Action Recognition with Spatio-Temporal Visual Attention on Skeleton Image Sequences

Zhengyuan Yang, *Student Member, IEEE,* Yuncheng Li, Jianchao Yang, *Member, IEEE,*
and Jiebo Luo, *Fellow, IEEE*

***Abstract*—Action recognition with 3D skeleton sequences became popular due to its speed and robustness. The recently proposed Convolutional Neural Networks (CNN) based methods shown good performance in learning spatio-temporal representations for skeleton sequences. Despite the good recognition accuracy achieved by previous CNN based methods, there existed two problems that potentially limit the performance. First, previous skeleton representations were generated by chaining joints with a fixed order. The corresponding semantic meaning was unclear and the structural information among the joints was lost. Second, previous models did not have an ability to focus on informative joints. The attention mechanism was important for skeleton based action recognition because different joints contributed unequally towards the correct recognition. To solve these two problems, we proposed a novel CNN based method for skeleton based action recognition. We first redesigned the skeleton representations with a depth-first tree traversal order, which enhanced the semantic meaning of skeleton images and better preserved the associated structural information. We then proposed the general two-branch attention architecture that automatically focused on spatio-temporal key stages and filtered out unreliable joint predictions. Based on the proposed general architecture, we designed a Global Long-sequence Attention Network (GLAN) with refined branch structures. Furthermore, in order to adjust the kernel's spatio-temporal aspect ratios and better capture long term dependencies, we proposed a Sub-Sequence Attention Network (SSAN) that took sub-image sequences as inputs. We showed that the two-branch attention architecture could be combined with the SSAN to further improve the performance. Our experiment results on the NTU RGB+D dataset and the SBU Kinetic Interaction dataset outperformed the state-of-the-art. The model was further validated on noisy estimated poses from the subsets of the UCF101 dataset and the Kinetics dataset.**

***Index Terms*—Action and Activity Recognition, Video Understanding, Human Analysis, Visual Attention.**

I. Introduction

THE major modalities used for action recognition include RGB videos [1], [2], [3], optic flow [4], [5], [6] and skeleton sequences. Compared to RGB videos and optic flow, skeleton sequences are computationally efficient. Furthermore, skeleton sequences have a better ability to represent dataset-invariant action information since no background context is included. One limitation is that manually labeling skeleton sequences is too expensive, while automatic annotation methods may yield inaccurate predictions. Given the above advantages and the fact that skeletons can now be more reliably predicted [7], [8], [9], 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.

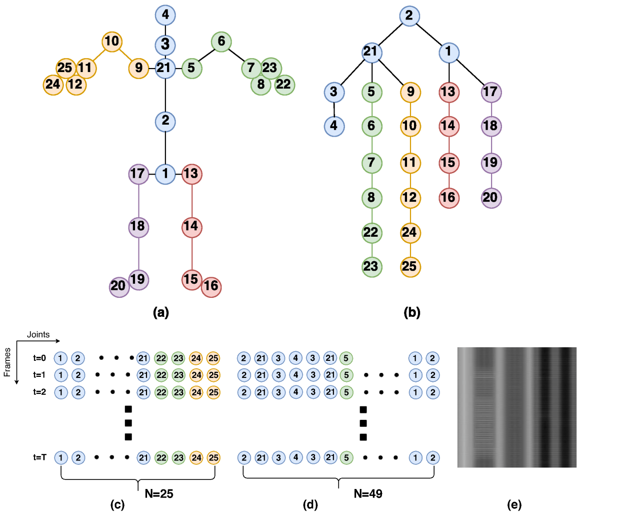


Fig. 1. Tree Structure Skeleton Image (TSSI): (a) The example skeleton structure and order in NTU RGB+D, (b) One possible skeleton tree for TSSI generating, (c) Joint arrangements of naive skeleton images, (d) Joint arrangements of TSSI based on the shown skeleton tree, and (e) An example frame of TSSI.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

With a strong ability of modeling sequential data, Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) neurons outperform the previous hand-crafted feature based methods [10], [11]. 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 [12], [13] that feeds 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 [14] is proposed to learn the spatial and temporal relations simultaneously. Furthermore, a two-stream RNN structure [15] is proposed to learn the spatio-temporal relations with two RNN branches.

CNN has a natural ability to learn representations from 2D

Z. Yang and J. Luo are with the Department of Computer Science, University of Rochester, Rochester, NY, 14627 USA (e-mail: zyang39@cs.rochester.edu; jluo@cs.rochester.edu).

Y. Li is with Snap Inc., Venice, CA, 90291 USA (e-mail: yuncheng.li@snapchat.com).

J. Yang is with Toutiao AI Lab (e-mail: jcyangenator@gmail.com).

Corresponding author: Jiebo Luo.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2864148, IEEE Transactions on Circuits and Systems for Video Technology

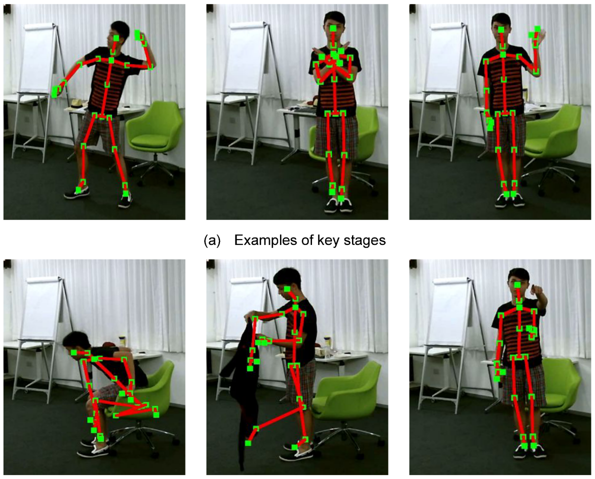

(a) Examples of key stages
(b) Examples of inaccurate keypoint predictions

Fig. 2. Examples of temporal key stages and inaccurate keypoint predictions on NTU RGB+D. The actions in (a) are: throwing, crossing hands in front and hand waving. Keypoint errors in (b) could lead to incorrect action predictions.

