

# A Fine-to-Coarse Convolutional Neural Network for 3D Human Action Recognition

Thao Minh Le  
thao@ks.cs.titech.ac.jp

Nakamasa Inoue  
inoue@ks.cs.titech.ac.jp

Koichi Shinoda  
shinoda@c.titech.ac.jp

Department of Computer Science  
Tokyo Institute of Technology  
Tokyo, Japan

---

## Abstract

This paper presents a new framework for human action recognition from a 3D skeleton sequence. Previous studies do not fully utilize the temporal relationships between video segments in a human action. Some studies successfully used very deep Convolutional Neural Network (CNN) models but often suffer from the data insufficiency problem. In this study, we first segment a skeleton sequence into distinct temporal segments in order to exploit the correlations between them. The temporal and spatial features of a skeleton sequence are then extracted simultaneously by utilizing a fine-to-coarse (F2C) CNN architecture optimized for human skeleton sequences. We evaluate our proposed method on NTU RGB+D and SBU Kinect Interaction dataset. It achieves 79.6% and 84.6% of accuracies on NTU RGB+D with cross-object and cross-view protocol, respectively, which are almost identical with the state-of-the-art performance. In addition, our method significantly improves the accuracy of the actions in two-person interactions.

## 1 Introduction

In the past few years, human action recognition has become an intensive area of research, as a result of the dramatic growth of societal applications for a number of areas including security surveillance systems, human-computer-interaction-based games, and healthcare industry. The conventional approach based on RGB data was not robust against intra-class variations and illumination variations. With the advancement of 3D sensing technologies, in particular, affordable RGB-D cameras such as Microsoft Kinect, these problems have been remedied to some extent. Human action recognition studies utilizing 3D skeleton data have drawn a great deal of attention [4, 22].

Human action recognition based on 3D skeleton data is a time series problem, and accordingly, a great body of previous studies have focused on extracting motion patterns from a skeleton sequence. Earlier methods utilized hand-crafted features for representing the intra-frame relationships through the skeleton sequences [24, 30]. Some studies utilized the deep learning, end-to-end learning based on Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) has been utilized to learn the temporal dynamics [0, 24, 15, 24, 26, 36]. Recent studies have shown the superiority of Convolutional Neural

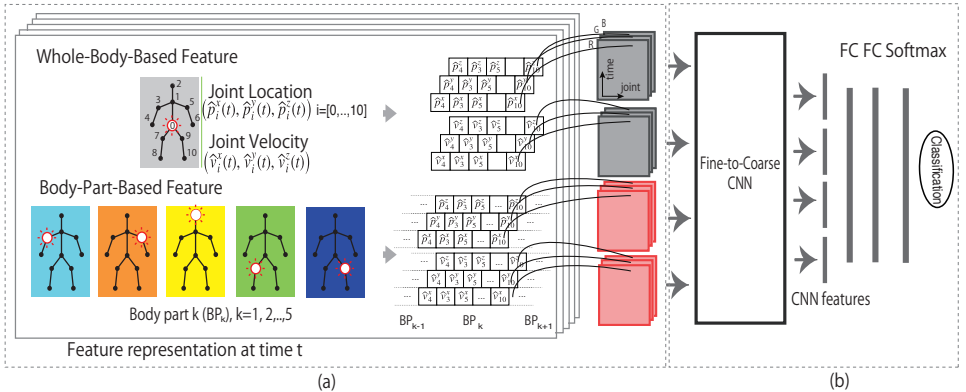


Figure 1: Overview of The Proposed Method. It consists of two parts: (a) feature representation and (b) high-level feature learning with a F2C CNN-based network architecture. A skeleton from a video input sequence is represented by whole-body-based features (WB) and body-part-based features (BP). These features are then transformed into a skeleton image that contains both the spatial structure information of a human body as well as the temporal sequence feature of a human action. The skeleton images are then fed into an F2C convolutional neural network for high-level feature learning. Finally, CNN features are concatenated before being passed to two subsequent fully connected layers, and a soft-max layer for final classification.

