論文 / 著書情報
Article / Book Information

論題	
Title	Skeleton-based Human Action Recognition with Fine-to-Coarse Convolutional Neural Network
著者	LE THAO MINH, 井上 中順, 篠田 浩一
Authors	Thao Minh Le, Nakamasa Inoue, Koichi Shinoda
出典	信学技報, vol. 118, no. 362, pp. 61-64
Citation	IEICE Technical Report, vol. 118, no. 362, pp. 61-64
発行日 / Pub. date	2018, 12
URL	http://search.ieice.org/
権利情報	本著作物の著作権は電子情報通信学会に帰属します。
Copyright	(c) 2018 Institute of Electronics, Information and Communication Engineers

社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

Skeleton-based Human Action Recognition with Fine-to-Coarse Convolutional Neural Network

Thao MINH LE†, Nakamasa INOUE†, and Koichi SHINODA†

† Tokyo Institute of Technology, Japan

Abstract This work introduces a new framework for skeleton-based human action recognition. Existing approaches using Convolutional Neural Network (CNN) often suffer from the insufficiency problem of training data. In this study, we utilize a fine-to-coarse (F2C) CNN architecture that is come up based on the special structure of human skeletal data. We evaluate our proposed method on two skeletal datasets publicly available, namely 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.

Key words Action Recognition, Human Skeleton, Fine-to-Coarse CNN

1. Introduction

Human action recognition studies utilizing 3D skeleton data have drawn a great deal of attention [1] due to its applications in a number of areas including security surveillance systems, human-computer-interaction-based games, and the healthcare industry.

Earlier methods of 3D human action recognition utilized hand-crafted features for representing the intra-frame relationships through the skeleton sequences [2]. Some studies are built upon 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 [3], [4]. Recent studies have shown the superiority of Convolutional Neural Networks (CNNs) over RNN with LSTM for this task [5], [6]. These CNN-based methods are, 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 [7], we believe 3D human action recognition systems can also benefit from a specific network structure for this application domain. In particular, we introduce a fine-to-coase (F2C) CNN architecture that utilizes 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. We are unaware of any attempt to use F2C network for 3D human action recognition.

2. Fine-to-Coarse CNN for Skeleton-based Human Action Recognition

Figure 1 shows an overview of our framework. It consists of two phases: feature representation and high-level feature learning with a F2C network architecture.

2. 1 Feature Representation

We encode the geometry of human body originally given in an image space into local coordinate systems. Motivated by Ke at al. [5], six joints located in different body parts are selected as reference joints in order to generate whole-body-based (WB) features and body-part-based (BP) features. In other words, 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 used 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.

Different from the other studies using BP features [3], [5], [8], 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 handcrafted-feature-based approaches [9]. 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:

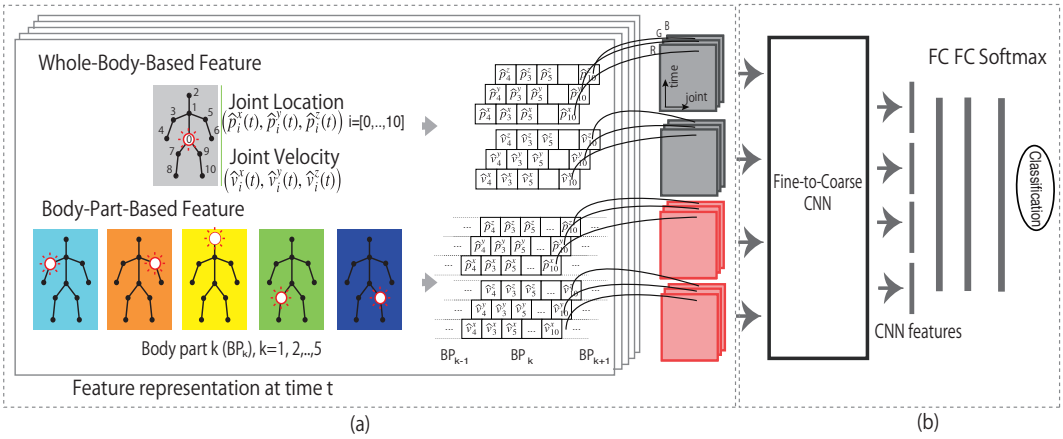

Figure 1 System Overview.