arrays. The works in [16], [17] 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. The 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 as shown in Figure 2. Therefore, it is important to include attention mechanisms. Attention masks [18], [19] learned from natural images are 2D weight matrices that amplify the visual features from the regions of importance and depress the others. Similarly, the idea of learning attention masks can be adopted on 'skeleton images'. The skeleton image representation has a natural ability to represent spatio-temporal importance jointly with 2D attention masks, where each row represents the spatial importance of key joints and each column represents the temporal importance of key frames. Based on this, we propose a two-branch architecture for visual attention on single skeleton images. One branch of the architecture is designed with a larger receptive field and generates the predicted attention mask. The other branch maintains and refines the CNN feature. We first introduce the two-branch architecture with a base attention model. A Global Long-sequence Attention Network (GLAN) is then 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. The GLAN alone achieves an accuracy of 80.1% on NTU RGB+D and 95.6% on SBU Kinect Interaction, compared to the 79.6% and 93.6% reported in CNN+MTLN [16].

Despite the effectiveness of the two-branch attention structure, representing an entire sequence as one skeleton image lacks the ability to adjust kernels' spatio-temporal resolutions and learn the long-term dependencies. The original resolutions are determined by the number of joints and the length of the sequence. Furthermore, there is information loss with a sequence longer than the height of the skeleton image. Therefore, we represent the skeleton sequence as several overlapped sub skeleton images and propose a Sub-Sequence Attention Network (SSAN) based on the representation. Furthermore, we show that the GLAN can be combined with the SSAN to further improve the performance.

Our main contributions include the following:

- We propose Tree Structure Skeleton Image (TSSI) to better preserve the spatial relations in skeleton sequences. TSSI improves the previous concatenation skeleton image generation with the depth-first tree traversal.
- 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.
- We propose a Sub-Sequence Attention Network (SSAN) to adjust spatio-temporal aspect ratios and better learn long-term dependencies. We further show that the GLAN and the SSAN can be well combined.
- The proposed method is compatible with both 3D skeletons and 2D poses. We evaluate the model on both Kinect recorded skeletons and noisy estimated poses. Experiments prove the effectiveness of the method and the robustness against noisy inputs.

This paper contains published contents from an early conference paper [20]. The key differences include the followings. The conference version discusses visual attention on a single skeleton image. Although the attention framework is effective, it lacks the ability to adjust the spatio-temporal resolution along the two axis of skeleton images. In this paper, we propose to split a single skeleton image into a skeleton image sequence and extend the GLAN with a temporal attention network. Experiments show the effectiveness of the proposed model and the importance of a proper spatio-temporal resolution. Furthermore, the proposed model is evaluating on estimated pose data with self-paced learning techniques. Finally, more experiment results are included.

II. Related Work

Historically, RGB video action recognition [21] consisted of a feature extraction stage, a feature encoding stage and a classification stage. Hand-crafted features including HOG [22], MBH [23] and etc were designed to represent activity features. Various feature encoding methods [24], [25] were also studied. With the developments of deep neural networks [26], [27], more deep models were designed to solve the action recognition problem. Adopting ConvNets to encode frame-level feature and using LSTM to learn the temporal evolution had been proved to be an effective approach [1], [28], [29]. Optical flow [4], [5], [6] was another way of representing temporal information. Furthermore, C3D [2] was proposed to learn the spatial and temporal information simultaneously with 3D convolutional kernels.

Compared to other frequently used modalities including RGB videos [1], [2], [3] and optical flow [4], [5], [6], skeleton sequences required much less computation and were more robust across views and datasets. With the advanced methods to acquire reliable skeletons from RGBD sensors [7] or even a single RGB camera [8], [30], [31], [32], skeleton-based action recognition became increasingly popular. Many previous skeleton-based action recognition methods [33] modeled the temporal pattern of skeleton sequences with Recurrent Neural Networks (RNN). Hierarchical structures [12], [13] better represented the spatial relations between body parts. Other works [34], [35] adopted attention mechanisms to locate spatial key joints and temporal key stages in skeleton sequences. Liu et al. [14] proposed a 2D LSTM network to learn spatial and temporal relations simultaneously. Wang et al. [15] modeled spatio-temporal relations with a two-stream RNN structure. Other effective approaches included lie groups [11], [36] and nearest neighbor search [37]. Recently, graphical neural networks [38], [39] achieved the state-of-the-art performance on the skeleton based recognition. A performance summary on two frequently used datasets NTU RGB+D [13] and SBU Kinect Interaction [40] was shown in Table I.

As shown in Table I, the recently proposed CNN based approaches showed a better performance in learning skeleton representations compared to RNN based methods. Ke et al. [16] proposed to convert human skeleton sequences into gray scale images, where the joint coordinates were represented by the intensity of pixels. Liu et al. [41] proposed to generate skeleton images with 'Skepxels' to better represent the joint correlations. In this paper, we further improved the design of skeleton images with a depth-first traversal on skeleton trees.

TABLE I
The performance summary of the state-of-the-art on skeleton based action recognition.

Methods	Approach	NTU RGB+D Cross Subject	SBU Kinect Interaction
Two-stream RNN [15]	RNN	71.3%	94.8%
Ensemble TS-LSTM [33]	RNN	76.0%	-
Clips+CNN+MTLN [16]	CNN	79.6%	93.6%
Skepxels [41]	CNN	81.3%	-
GLAN [20]	CNN	80.1%	95.6%
A^{2}GNN [38]	Graphic NN	72.7%	-
ST-GCN [39]	Graphic NN	81.5%	-

Attention mechanisms were important for skeleton based action recognition. Previous LSTM based methods [34], [35] learned attention weights between the stacked LSTM layers. For CNN based methods, we proposed that general visual attention can be directly adopted to generate 2D attention masks, where each row represented the spatial importance and each column represented the temporal importance. Visual attention had achieved successes in many areas, including image captioning [18], [42], RGB based action recognition [19], [43], image classification [44], [45], sentiment analysis [46] and etc. Many visual attention methods took an image sequence as input [19], or used extra information from another modality like text [18], [42], [43]. Because a single skeleton image already represented a spatio-temporal sequence and there was no need for an extra modality, we proposed a single frame based visual attention structure.