Networks (CNNs) over RNN with LSTM for this task [8, 9, 17, 18]. Most of the CNN-based studies encoded the trajectories of human joints in an image space representing the spatio-temporal information of the skeleton data. The encoded feature is then fed into a deep CNN pre-trained on large scale image datasets, for example, ImageNet [23], under the notion of transfer learning [24]. This CNN-based method is, however, weak in handling long temporal sequences. And thus, it usually fails to distinguish actions with similar distance variations but with different durations, such as “handshaking” and “giving something to other persons”.

Motivated by the success of the generative model for CAPTCHA images [4], we believe 3D human action recognition systems can also benefit from a specific network structure for this application domain. The first step is to segment a given skeleton sequence into different temporal segments. Here, we assume that temporal features of different time-steps have different correlations. We further utilize a tailor-made F2C CNN-based network architecture to model high-level features. By utilizing both the temporal relationships between temporal segments and spatial connectivities among human body parts, our method is expected to have a superior performance to the naive deep CNN networks. To the best of our knowledge, this is the first attempt to use F2C network for 3D human action recognition.

The paper is organized as follows. In Section 2, we discuss the related studies. In Section 3, we explain our proposed network architecture in detail. We then show the experimental results to justify our motivations in Section 4. Finally, we conclude our study in Section 5.

## 2 Related Studies

Deep learning techniques drew a great attention in the field of 3D human action recognition. Especially, the end-to-end network architectures can discriminate actions from raw skeleton data without any handcrafted features. Zhu et al. [56] adopted three LSTM layers to exploit the co-occurrence features of skeleton joints at different layers. Du et al. [9] proposed a hierarchical RNN to exploit the spatio-temporal feature of a skeleton sequence. They divided skeleton joints into five subsets corresponding to five body parts before independently feeding them into five bidirectional recurrent neural networks for local feature extraction. The relationships between body parts were then modeled in later layers by hierarchically fusing them together. LSTMs were deliberately used in the last layer to tackle the vanishing problem of a vanilla RNN.

The use of deep learning techniques for this area of research was exploded when NTU RGB+D dataset [24] was released. Shahroudy et al. [24] introduced a part-aware LSTM to learn the long-term dynamics of a long skeleton sequence from multimodal inputs extracted from human body parts. Liu et al. [12], on the other hand, employed a spatio-temporal LSTM (ST-LSTM) to handle both the spatial dependencies and the temporal dependencies. ST-LSTM is also enhanced with a tree-structure based traversal method for transmitting input data of each frame into the network. In addition, this method used a trust gate mechanism to exclude noisy data from the input. Zhang et al. [33] proposed a view adaptation scheme for 3D skeleton data and further integrated it into an end-to-end LSTM network for sequential data modeling and feature extraction.

CNNs are powerful for the task of object detection from images. Transfer learning techniques enable them to perform well even with a limited number of data samples [19, 28]. Motivated by this, Ke et al. [8] was the first to apply transfer learning for 3D human action recognition. They used a VGG model [9] pre-trained with ImageNet to extract high-level features from cosine distance features between joint vectors and their normalized magnitude. Ke et.al 2017b [9] further transformed the cylindrical coordinates of an original skeleton sequence into three clips of gray-scale images. The clips are then processed by pre-trained VGG19 model [25] to extract image features. Multi-task learning was also proposed by [9] for the final classification, which achieved the state-of-the-art performance on NTU RGB+D dataset.

Our study addresses two problems of the previous studies: (1) the loss of temporal information of a skeleton sequence during training and, (2) the need for a specific CNN structure for skeleton data. We believe that a very deep CNN model such as VGG [25], AlexNet [13] or ResNet [5] are overqualified for such sparse data as human skeleton. Moreover, the available skeleton datasets are relatively small compared to image datasets. Thus, we believe a network architecture which is able to leverage the geometric dependencies of human joints is promising for solving this issue.

## 3 Fine-to-Coarse CNN for 3D Human Action Recognition

This section presents our proposed method for 3D skeleton-based action recognition which exploits the geometric dependency of human body parts and the temporal relationship in a time sequence of skeletons (Figure 1). It consists of two phases: feature representation and high-level feature learning with a F2C network architecture.