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

The relative inter-joint positions are highly discriminative for human actions. 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. [9] 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:

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

Considering each component of WB and BP features are 2D array features, we finally 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. We call these two RGB images as skeleton images.

2.2 Fine-to-Coarse Network Architecture

Our F2C network, as illustrated in Figure 2, takes three color channels of skeleton images as inputs. Accordingly, the input of our F2C network consists of two dimensions: the spatial dimension describing 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. Assume T_{seg} (T=m$\times T_{\text{seg}}$) is the number of frames in one temporal segment, l_{bp} is the dimension of one body part along the

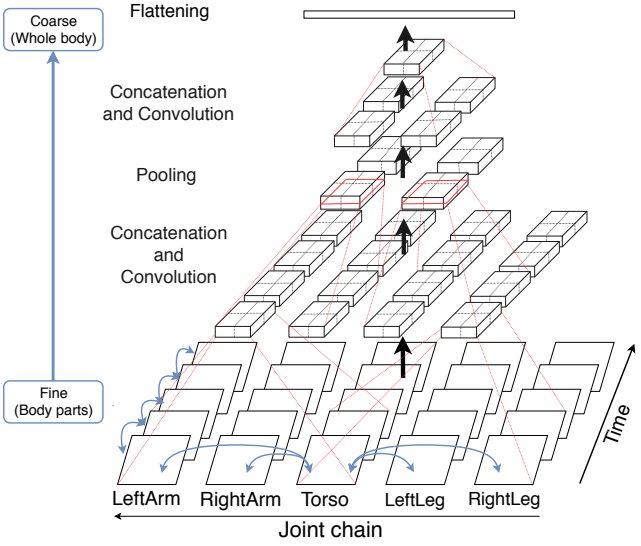

Figure 2 Fine-to-Coarse Network Architecture. Blue arrows show pair slices which are concatenated along each dimension before passing to a convolutional block.

spatial dimension, each input slide has size of $l_{\text{bp}} \times T_{\text{seg}}$. We then simultaneously concatenate the slices over both spatio-temporal axes. In other words, we first concatenate each body part which belongs to human limbs with the torso, while concatenating two consecutive temporal segments together. Each concatenated 2D array feature is further passed through a convolution layer followed by a max pooling layer. The same procedure is applied in the next step. In short, our F2C network composes of three layer-concatenation steps, and three convolution blocks accordingly.

Both WB-based and BP-based skeleton images are gone through 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 through the network since the spatial dimension of WB features, which are formed by a pre-defined joint chain, contains the intrinsic relationships between body

Table 1 Classification Performance on NTU RGB+D Dataset

Methods	CS	CV
Enhanced skeleton visualization [10]	76.0	82.6
Temporal CNNs [11]	74.3	83.1
Clips+CNN+Concatenation [6]	77.1	81.1
Clips+CNN+MTLN [6]	**79.6**	84.8
VA-LSTM [4]	79.4	**87.6**
SkeletonNet [5]	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
F2CSkeleton (Proposed)	**79.6**	84.6

parts.

Our network can be viewed as a procedure to eliminate unwanted connections between layers from the CNN. We believe conventional 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, pick up something). For this reason, the bottom layers in our method can discriminate "fine" actions which require the interactions of some certain body parts, while the top layers are discriminative for "coarse" actions using the movements of the whole body.

3. Experimental Results

3.1 Datasets and Experimental Conditions

We conduct experiments on two skeleton benchmark datasets publicly available: NTU RGB+D [3] and SBU Kinect Interaction Dataset [12].

NTU RGB+D Dataset: contains 56,880 skeleton sequences. Each skeleton contains 25 human joints. This dataset is collected by 40 human subjects performing 60 distinct action classes of three human-action groups: daily actions, health-related actions, and two-person interactive actions. It is challenging due to the large variations of viewpoints and sequence lengths. We use the two standard evaluation protocols proposed by the original study [3], namely, cross-subject (CS) and cross-view (CV).