III. Methodology

In this section, we first introduced the previous design of skeleton images and the base CNN structure, before an improved Tree Structure Skeleton Image (TSSI) was proposed. Later, we proposed the idea of two-branch visual attention and introduced a Global Long-sequence Attention Network (GLAN) based on the idea. Finally, we introduced the Sub-Sequence Attention Network (SSAN) to learn long-term dependencies.

A. Base Model

In CNN based skeleton action recognition, joint sequences were arranged as 2D arrays that were processed as gray scale images. We named such a generated image the 'Skeleton Image'. For a channel in skeleton images, each row contained the chaining of joint coordinates at a certain time-stamp. Each column represented the coordinates of a certain joint throughout the entire video clip. The chain order of joints was pre-defined and fixed. A typical arrangement of the 2D array was shown in Figure 1 (c). The generated 2D arrays were then scaled into 0 to 255, and resized into a fixed size of $224 * 224$. The processed 2D arrays were processed as gray scale images, where each channel represented an axis of joint coordinates. The skeleton images were then fed into CNNs for action recognition. ResNet-50 [47] was adopted as the base ConvNet model. Compared to RNN based or graph neural network based method, CNN based methods could better learn the spatio-temporal relations among joints.

B. Tree Structure Skeleton Image

A shortcoming in the previous skeleton images was that each row was arranged by an improper fixed order. Each row contained the concatenation of all joints with a pre-defined chain order. CNN had a feature that the receptive field grows larger at higher levels. Therefore, the adjacent joints in each row or column were learned first at lower levels. This implied that the adjacent joints shared more spatial relations in original skeleton structure, which did not hold frequently. In previous skeleton images, a generated array had 25 columns representing the joint coordinates 1 to 25 with a joint index shown in Figure 1 (a). An arrangement of the skeleton image was shown in Figure 1 (c). In this case, a convolutional kernel would cover joints [20, 21, 22, 23, 24] at a certain level since these joints were adjacent in skeleton images. However, these joints had less spatial relations in original skeleton structures and should not be learned together at lower levels.

To solve this problem, we proposed a Tree Structure Skeleton Image (TSSI) inspired by a recent LSTM based study [14]. The basic assumption was that the spatially related joints in original skeletons had direct graph links between them. The less edges required to connect a pair of joints, the more related was the pair. The human structure graph was defined with semantic meanings as shown in 1 (a). In the proposed TSSI, the direct concatenation of joints was replaced by a depth-first tree traversal order. One possible skeleton tree was defined in Figure 1 (b) and the corresponding arrangement of TSSI was shown in Figure 1 (d). Based on the shown skeleton tree, the depth-first tree traversal order for each row was [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]. It is worth noticing that other torso nodes could also be selected as the root of skeleton trees and the corresponding traversal order would be different. Using different torso nodes (node 1, 2 or 21) as tree roots generated a maximum accuracy difference of 0.6% on NTU RGB+D with the cross subject setting.

With the depth-first tree traversal order, the neighboring columns in skeleton images were spatially related in original skeleton structures. This proved that the TSSI better preserved the spatial relations. With TSSI, the spatial relations between related joints were learned first at lower levels of CNN and the relations between less relevant joints were learned later at high levels when the receptive field became larger. An example of the generated TSSI was shown in Figure 1 (e).

C. Attention Networks

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

To alleviate the effect of data noises and to focus on informative stages, skeleton based methods should adjust weights for different inputs automatically. We proposed the idea of two-branch visual attention and further designed a Global Long-sequence Attention Network (GLAN) based on the idea. In this section, we first introduced the basic idea of the two-branch attention architecture with a base attention model. Then the structure of Global Long-sequence Attention Network (GLAN) was introduced.

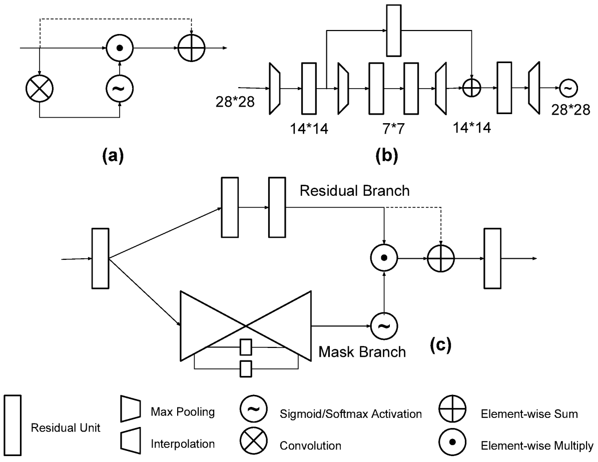


Fig. 3. A base attention module and a GLAN module: (a) A base attention block, (b) An expanded plot for the Hourglass mask branch in GLAN, (c) An attention block with GLAN structure, short for 'GLAN block'.

Base Attention Model. Skeleton images naturally represented both spatial and temporal information of skeleton sequences. Therefore a 2D attention mask could represent spatio-temporal importance simultaneously, where the weights in each row represented the spatial importance of joints and the weight in each column represented the temporal importance of frames. In order to generate the attention masks, we proposed a two-branch attention architecture that learned attention masks from a single skeleton image. The two-branch structure contained a 'mask branch' and a 'residual branch'. Taking previous CNN feature blocks as inputs, the mask branch learned a 2D attention mask and the residual branch refined previous CNN feature. The two branches were then merged to generate the weighted CNN feature block. To be specific, the mask branch learned attention masks with structures that had larger receptive fields. The residual branch was designed to maintain and refine the input CNN features. The two branches were fused at the end of each attention block with element-wise multiplication and summation.