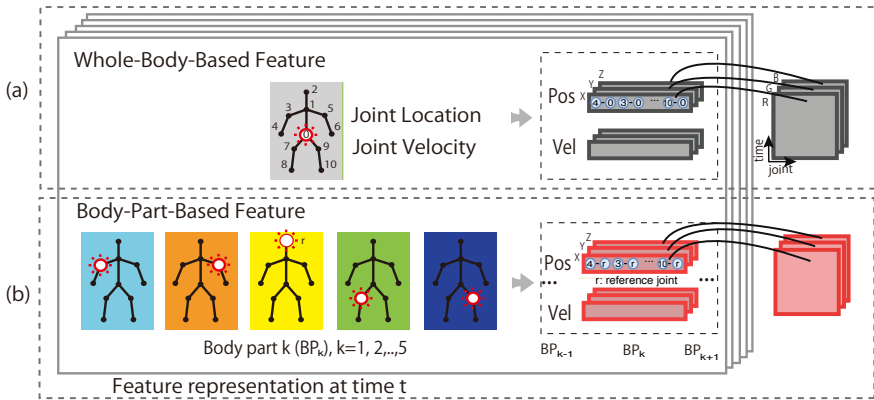


Figure 2: Feature Generation. Figure (a) illustrates the procedure of generating WB features obtained by transforming the joint positions in the camera coordinate system to the hip-based coordinate system. In Figure (b), we arrange BP features side by side to obtain one unique feature 2D array before projecting the coordinates in Euclidean space into RGB image space using a linear transformation and further up-scaling by using cubic interpolation transformation.

### 3.1 Feature Representation

We encode the geometry of human body originally given in an image space into local coordinate systems to extract the relative geometric relationships among human joints in a video frame. We select six joints in a human skeleton as reference joints in order to generate whole-body-based (WB) features and body-part-based (BP) features. The hip joint is chosen as the origin of the coordinate system presenting the WB features, while the other reference joints, namely the head, the left shoulder, the right shoulder, the left hip, and the right hip, are selected exactly the same as [8] to represent the BP features. The WB features represent the motions of human joints around the base of the spine, while the BP features represent the variation of appearance and deformation of the human pose when viewed from different body parts. We believe that the combined use of WB and BP is robust against coordinate transformations.

Different from the other studies using BP features [8, 14, 24], we extract a velocity together with a joint position from each joint of the raw skeleton. The velocity represents the variations over the time and has been widely employed in many previous studies, mostly in the handcrafted-feature-based approaches [11, 52, 54]. It is robust against the speed changes; and accordingly, is effective to discriminate actions with similar distance variations but with different speeds, such as punching and pushing.

In the  $t$ -th frame of sequence of skeletons with  $n$  joints, the 3D position of the  $i$ -th joint is depicted as:

$$p_i(t) = [p_i^x(t), p_i^y(t), p_i^z(t)]^\top. \quad (1)$$

The relative inter-joint positions are highly discriminative for human actions [20]. The relative position of joint  $i$  at time  $t$  is described as:

$$\hat{p}_i(t) = p_i(t) - p_{\text{ref}}(t), \quad (2)$$

where  $p_{\text{ref}}(t)$  depicts the position of a selected reference joint. The velocity feature  $\hat{v}_i(t)$  at time frame  $t$  is defined as the first derivatives of the relative position feature  $\hat{p}_i(t)$ . Zanfir et al. [52] showed that it is effective to compute the derivatives of human instantaneous pose which is represented by joints' location at a given time frame  $t$  over a time segment. The velocity feature, therefore, is formulated as:

$$\hat{v}_i(t) \approx \hat{p}_i(t+1) - \hat{p}_i(t-1). \quad (3)$$

### 3.1.1 Whole-body-based Feature

As mentioned above, we choose the hip joint as the reference joint in order to represent WB features (See figure 2(a)). In addition, we follow the limb normalization procedure [52] to reduce the problem caused by the variations in human body size among human subjects. We first compute the average bone lengths of each two connected joints over the training dataset, and then use them to normalize each human subject's bones. To put it differently, we stretch each bone of a certain human subject with a normalized length while keeping the joint angle between any two bones unchanged.