SBU Kinect Interaction Dataset: 282 skeleton sequences divided into 21 subsets 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 [5] before doing five-fold cross-validation. Eventually, we obtain a dataset of 11,280 samples.

Table 2 Two-person-Interactions, RGB+D dataset (CV)

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

* Prec.: Precision Rec.: Recall

Implementation Details 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 of 224×224. 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.

For NTU RGB+D dataset, 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 [13]. 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.

3.2 Experimental Results

NTU RGB+D Dataset We compare the performance of our method with the previous studies in Table 1. The classified accuracy is chosen as the evaluation metric. *(WB + BP) + VGG* uses VGG16 pre-trained on ImageNet dataset instead of our F2C network. This examines the significance of the proposed F2C network for high-level feature learning against the conventional deep CNN models. *BP + F2C network* only adopts 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* only uses joint position features which are put into the proposed F2C network to examine the importance of incorporating velocity feature to the final classification performance. Finally, *WB + BP + F2C network (F2CSkeleton)* is our proposed method.

As shown in Table 1, our proposed method outperforms results reported by [5], [6], [10], [11] with the same testing con-

Table 3 Classification Performance on SBU Dataset

Methods	Acc.
SkeletonNet [5]	93.5
Clips+CNN+Concatenation [6]	92.9
Clips+CNN+MTLN [6]	93.6
Context-aware attention LSTM [15]	94.9
VA-LSTM [4]	97.2
F2CSkeleton (Proposed)	**99.1**

dition. In particular, we gain over 3.0% improvement from our baseline [5] on both two testing protocols, and around 2.5 points better than Ke et al. [6] with feature concatenation. However, Ke et al. [6] with Multi-Task Learning Network (MTLN) obtained a slightly better performance than ours with the CV protocol. The learning paradigm MTLN works as a hierarchical method to effectively learn the intrinsic correlations between multiple related tasks [14], thus, outperforms a mere concatenation.

Table 1 also shows that our F2C network performs significantly better than VGG16 by approximately 12 points, while the incorporation of velocity improves the performance about 3.0 points with 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 CS and CV protocol, respectively.

Our method outperforms SkeletonNet on all the two-person interactions (Table 2). Two-person interactions usually require the movement of the whole body. We argue that 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.

SBU Kinect Interaction Dataset Table 2 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 [5], [6], about 4.0 points better than [15], and approximately 2.0 points better than [4]. These results again confirm that our method has superior performance on two-person interaction actions.

4. Conclusion

This work addresses two problems of the previous studies: the loss of temporal information in a skeleton sequence when modeling with 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 is competitive with the state-of-the-art. In the future, as has been noted, we will adopt the notion of multi-task learning for better performance.

Acknowledgement

This work was supported by JSPS KAKENHI 15K12061 and by JST CREST Grant NumberJPMJCR1687, Japan.

References

[1] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: A review," *Elsevier CVIU*, vol. 158, pp. 85–105, 2017.

[2] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints," *Elsevier VCIR*, vol. 25, no. 1, pp. 2–11, 2014.

[3] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," *Proc. CVPR*, 2016.

[4] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," *Proc. ICCV*, pp. 2136–2145, 2017.

[5] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "Skeletonnet: Mining deep part features for 3-d action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 731–735, 2017.

[6] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," *Proc. CVPR*, pp. 4570–4579, 2017.

[7] D. George, W. Lehrach, K. Kansky, M. Lázaro-Gredilla, C. Laan, B. Marthi, X. Lou, Z. Meng, Y. Liu, H. Wang *et al.*, "A generative vision model that trains with high data efficiency and breaks text-based captchas," *Science*, vol. 358, no. 6368, p. 2612, 2017.

[8] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," *Proc. ECCV*, pp. 816–833, 2016.

[9] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," *Proc. ICCV*, pp. 2752–2759, 2013.

[10] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[11] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," *Proc. CVPR BNM Workshop*, pp. 1623–1631, 2017.

[12] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," *Proc. CVPR Workshops*, pp. 28–35, 2012.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015.

[14] Y. Zhang and D.-Y. Yeung, "A regularization approach to learning task relationships in multitask learning," *ACM Trans. on TKDD*, vol. 8, no. 3, p. 12, 2014.

[15] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Trans. on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2018.