We first introduced the base attention model, which was the simplest version of two-branch attention structures. As shown in Figure 3 (a), the mask branch in the base model gained a larger receptive field with a single convolutional layer. Softmax or Sigmoid functions were used for mask generation. The residual branch preserved the input CNN features with a direct link. The 'attention block' was defined as a module with one mask branch and one residual branch as in Figure 3 (a). The whole framework was built by mixing the proposed attention blocks with the convolutional blocks from ResNet. In the base attention model, attention blocks were 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 improved the designs of both branches to learn attention masks and CNN features more effectively. Inspired by the hourglass structure

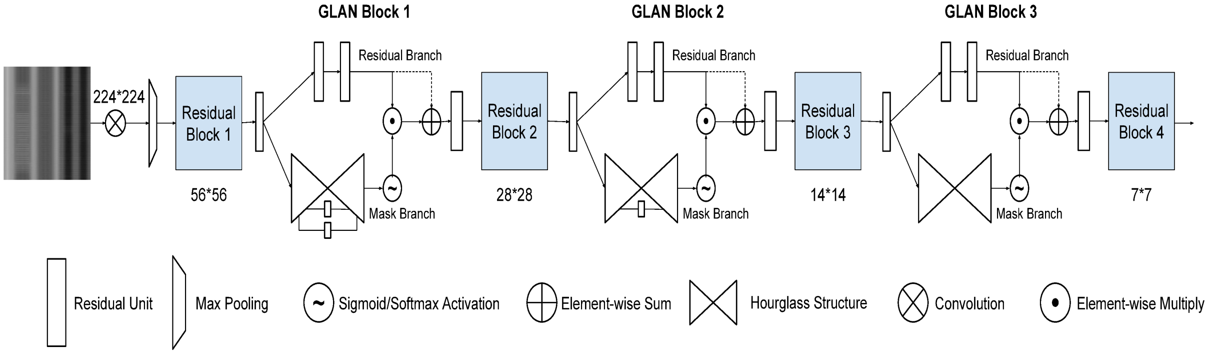


Fig. 4. The structure of the Global Long-sequence Attention Network (GLAN).

[48], [45], we proposed a Global Long-sequence Attention Network (GLAN) as shown in Figure 4. The hourglass structure was adopted in mask branches to quickly adjust the feature size and efficiently gain a larger receptive field. As shown in Figure 3 (b), the hourglass structure consisted of a series of down-sampling units followed by up-sampling units. In each hourglass mask branch, input CNN features were first down-sampled to the lowest spatial resolution of $7 * 7$ and recovered back to the original size. Max pooling was used for down-sampling and bilinear interpolation was used for up-sampling. Each down-sampling unit included a max pooling layer, a followed residual unit and a link connection to the recovered feature with a same size. Each up-sampling unit contained a bilinear interpolation layer, a residual unit and a element-wise sum with the link connection. We showed that the Convolution-Deconvolution structure gained a large receptive field effectively and therefore could better learn an attention mask. For residual branches, we added two residual units to further refine the learned CNN features. All residual units were the same as ResNet [47], which contains three convolutional units and a direct residual link.

As shown in Figure 4, three GLAN attention blocks were added between the four residual blocks in ResNet-50 to build the GLAN network. The depth of each GLAN blocks varied due to the different input feature sizes. Furthermore, we reduced the number of residual units in each residual block to keep a proper depth of the GLAN network, since GLAN blocks were much deeper than the base attention blocks. Only one residual unit was kept for the first three residual blocks. The final residual block kept all three residual units as in ResNet-50.

D. Long-term Dependency Model

Although single-frame-based two-branch attention structure achieved a good performance, it lacked the ability to learn long-term dependences. The generated skeleton image had a fix height of 224. This implied an information loss with a sequence longer than 224 frames, which is around 7 seconds with a frame rate of 30 fps. To better learn long-term dependencies, we proposed a CNN + LSTM model with sub skeleton image sequences. We first split skeleton sequences into several overlapped sub-sequences and generated a series of sub skeleton images for a skeleton sequence. CNN features were first extracted from each sub-sequence skeleton image and the long-term dependencies were modeled with RNNs.

Furthermore, in the original two-branch attention structures, both the spatial and temporal resolutions in skeleton images were fixed by the number of joints and the length of the sequence. However, the kernel should be able to adjust the number of joints and frames it jointly looked at to achieve the best performance. The proposed sub-image model could adjust the relative resolution flexibly by adjusting the number of sub-images and the overlapping rate. This adjustment was equivalent to adjusting the width and height of CNN kernels, while it did not require retraining the model for each dataset.

Sub-Sequence Attention Network (SSAN). To further improve the performance of the proposed sub-image model, an long-term attention module was adopted. Inspired by [19], [18], a Sub-Sequence Attention Network (SSAN) was proposed with a structure shown in Figure 5. Long Short-Term Memory (LSTM) was adopted as the RNN cells. The LSTM implementation was based on [49], [18]:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{d+D,4d} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \tag{1}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{2}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3}$$

where i_t, f_t, o_t, c_t, h_t were the input gates, forget gates, output gates, cell states and hidden states of the LSTM. g_t was an intermediate representation for updating cell states c_t. T was an affine transformation, where D was the depth of the CNN feature block and d was the dimension of all LSTM states. x_t was the weighted CNN feature that input to the LSTM at time t with length D.

Based on the LSTM model, the 2D attention map l_t at time t was defined as a $K*K$ mask, where K was the output width and height of the CNN feature block:

$$l_{t,i} = \frac{\exp(W_i^\top h_{t-1})}{\sum_{j=1}^{K*K} \exp(W_j^\top h_{t-1})} \quad i \in 1 \dots K^2 \tag{4}$$

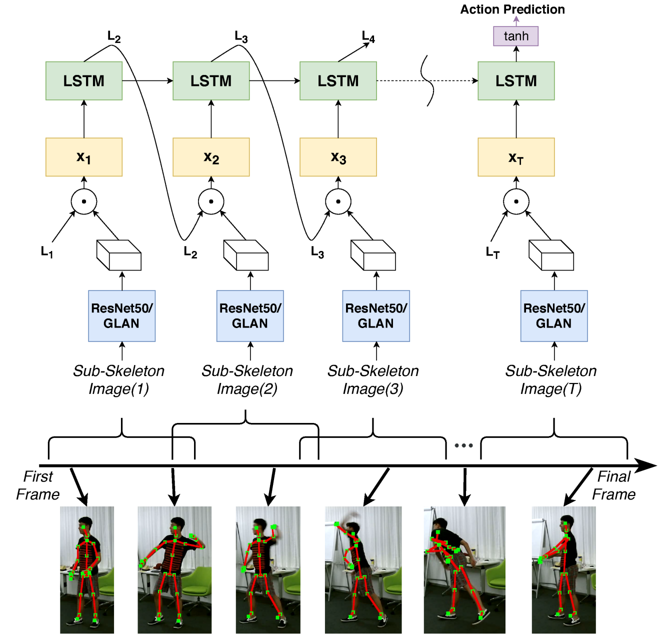


Fig. 5. The structure of the Sub-Sequence Attention Network (SSAN).