In order to extract the spatial features of a human skeleton at time  $t$  over the set of joints, we first define a spatial configuration of a joint chain. We believe that the order of joints greatly affects the learning ability of 2D CNN since the joints in adjacent body parts share more spatial relations than a random pair of joints. For example, in most actions, the joints of the right arm are more correlated to those of the left arm than those of the left leg are. With this intention, we concatenate joints in the following order: left arm, right arm, torso, left leg, right leg. Note that the torso in the context of this paper includes the head joint of the human skeleton. Let  $T$  be the number of frames in a given skeleton sequence. In the next step, we compute each feature of skeleton data over  $T$  frames and stack them as a feature row. Consequently, we obtain the WB features of two 2D arrays; each corresponds to the joint location and velocity. Finally, we project these 2D array features into RGB image space using a linear transformation. In particular, each of three components  $(x, y, z)$  of each skeleton joint is represented as one of the three corresponding components  $(R, G, B)$  of a pixel in a color image; by normalizing the  $(x, y, z)$  values to the range 0 to 255. The two sets of color images are further up-scaled by using a cubic spline interpolation. Cubic spline interpolation is a commonly used technique in image processing to minimize the interpolation error [6]. We call these two RGB images as skeleton images.

### 3.1.2 Body-part-based Feature

In order to represent the BP features, we choose five joints corresponding to five human body parts as the reference joints: the head, the left shoulder, the right shoulder, the left hip, and the right hip, as in [8]. They are relatively stable in most actions. Provided that, we calculate joint position features and velocity features for each reference joint in the above order dependently. As a result, with each skeleton at time  $t$ , we obtain five feature vectors of a joint location and five vectors of a velocity corresponding to five distinct reference joints. We then place all BP features side by side to produce one unique row feature and place them along the temporal axis to obtain a 2D array feature. Finally, we apply a linear transformation to represent these array features as RGB images and further up-scale them by using a cubic spline interpolation. After all, we obtain two BP-base skeleton images; one corresponding to the joint location and the other to the velocity from each skeleton sequence. The whole process is illustrated in Figure 2(b).

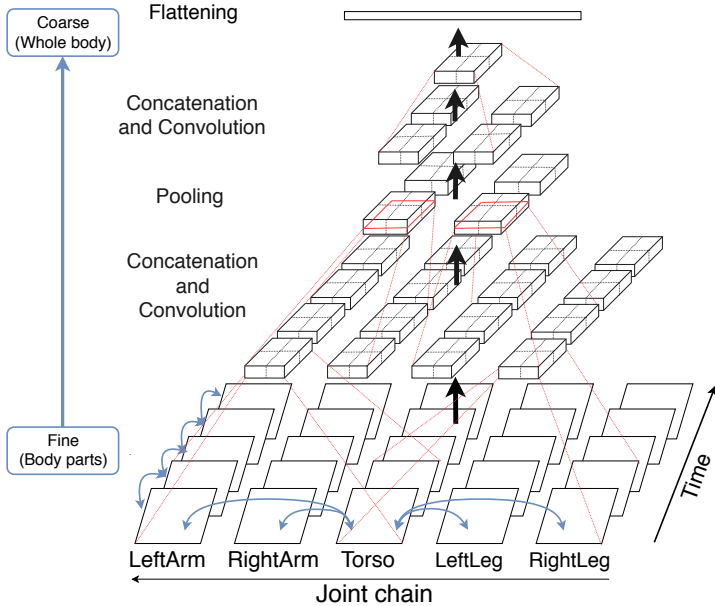


Figure 3: Proposed Fine-to-Coarse Network Architecture. Blue arrows show pair slices which are concatenated along each dimension before passing to a convolutional block.

### 3.2 Fine-to-Coarse Network Architecture

In this section, we explain the detail of our proposed F2C network architecture for high-level feature learning. Figure 3 illustrates our network structure in three dimensions.

