the top 1 accuracy between the baseline and TSSI + GLAN is similar. The improvements are obtained through the increases in the more challenging classes.

V. CONCLUSION

Using CNN for skeleton based action recognition is a promising approach. In this work, we address the two major problems with previous CNN based methods, that is the improper design of skeleton images and the lack of attention mechanisms. The design of skeleton images is improved by introducing the Tree Structure Skeleton Image (TSSI). The two-branch attention structure is then introduced for visual attention on skeleton images. A Global Long-sequence Attention Network (GLAN) is proposed based on the two-branch attention structure. Experiments show that the proposed enhancement modules greatly improve the recognition accuracy, especially on the challenging classes.

ACKNOWLEDGMENTS

We thank the support of New York State through the Goergen Institute for Data Science, our corporate research sponsors Snap and Cheetah Mobile, and NSF Award #1704309.

REFERENCES

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

TABLE III

THE STATISTICS AND NAMES OF CLASSES WITH THE HIGHEST AND LOWEST RECOGNITION ACCURACIES. EXPERIMENTS ARE CONDUCTED ON NTU RGB+D WITH CROSS SUBJECT SETTING. LEFT TABLE SHOWS THE CLASSES THAT CONSTANTLY HAVE A GOOD OR BAD PERFORMANCE. RIGHT TABLE SHOWS THE STATISTICS OF THE TOP AND BOTTOM CLASSES.

Selected Best Classes / Accu.	Base Model	TSSI	TSSI + GLAN
standing up	85.4	94.1	97.1
sitting down	91.6	91.6	93.8
walking apart	90.6	91.3	93.1
kicking something	80.8	91.7	92.4
Selected Worst Classes / Accu.	**Base Model**	**TSSI**	**TSSI + GLAN**
writing	52.2	26.5	39.7
reading	25.6	26.0	39.9
clapping	17.2	36.6	39.7
playing with phone	31.6	43.6	56.0
Overall	68.0	73.1	80.1

Best Classes Stat. / Accu.	Base Model	TSSI	TSSI + GLAN
Top 1	96.0	99.3	97.8
Top 3 Avg.	93.6	96.1	96.8
Top 5 Avg.	92.0	94.3	96.2
Top 10 Avg.	87.3	92.0	94.8
Worst Classes Stat. / Accu.	**Base Model**	**TSSI**	**TSSI + GLAN**
Top 1	17.2	25.3	39.7
Top 3 Avg.	23.8	25.9	45.2
Top 5 Avg.	27.9	31.6	49.5
Top 10 Avg.	39.6	42.2	56.4
Overall	68.0	73.1	80.1

[3] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[4] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[5] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *arXiv preprint arXiv:1705.07750*, 2017.

[7] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[8] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *arXiv preprint arXiv:1705.01583*, 2017.

[9] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*. IEEE, 2012, pp. 14–19.

[10] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[11] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[12] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.

[13] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833.

[14] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," *arXiv preprint arXiv:1704.02581*, 2017.

[15] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," *arXiv preprint arXiv:1703.03492*, 2017.

[16] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," *arXiv preprint arXiv:1704.04516*, 2017.

[17] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[18] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[19] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1012–1020.

[20] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3d action recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 1647–1656.

[21] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data." in *AAAI*, 2017, pp. 4263–4270.

[22] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," *arXiv preprint arXiv:1612.05877*, 2016.

[23] J. Weng, C. Weng, and J. Yuan, "Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition."

[24] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang, "Action-attending graphic neural network," *arXiv preprint arXiv:1711.06427*, 2017.

[25] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," *arXiv preprint arXiv:1711.05941*, 2017.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[27] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.

[28] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[29] W. Du, Y. Wang, and Y. Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3725–3734.

[30] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Int. Conf. on Computer Vision*, 2017.

[31] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," *arXiv preprint arXiv:1704.06904*, 2017.

[32] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions." in *AAAI*, 2017, pp. 231–237.

[33] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," *arXiv preprint arXiv:1703.10631*, 2017.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[36] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 28–35.

[37] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[38] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2248–2255.