Inspired by [45], we also adopted the spatial-channel attention with sigmoid activation, where $i \in 1 \dots K^2, z \in 1 \dots D$.

$$l_{t,i,z} = Sigmoid(W_{i,z}^{\top} h_{t-1}) \tag{5}$$

The weighted CNN feature x_t at time t was the element-wise multiplication of attention mask l_t and original CNN output X_t following Equation 6. In the SSAN, Resnet-50 was selected as the CNN model.

$$x_t = \sum_{j=1}^{K^2} l_{t,i} X_{t,i} \tag{6}$$

GLAN + SSAN. Furthermore, we showed that the GLAN could replace Resnet-50 as the CNN structure in the long-term dependency model. The combination of SSAN and GLAN enabled the framework to generate attentions both in CNN layers with a bottom-up approach and in LSTM layers with a top-down approach. Experiments showed the effectiveness of the proposed combination, and further proved the possibility of using the proposed modules as atomic parts in other frameworks.

IV. Experiments

The proposed method is evaluated on both clean datasets captured by Kinect and noisy datasets where the poses are estimated from RGB videos. We adopt the NTU RGB+D dataset [13] and the SBU Kinect Interaction Dataset [40] for clean dataset evaluation. The estimated poses on UCF101 [50] and Kinetics [51] are used to measure the performance with potentially incomplete and noisy poses. We further evaluate the effectiveness of each proposed module separately. The experiments show that both the TSSI and the attention network generate a large boost in the action recognition accuracy to outperform the state-of-the-art.

A. Datasets

NTU RGB+D. The NTU RGB+D dataset [13] 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 [13]. Samples with missing joints are discarded as in that paper.

SBU Kinect Interaction. The SBU Kinect Interaction dataset [40] 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. During training, random cropping is applied for data augmentation. Prediction scores from the five crops in center and four corners are averaged as testing prediction.

Kinetics-Motion. The Kinetics dataset [51] is the largest RGB action recognition dataset, containing around 300,000 video clips from 400 action classes. The videos are collected from YouTube and each clip is around 10 seconds long. To conduct joints based action recognition, we use the pre-calculated estimated poses provided by [39]. Videos are first converted into a fixed resolution of 340×256 with a frame rate of 30 FPS, and poses are then estimated with the OpenPose toolbox [31]. Because the joint sequence contains no background context or object appearance, it fails to distinguish certain classes defined in RGB action recognition datasets. To better evaluate skeleton based methods on estimated joints, Yan et al. [39] proposes a 'Kinetics-Motion' dataset, which is a 30-class-subset of Kinetics with action labels strongly related to body motion. We evaluate the proposed method on the Kinetics-Motion dataset. The selected 30 classes are: *belly dancing, punching bag, capoeira, squat, windsurfing, skipping rope, swimming backstroke, hammer throw, throwing discus, tobogganing, hopscotch, hitting baseball, roller skating, arm wrestling, snatch weight lifting, tai chi, riding mechanical bull, salsa dancing, hurling (sport), lunge, skateboarding, country line dancing, juggling balls, surfing crowd, dead lifting, clean and jerk, crawling baby, push up, front raises, pull ups*.

UCF101-Motion. The UCF101 dataset [50] contains 13,320 videos from 101 action categories. Videos have a fixed frame rate and resolution of 25 FPS and 320×240. Using RGB videos as inputs, we estimate 16-joint-poses with AlphaPose toolbox [52]. The toolbox provides 2D joint locations and the confidence values for predictions. Similar to Kinetics-Motion, we argue the problem also exists on UCF101 that certain pre-defined action classes such as 'cutting in kitchen' are more relevant to objects and scenes. To prove this, we follow the procedure in ST-GCN [39] and propose a subset from UCF-101 named 'UCF-Motion'. UCF-Motion contains 23 classes that are strongly related to body motions with 3172 videos in total. The selected classes are: *playing dhol, clean and jerk, writing on board, playing flute, playing cello, playing guitar, bowling, ice dancing, playing piano, punch, playing* *tabla, soccer juggling, tai chi, boxing-speed bag, salsa spins, jump rope, boxing-punching bag, hammer throw, rafting, push ups, juggling balls, golf swing, baby crawling*.

B. Ablation Studies

To prove the effectiveness of the TSSI and the proposed attention networks, we separately evaluate each proposed module with results shown in Table III. Each component of the framework is evaluated on NTU RGB+D with the cross subject setting. NTU RGB+D is selected for component evaluations because it is 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 [16] 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 [53], PennAction [54] and NTU RGB+D [13]. 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 experiment shows that even the simplest two-branch attention network can improve the recognition accuracy.

TSSI + GLAN. 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 4, the recognition accuracy increases from 73.1% to 80.1% compared to TSSI without attention mechanisms.

TSSI + SSAN. The SSAN is one of the two attention networks we proposed. The number of sub-sequences and the overlapping rate for the sub-sequences are two hyper-parameters that are tuned with validation set. With a sub-sequence number of 5 and an overlapping rate of 0.5, the attention network achieves an accuracy of 80.9% from 73.1% compared to the base TSSI structure.