Our F2C network takes three color channels of skeleton images generated from the feature representation phase as inputs. Accordingly, the input of our F2C network consists of two dimensions: the spatial dimension which describes the geometric dependencies of human joints along the joint chain, and the temporal dimension of the time-feature representation over  $T$  frames of a skeleton sequence. Let  $m$  be the number of segments along the temporal axis,  $n$  is the number of body parts ( $n = 5$ ), each image skeleton is considered as a set of  $m \times n$  slices (Figure 3). Assume  $T_{\text{seg}}$  ( $T = m \times T_{\text{seg}}$ ) is the number of frames in one temporal segment,  $l_{\text{bp}}$  is the dimension of one body part along the spatial dimension, each input slide has size of  $l_{\text{bp}} \times T_{\text{seg}}$ . In the next step, we simultaneously concatenate the slices over both the spatial axis and temporal axis. In other words, regarding the spatial dimension, we first concatenate each body part which belongs to human limbs (arms and legs) with the torso, while concatenating two consecutive temporal segments together. Each concatenated 2D array feature is further passed through a convolutional layer and a max pooling layer. The same fusion procedure is applied before passing the next convolutional layer. In short, our F2C network composes of three layer-concatenation steps, and three convolutional blocks accordingly. In the last step, the extracted image features are flattened to obtain an output of 1D array feature.

Both WB-based skeleton images and BP-based skeleton images are fed into the proposed F2C network in the same way. While it is conceivable for feeding BP features into our network for high-level feature learning, we believe WB features also benefit from going

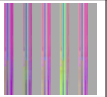

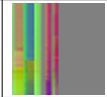

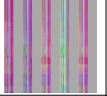

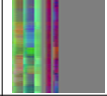

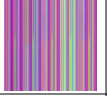

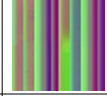

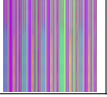

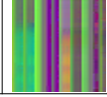

Actions	BP-based skeleton image		WB-based skeleton image	
	Joint Position	Velocity	Joint Position	Velocity
Standing up				
Take off jacket				
Point finger at the other person				
Handshaking				

Figure 4: Examples of Generated Skeleton Images. "Standing up" and "take off jacket" present single actions while "point finger at the other person" and "handshaking" are two-person interaction actions.

through the network since the spatial dimension of WB features, which are formed by the pre-defined joint chain, includes the intrinsic relationships between body parts.

Our network can be viewed as a procedure to eliminate unwanted connections between layers from the conventional CNN. We believe traditional CNN models include some redundant connections for capturing human-body-geometric features. Many actions only require the movement of the upper body (*e.g.* hand waving, clapping) or the lower body (*e.g.* sitting, kicking), while the other requires the movements of the whole body (*e.g.* moving towards another person, pick up something). For this reason, the bottom layers in our proposed method can discriminate "fine" actions which require the movements of some certain body parts, while the top layers are discriminative for "coarse" actions using the movements of the whole body.

## 4 Experiments and Discussion

### 4.1 Datasets and Experimental Conditions

We conduct experiments on two skeleton benchmark datasets publicly available: NTU RGB+D [24] and SBU Kinect Interaction Dataset [51]. As the method proposed by [8] is relatively related to this paper, we employ their method as our baseline. We also compare our proposed method with other state-of-the-art methods reported on the same datasets.

**NTU RGB+D Dataset** is the largest skeleton-based human action dataset for the time being with 56,880 sequences. The skeleton data were collected by utilizing Microsoft Kinect v2 sensors. Each skeleton contains 25 human joints. In this dataset, there are 60 distinct action classes of three human-action groups: daily actions, health-related actions, and two-person interactive actions. All the actions are performed by 40 distinct subjects. The actions are recorded simultaneously by three camera sensors located at different angles:  $-45^\circ$ ,  $0^\circ$ ,  $45^\circ$ . This dataset is challenging due to the large variations of viewpoints and sequence



Table 1: Network Configuration

conv3-64:  $3 \times 3$  convolution, 64 filters

Input of $224 \times 224$ RGB image
35 input slices of $32 \times 44$
24 input slices of $64 \times 88$
conv3-64
maxpool
conv3-64
maxpool
10 fused feature slices of $32 \times 44$
conv3-128
maxpool
conv3-128
maxpool
4 fused feature slices of $16 \times 22$
conv3-256
maxpool
conv3-256
maxpool
output $4 \times 5, 120$