TSSI + GLAN + SSAN. Individually, the GLAN and SSAN san achieve a similar improvement in recognition accuracy. Moreover, we show that the GLAN and SSAN can be well combined to further improve the performance. By replacing the Resnet-50 with the proposed GLAN, the framework achieved an accuracy of 82.4%. This experiment also shows that the proposed two branch attention structure can be adopted as atomic CNN structure in various frameworks to achieve a better performance.

TABLE II
GLAN + SSAN PERFORMANCE WITH DIFFERENT HYPER PARAMETERS ON THE NTU RGB+D DATASET.

Sub-image Lengths	Sub-image Numbers	Overlapping Rate	Accuracy (%)
$T/3$	3	0%	80.80
$T/3$	5	50%	82.42
$T/3$	9	75%	81.38
$T/5$	5	0%	78.56
$T/4$	5	25%	80.30
$T/3$	5	50%	82.42
$T/2$	5	75%	81.57

Furthermore, we analyze the hyper-parameters in the SSAN, i.e. the overlapping rate, the number and length of sub-images. The relation of these parameters is:

$$T = t_{sub} * [1 + (1 - \alpha) * (n - 1)] \tag{7}$$

where t_{sub} is the number of frames in each sub-image or the sub-image length, T is the number of frames in the whole sequence, α is the overlapping rate and n is the number of sub-images. We design two sets of experiments with fixed sub-image lengths or fixed sub-image numbers to interpret the effectiveness of the SSAN and find the best set of hyper parameters. Experiments are conducted on NTU RGB+D with the TSSI + GLAN + SSAN framework.

Starting from the optimal hyper-parameters of a 50% overlapping rate and 5 sub-images, we report the performances under different hyper parameters with either sub-image numbers or sub-image lengths unchanged. For the fixed length experiment as shown in Table II, the length of sub-images are fixed and the number of sub-images changes from 3 to 9 by adjusting the overlapping rate. We observe the accuracy drops 1.6% from 82.4% to 80.8%. In the fixed sub-image number experiment, the number of sub-images is fixed as five where the best performance is achieved. The length of sub-images varies from $T/5$ to $T/2$ with different overlapping rates. A larger drop of accuracy of 3.8% is observed in the fixed sub-image number experiment.

According to the experiment results, the length of sub-images influence the performance of the SSAN most. This implies that the SSAN produces a large boost in accuracy mainly with its ability to flexibly adjusting the spatial-temporal resolutions. The optimal hyper parameters of a 50% overlapping rate, 5 sub-images and the $T/3$ temporal length works best with the 25 joints on NTU RGB-D. The optimal hyper parameters vary on different datasets with different averaged sequence lengths and joint numbers. Furthermore, the SSAN also better learns the long term dependencies. Most results with the SSAN as shown in Table II outperform the methods with single skeleton frames such as TSSI + GLAN.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2864148, IEEE Transactions on Circuits and Systems for Video Technology

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

State-of-the-art	Cross Subject	Cross View
Lie Group [11]	51.0	52.8
HBRNN [12]	59.1	64.0
Part-aware LSTM [13]	62.9	70.3
Trust Gate LSTM [14]	69.2	77.7
Two-stream RNN [15]	71.3	79.5
TCN [17]	74.3	83.1
Global Attention LSTM [34]	74.4	82.8
A^{2}GNN [38]	72.7	82.8
Clips+CNN+MTLN [16]	79.6	84.8
Ensemble TS-LSTM [33]	76.0	82.6
Skepxels [41]	81.3	**89.2**
ST-GCN [39]	81.5	88.3
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 [20]	80.1	85.2
TSSI + SSAN	80.9	86.1
TSSI + GLAN + SSAN	**82.4**	89.1

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

State-of-the-art	Accuracy (%)
Raw Skeleton [40]	49.7
HBRNN [12]	80.4
Trust Gate LSTM [14]	93.3
Two-stream RNN [15]	94.8
Global Attention LSTM [34]	94.1
Clips+CNN+MTLN [16]	93.6
Proposed Model	**Accuracy (%)**
Base Model	82.0
With TSSI	89.2
TSSI + Base Attention	93.6
TSSI + GLAN [20]	95.6
TSSI + SSAN	94.0
TSSI + SSAN + GLAN	**95.7**

C. Evaluations on Clean Datasets

NTU RGB+D. The middle column of Table III shows the results of the NTU RGB+D cross subject setting. The base model with naive skeleton images already outperforms a number of previous LSTM based method, without adopting the 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. The proposed two-branch attention architecture further improves the performance and the GLAN outperforms the state-of-the-art. Experiments prove the effectiveness of the proposed CNN based action recognition method.

Furthermore, we show that generating sub-sequences and adopting the long-term dependency model (SSAN) can achieve a better results. The SSAN with ResNet-50 achieves a cross subject accuracy of 80.9%. By replacing the ResNet with the proposed GLAN to provide the spatial attention, the framework improves the state-of-the-art to 82.4%. Similar results are observed in the NTU RGB+D cross view setting, as shown in the right column of Table III.

SBU Kinect Interaction. Similar to the performance on the NTU RGB+D dataset, the proposed TSSI and attention framework generates a large boost in the recognition accuracy on the SBU Kinect Interaction dataset that outperforms the state-of-the-art. The performances are shown in Table IV. The proposed TSSI+SSAN+GLAN achieves an accuracy of 95.7± 1.7% on the five splits provided by SBU Kinect Interaction.

D. Error Case Analysis

To better understand the successful and failure cases, experiments are conducted to analyze the performance of each class in NTU RGB+D. As shown in Table VII, two parts of analysis are conducted. First, eight classes that constantly perform the best or worst are selected on the left side of Table VII. Results show that the actions with dynamic body movements, such as standing, sitting and walking, can be well classified with skeletons, while the classes with less motion like reading, writing and clapping usually have a poor result. The first, middle and last frames from these classes are visualized in the first row of Figure 6. 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 motion.

The results also show that the proposed TSSI, GLAN and SSAN all generate a large boost in performance in all the listed classes. On the righthand side of the table, statistics of the best and worst classes are listed. Results show that TSSI + GLAN + SSAN greatly improve the accuracy in challenging classes. The top 1 worst class in TSSI + GLAN + SSAN has an accuracy of 42.8%, 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 + SSAN are similar. The improvements are mainly obtained through the improvements in the challenging classes.