Table 2: Classification Performance on NTU RGB+D Dataset

Methods	CS	CV
Lie Group [24]	50.1	52.8
Part-aware LSTM [25]	62.9	70.3
ST-LSTM + Trust Gate [26]	69.2	77.7
Temporal Perceptive Network [9]	75.3	84.0
Context-aware attention LSTM [16]	76.1	84.0
Enhanced skeleton visualization [13]	76.0	82.6
Temporal CNNs[10]	74.3	83.1
Clips+CNN+Concatenation [9]	77.1	81.1
Clips+CNN+MTLN [9]	79.6	84.8
VA-LSTM [5]	79.4	<b>87.6</b>
SkeletonNet [8]	75.9	81.2
(WB + BP) + VGG	68.1	72.4
BP + F2C network	78.2	81.9
(WB + BP) w/o velocity + F2C network	76.6	81.7
<b>F2CSkeleton (Proposed)</b>	<b>79.6</b>	84.6

lengths. In our experiments, we use the two standard evaluation protocols proposed by the original study [24], namely, cross-subject (CS) and cross-view (CV).

**SBU Kinect Interaction Dataset** is another skeleton-based dataset collected using the Microsoft Kinect sensor. There are 282 skeleton sequences divided into 21 subsets, which are collected from eight different types of two-person interactions including approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. Each skeleton contains 15 joints. There are seven subjects who performed the actions in the same laboratory environment. We also augment data as in [8] before doing five-fold cross-validation. Each skeleton image is first resized to  $250 \times 250$  and then is randomly cropped into 20 sub-images with the size of  $224 \times 224$ . Eventually, we obtain a dataset of 11,280 samples.

**Implementation Details** The proposed model was implemented using Keras<sup>1</sup> with TensorFlow backend. For a fair comparison with the previous studies, transfer learning is applied in order to improve the classification performance. To be more specific, our proposed F2C network architecture is first trained with ImageNet with the input image dimension is set to  $224 \times 224$ . The pre-trained weights are then applied to all experiments.

Regarding input skeletons at each time step, we consider up to two distinct human subjects at once. This means that in case of two-person interactions, joint position features of the two subjects at a certain frame are read simultaneously and place side by side. In the case of single actions, we use zero matrices in the presentation of the second subject. Figure 4 shows some examples of skeleton images generated from NTU RGB+D dataset.

For NTU RGB+D dataset, we first remove 302 missing skeletons reported by [24]. 20% of training samples are used as a validation set. The first fully connected layer has 256 hidden units, while the output layer has the same size as the number of actions in the datasets. The network is trained using Adam for stochastic optimization [10]. The learning rate is set to 0.001 and exponentially decayed over 25 epochs. We use a batch size of 32. The same experimental settings are applied to all the experiments.

We set the number of temporal segments to seven, because it shows the best performance

<sup>1</sup><https://github.com/keras-team/keras>



Table 3: Classification Performance with Two-person Interactions, RGB+D Dataset, CV Protocol

Actions	SkeletonNet		F2CSkeleton	
	Prec.	Rec.	Prec.	Rec.
Punching/slapping	59.2	56.0	<b>80.6</b>	<b>82.2</b>
Kicking	46.8	64.9	<b>90.4</b>	<b>91.3</b>
Pushing	69.7	72.2	<b>88.0</b>	<b>86.1</b>
Pat on back	54.7	46.2	<b>82.8</b>	<b>80.7</b>
Point finger	42.8	72.8	<b>88.3</b>	<b>91.1</b>
Hugging	77.6	83.5	<b>92.9</b>	<b>83.8</b>
Giving something	72.5	72.5	<b>88.7</b>	<b>91.8</b>
Touch other’s pocket	66.9	50.6	<b>90.9</b>	<b>95.3</b>
Handshaking	83.1	82.6	<b>95.8</b>	<b>94.9</b>
Walking towards	66.2	82.3	<b>96.9</b>	<b>97.8</b>
Walking apart	61.8	<b>78.5</b>	<b>76.2</b>	77.7

\* Prec.: Precision      Rec.: Recall

Table 4: Classification Performance on SBU Dataset

Methods	Acc.
Deep LSTM+Co-occurrence [67]	90.4
ST-LSTM+Trust Gate [68]	93.3
SkeletonNet [8]	93.5
Clips+CNN+Concatenation [9]	92.9
Clips+CNN+MTLN [9]	93.6
Context-aware attention LSTM [66]	94.9
VA-LSTM [63]	97.2
<b>F2CSkeleton (Proposed)</b>	<b>99.1</b>

on NTU RGB+D dataset. Considering body part features have different contributions to an action, we do not share weights between input slices during training. This might increase the number of parameters but gain better generalization ability of the network. Table 1 shows the detail of our network configuration.