E. Self-Paced Learning on Noisy Datasets

The model is then evaluated on large scale RGB datasets with estimated and thus noisy poses. For a fair comparison, we do not use any pre-processing methods like interpolation to reduce the noise. To better learn a noise robust system, we adopt self-paced learning [55], [56], [57], [58] during the training process. The model is first trained with a small portion of reliable pose estimates and then gradually take more noisy data as inputs. The average pose estimation confidence values provided by Openpose is used as the indication of reliability and the level of noises. The model starts with a high confidence threshold of 0.5, i.e. all estimated pose sequences with an average confidence lower than 0.5 are eliminated in the training process. We then fine-tune the model step by step by feeding more unreliable noisy data. Experiments show that self-paced learning can both accelerate the convergence speed and improve the final accuracy.

Kinetics-Motion. As shown in Table V, the proposed long term dependency model with attention is comparable to the state-of-the-art performances. The recognition accuracy also similar to the methods using other modalities including RGB

TABLE V
THE RECOGNITION ACCURACY COMPARED TO THE STATE-OF-THE-ART METHODS ON THE KINETICS-MOTION DATASET.

State-of-the-art	Accuracy (%)
RGB CNN [51]	70.4
Flow CNN [51]	72.8
ST-GCN [39]	72.4
Proposed Model	Accuracy (%)
With TSSI	58.8
TSSI + GLAN [20]	67.2
TSSI + SSAN + GLAN	68.7

TABLE VI
THE RECOGNITION ACCURACY COMPARED TO THE STATE-OF-THE-ART METHODS ON THE UCF-MOTION DATASET.

State-of-the-art	Accu. (%)	RGB	Flow	Keypoints
HLPF [53]	71.4	-	-	✓
LRCN [1]	81.6	✓	-	-
3D-ConvNet [2]	75.2	✓	-	-
Flow CNN [51]	85.1	-	✓	-
Two-Stream [51]	91.3	✓	✓	-
Proposed Model	Accu. (%)			
With TSSI	87.9	-	-	✓
TSSI + GLAN [20]	91.0	-	-	✓
TSSI + SSAN + GLAN	**91.7**	-	-	✓

and optical flow. This experiment proves that the proposed GLAN + SSAN framework is noise robust. The first, middle and last frames from example videos are shown for success and failure cases in the third row of Figure 6. Observations show that failure cases are mostly caused by missing or incorrect pose estimates.

UCF101-Motion. As shown in Table VI, we evaluate the framework on the proposed UCF-Motion dataset. The proposed framework outperforms previous methods that use a single modality [1], [2] or both appearance feature and optical flow [4]. The experiment proves that joint is an effective modality for recognizing motion related actions, although joints alone are insufficient for distinguishing all defined action classes since recognizing certain classes requires object and scene appearances. Furthermore, the recognition accuracy is still limited by the imperfect pose estimations. Example frames are shown in the second row of Figure 6. The compared performances are based on released codes or our reimplementation.

V. CONCLUSIONS

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, i.e., 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 the skeleton image. A Global Long-sequence Attention Network (GLAN) is proposed based on the two-branch attention structure. We further propose the long-term dependency model with a Sub-Sequence Attention Network (SSAN). The effectiveness of combining the GLAN and the SSAN is also validated. Experiments show that the proposed enhancement modules greatly improve the recognition accuracy, especially on the challenging classes. Extended ablation studies prove that the improvements are achieved by better retaining the skeletal information and focusing on informative joints or time stamps. Moreover, the model shows the robustness against noisy estimated poses. Next on our agenda is to further improve the robustness against incomplete and inaccurate estimated poses. Another direction for further exploration is extending TSSI to automatically learn refined joint relations.

ACKNOWLEDGMENTS

The authors would like to 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 *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

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

[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 *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," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4724–4733.

[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," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 44, 2017.

[9] Z. Yang and J. Luo, "Personalized pose estimation for body language understanding," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 126–130.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Conference on Computer Vision and Pattern Recognition*, 2017.

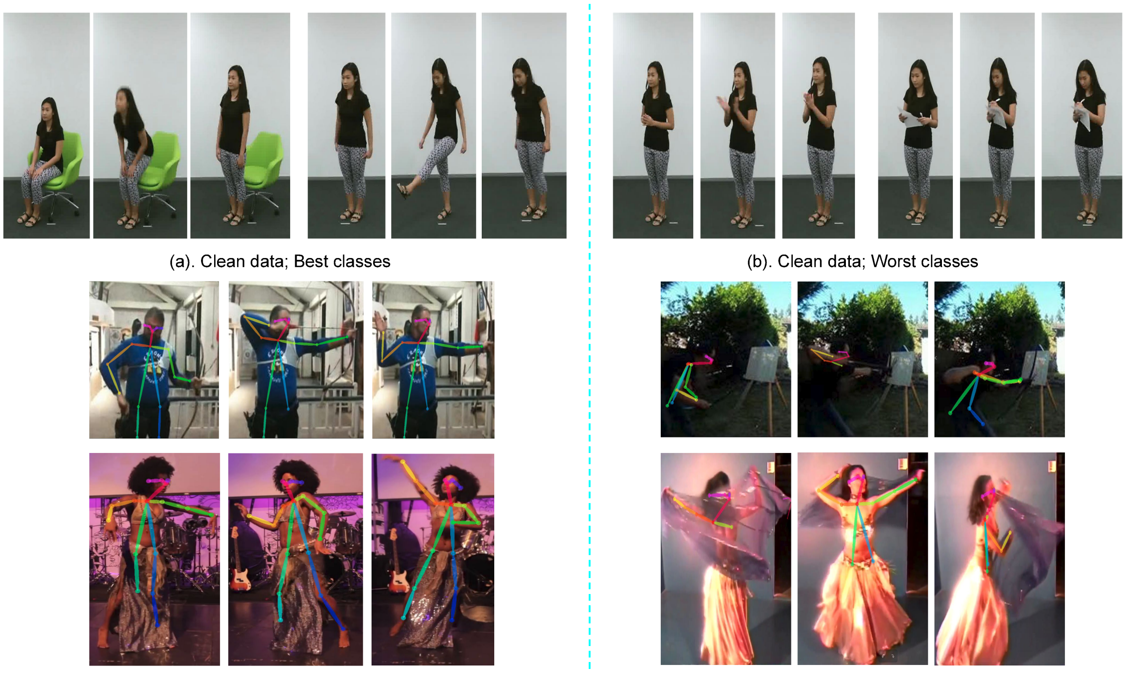
[16] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4570–4579.