## 4.2 Experimental Results

**NTU RGB+D Dataset** We compare the performance of our method with the previous studies in Table 2. The classified accuracy is chosen as the evaluation metric.

*(WB + BP) + VGG* In this experiment, we use VGG16 pre-trained on ImageNet dataset instead of our F2C network. This experiment examines the significance of the proposed F2C network for high-level feature learning against the conventional deep CNN models.

*BP + F2C network* In this experiment, we only adopt the skeleton images generated by BP features to feed into the proposed F2C network architecture. This aims to justify the contribution of WB features going through our F2C network.

*(WB + BP) w/o velocity + F2C network* In this experiment, only joint position features are put into the proposed F2C network architecture for the purpose of examining the importance of incorporating velocity feature to the final classification performance.

*WB + BP + F2C network (F2CSkeleton)* This is our proposed method.

As shown in Table 2, our proposed method outperforms results reported by [7, 8, 14, 16, 18, 24, 27] with the same testing condition. In particular, we gain over 3.0% improvement from our baseline [8] on both CS and CV testing protocols. Similarly, our method is around 2.5% better than the method with feature concatenation [9]. However, [9] with Multi-Task Learning Network (MTLN) obtained a slightly better performance than our method with the CV protocol. The learning paradigm MTLN works as a hierarchical method to effectively learn the intrinsic correlations between multiple related tasks [55], thus, outperforms a mere concatenation. We believe our method also can benefit from MTLN. We will include this as a part of our future work to improve our network. It is also to note that while our method with the CS protocol outperforms [63], they achieved a better performance of 3% when coming to the CV protocol handled by a view adaption scheme in a multiple-view environment.

Table 2 also shows that our F2C network performs significantly better than VGG16. In particular, our F2C network improves the accuracy from 68.1% to 79.6% with the CS protocol and from 72.4% to 84.6% with the CV protocol. The incorporation of velocity

improves the performance about 3.0 points in both testing protocols. Besides, the use of WB and BP features in combination improves the accuracies from 78.2% to 79.6% and 81.9% to 84.6% with the CS and CV protocol, respectively.

Our method outperforms SkeletonNet on all the two-person interactions. Table 3 shows our classification performance with the CV protocol. Two-person interactions usually require the movement of the whole body. Top layers of our tailored network architecture can learn the whole body motion better than the naive CNN models originally designed for detecting generic objects in a still image.

On the other hand, it appears that our method performs poorly on two classes, namely “brushing teeth” (58.3%) and “brushing hair” (47.6%). Confusion matrix reveals that “brushing teeth” is often misclassified as either “cheer up” and “hand waving”, while the “brushing hair” is misclassified as “hand waving”. This may be because the “head joint”, which is selected as the reference joint for the torso, is not stationary enough compared to the other reference joints in these action types.

**SBU Kinect Interaction Dataset** Table 4 shows the comparisons of our proposed method with the previous studies on SBU dataset. As can be seen, our proposed method achieved the best performance on this dataset over all the other previous methods. In particular, our method gains more than 5.0 points improvement compared to the two state-of-the-art CNN-based methods [8, 9], about 4.0 points better than [16], and approximately 2.0 points better than [33]. These results again confirm that our method has superior performance on two-person interaction actions.

## 5 Conclusion

This paper addresses two problems of the previous studies: the loss of temporal information in a skeleton sequence when modeling using CNNs and the need for a network model specific to a human skeleton sequence. We first propose to segment a skeleton sequence to retrieve the dependencies between temporal segments in an action. We also propose an F2C CNN architecture for exploiting the spatio-temporal feature of skeleton data. As a result, our method with only three network blocks shows the superior generalization ability across very deep CNN models. We achieve a performance of 79.6% and 84.6% of accuracies on the large skeleton dataset, NTU RGB+D, with cross-object and cross-view protocol, respectively, which reaches the state-of-the-art.