TABLE VII

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

Selected Best Classes	Base.	TSSI	GLAN	GLAN + SSAN
standing up	85.4	94.1	**97.1**	96.3
sitting down	91.6	91.6	93.8	**93.8**
walking apart	90.6	91.3	93.1	**96.0**
kicking something	80.8	91.7	92.4	**92.8**
Selected Worst Classes	Base.	TSSI	GLAN	GLAN + SSAN
writing	**52.2**	26.5	39.7	45.6
reading	25.6	26.0	39.9	**42.8**
clapping	17.2	36.6	39.7	**63.0**
playing with phone	31.6	43.6	56.0	**66.2**
Overall	68.0	73.1	80.1	**82.4**

Best Classes Stat.	Base.	TSSI	GLAN	GLAN + SSAN
Top 1	96.0	**99.3**	97.8	97.1
Top 3 Avg.	93.6	96.1	96.8	**96.8**
Top 5 Avg.	92.0	94.3	96.2	**96.6**
Top 10 Avg.	87.3	92.0	94.8	**95.5**
Worst Classes Stat.	Base.	TSSI	GLAN	GLAN + SSAN
Top 1	17.2	25.3	39.7	**42.8**
Top 3 Avg.	23.8	25.9	45.2	**49.9**
Top 5 Avg.	27.9	31.6	49.5	**55.0**
Top 10 Avg.	39.6	42.2	56.4	**61.2**
Overall	68.0	73.1	80.1	**82.4**



(a). Clean data; Best classes

(b). Clean data; Worst classes

(c). Success cases with noisy data

(d). Failure cases with noisy data

Fig. 6. Example frames from clean and noisy datasets. In the first row from left to right contains classes from NTU RGB+D: standing up, kicking something, clapping and writing. The second and third row contains predicted noisy poses of success and failure cases. The second row is from UCF101 and the third row is from Kinetics. Success cases are shown on the left side and the failure cases are shown on the right side.

[17] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1623–1631.

[18] 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.

[19] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," in *NIPS Time Series Workshop*, 2015.

[20] Z. Yang, Y. Li, J. Yang, and J. Luo, "Action recognition with visual attention on skeleton images," in *The International Conference on Pattern Recognition (ICPR)*, 2018.

[21] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*. IEEE, 2009, pp. 1996–2003.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR), 2005.*, vol. 1. IEEE, 2005, pp. 886–893.

[23] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*. Springer, 2006, pp. 428–441.

[24] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *European Conference on Computer Vision*. Springer, 2014, pp. 581–595.

[25] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[28] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning*, 2015, pp. 843–852.

[29] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

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

[31] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, vol. 1, no. 2, 2017, p. 7.

[32] R. Alp Gler, N. Neverova, and I. Kokkinos, "Densepose: Dense human

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2864148, IEEE Transactions on Circuits and Systems for Video Technology

pose estimation in the wild," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] 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.

[34] 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.

[35] 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.

[36] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE computer Society, 2017, pp. 6099–6108.

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

[38] C. Li, Z. Cui, W. Zheng, C. Xu, R. Ji, and J. Yang, "Action-attending graphic neural network," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3657–3670, 2018.

[39] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

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

[47] 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.

[48] 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.

[49] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *CoRR*, vol. abs/1409.2329, 2014.

[50] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.

[51] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[52] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *ICCV*, 2017.

[53] 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.

[54] 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.

[55] J. Luo and A. E. Savakis, "Self-supervised texture segmentation using complementary types of features," *Pattern Recognition*, vol. 34, no. 11, pp. 2071–2082, 2001.

[56] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems*, 2010, pp. 1189–1197.

[57] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 547–556.

[58] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Advances in Neural Information Processing Systems*, 2014, pp. 2078–2086.

Zhengyuan Yang received the BE degree in electrical engineering from the University of Science and Technology of China in 2016. He is currently pursuing the PhD degree with the Computer Science Department, University of Rochester, under the supervision of Prof. Jiebo Luo. His research interests mainly include action recognition, pose estimation and video analysis.

Yuncheng Li received the BE degree in electrical engineering from the University of Science and Technology of China in 2012, and the PhD degree in computer sciences from the University of Rochester in 2017. He is currently a Research Scientist with Snapchat Inc.

Jianchao Yang (M'12) received the MS and PhD degrees from the ECE Department, University of Illinois at UrbanaChampaign, under the supervision of Prof. Thomas S. Huang. He was a Research Scientist with Adobe Research and a Principal Research Scientist with Snapchat Inc. He is currently the director of Toutiao AI Lab in the bay area. He has authored over 80 technical papers over a wide variety of topics in top tier conferences and journals, with over 12,000 citations per Google Scholar. His research focuses on computer vision, deep learning, and image and video processing. He received the Best Student Paper award in ICCV 2010, the Classification Task Prize in PASCAL VOC 2009, first position for object localization using external data in ILSVRC ImageNet 2014, and third place in the 2017 WebVision Challenge. He has served on the organizing committees of the ACM Multimedia Conference in 2017 and 2018.

Jiebo Luo (S'93-M'96-SM'99-F'09) joined the Department of Computer Science, University of Rochester, in 2011, after a prolific career of over 15 years with Kodak Research. He has authored over 400 technical papers and holds over 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media, and biomedical informatics. He has served as the Program Chair of the ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017, and on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BIG DATA, Pattern Recognition, Machine Vision and Applications, and ACM Transactions on Intelligent Systems and Technology. He is a Fellow of the SPIE and IAPR.