In the future, as has been noted, we will adopt the notion of multi-task learning. In addition, since we do not share weights between input slices during training, our network has more trainable parameters compared to general CNN models with the same input size and the number of filters. We believe our method will work better if we reduce the number of feature maps in convolutional layers. The current skeleton data is very challenging due to noisy joints. For example, by manually checking skeleton data from the first data collection setup of NTU RGB+D, we find that there were about 8.8% of noisy detections. Because our method did not apply any algorithms to remove these noises from the input, it is promising to take this into consideration for better performance.

## Acknowledgments

This work was supported by JSPS KAKENHI 15K12061 and by JST CREST Grant Number JPMJCR1687, Japan.

## References

- [1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference (BMVC)*, 2014.
- [2] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.
- [3] Dileep George, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan, Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 358(6368):eaag2612, 2017.
- [4] Fei Han, Brian Reily, William Hoff, and Hao Zhang. Space-time representation of people based on 3d skeletal data: A review. *Proc. of Computer Vision and Image Understanding*, 158:85–105, 2017.
- [5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [6] Hsieh Hou and H Andrews. Cubic splines for image interpolation and digital filtering. *IEEE Transactions on acoustics, speech, and signal processing*, 26(6):508–517, 1978.
- [7] Yueyu Hu, Chunhui Liu, Yanghao Li, Sijie Song, and Jiaying Liu. Temporal perceptive network for skeleton-based action recognition. In *Proc. of British Machine Vision Conference (BMVC)*, pages 1–2, 2017.
- [8] Qihong Ke, Senjian An, Mohammed Bennamoun, Ferdous Sohel, and Farid Bous-said. Skeletonnet: Mining deep part features for 3-d action recognition. *IEEE signal processing letters*, 24(6):731–735, 2017.
- [9] Qihong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4570–4579. IEEE, 2017.
- [10] Tommi Kerola, Nakamasa Inoue, and Koichi Shinoda. Graph regularized implicit pose for 3d human action recognition. In *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.
- [11] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1623–1631. IEEE, 2017.
- [12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Proc. of International Conference on Learning Representations (ICLR)*, 2015.
- [13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

- [14] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *Proc. of European Conference on Computer Vision*, pages 816–833. Springer, 2016.
- [15] Jun Liu, Amir Shahroudy, Dong Xu, Alex Kot Chichung, and Gang Wang. Skeleton-based action recognition using spatio-temporal lstm network with trust gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [16] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2018.
- [17] Mengyuan Liu, Chen Chen, and Hong Liu. 3d action recognition using data visualization and convolutional neural networks. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 925–930. IEEE, 2017.
- [18] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.
- [19] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *Proc. of the 32nd International Conference on Machine Learning, ICML*, pages 97–105, 2015.
- [20] Jiajia Luo, Wei Wang, and Hairong Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *Proc. of Computer vision (ICCV)*, pages 1809–1816. IEEE, 2013.
- [21] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
- [22] Liliana Lo Presti and Marco La Cascia. 3d skeleton-based human action classification: A survey. *Proc. of Pattern Recognition*, 53:130–147, 2016.
- [23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- [24] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Proc. of CoRR*, 2014.
- [26] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, volume 1, page 7, 2017.

- [27] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proc. of Computer Vision and Pattern Recognition*, pages 588–595, 2014.
- [28] Raimar Wagner, Markus Thom, Roland Schweiger, Gunther Palm, and Albrecht Rothmel. Learning convolutional neural networks from few samples. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–7. IEEE, 2013.
- [29] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1297. IEEE, 2012.
- [30] Xiaodong Yang and YingLi Tian. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1):2–11, 2014.
- [31] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 28–35. IEEE, 2012.
- [32] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proc. of the IEEE International Conference on Computer Vision*, pages 2752–2759, 2013.
- [33] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. *International Conference on Computer Vision*, pages 2136–2145, 2017.
- [34] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 148–157. IEEE, 2017.
- [35] Yu Zhang and Dit-Yan Yeung. A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):12, 2014.
- [36] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, Xiaohui Xie, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Proc. of Association for the Advancement of Artificial Intelligence (AAAI)*, volume 2, page 8, 2016.