

Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation

Chao Li, Qiaoyong Zhong, Di Xie, Shiliang Pu

Hikvision Research Institute

{lichao15, zhongqiaoyong, xiedi, pushiliang}@hikvision.com

Abstract

Skeleton-based human action recognition has recently drawn increasing attentions with the availability of large-scale skeleton datasets. The most crucial factors for this task lie in two aspects: the intra-frame representation for joint co-occurrences and the inter-frame representation for skeletons' temporal evolutions. In this paper we propose an end-to-end convolutional co-occurrence feature learning framework. The co-occurrence features are learned with a hierarchical methodology, in which different levels of contextual information are aggregated gradually. Firstly point-level information of each joint is encoded independently. Then they are assembled into semantic representation in both spatial and temporal domains. Specifically, we introduce a *global* spatial aggregation scheme, which is able to learn superior joint co-occurrence features over local aggregation. Besides, raw skeleton coordinates as well as their temporal difference are integrated with a two-stream paradigm. Experiments show that our approach consistently outperforms other state-of-the-arts on action recognition and detection benchmarks like NTU RGB+D, SBU Kinect Interaction and PKU-MMD.

1 Introduction

Analysis of human behavior such as action recognition and detection is one of the fundamental and challenging tasks in computer vision. It has a wide range of applications such as intelligent surveillance system, human-computer interaction, game-control and robotics. Articulated human pose, being also referred to as skeleton, provides a very good representation for describing human actions. On one hand, skeleton data are inherently robust against background noise and provide abstract information and high-level features of human action. On the other hand, compared with RGB data, skeleton data are extremely small in size, which makes it possible to design lightweight and hardware friendly models.

In this paper, we focus on the problem of skeleton-based human action recognition and detection (Figure 1). The interactions and combinations of skeleton joints play a key

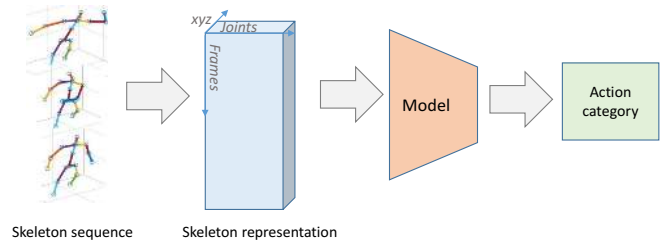


Figure 1: Workflow for skeleton-based human action recognition.

role to characterize an action. Many of the early works attempted to design and extract co-occurrence features from skeleton sequences, such as pairwise relative position of each joint [Wang *et al.*, 2014], spatial orientation of pairwise joints [Jin and Choi, 2012], and the statistics-based features like Cov3DJ [Hussein *et al.*, 2013] and HOJ3D [Xia *et al.*, 2012]. On the other hand, the Recurrent Neural Networks (RNNs) with Long-Short Term Memory (LSTM) neurons are used to model the time series of skeleton prevalently [Shahrudy *et al.*, 2016; Song *et al.*, 2017; Liu *et al.*, 2016]. Although the LSTM networks were designed to model long-term temporal dependency, it is difficult for them to learn high-level features from skeletons directly since the temporal modeling is done on the raw input space [Sainath *et al.*, 2015]. While fully connected layers are able to learn co-occurrence features for their ability of aggregating global information from all input neurons. In [Zhu *et al.*, 2016], an end-to-end fully connected deep LSTM network was proposed to learn co-occurrence features from skeleton data.

CNN models are equipped with excellent ability to extract high-level information, and they have been used to learn spatial-temporal features from skeletons [Du *et al.*, 2016; Ke *et al.*, 2017]. These CNN-based methods represent a skeleton sequence as an image by encoding the temporal dynamics and the skeleton joints as rows and columns respectively, and then feed it into a CNN to recognize the underlying action just like image classification. However, in that case, only the neighboring joints within the convolutional kernel are considered to learn co-occurrence features. Although the receptive field covers all joints of a skeleton in later convolution layers, it is difficult to mine co-occurrences from all joints efficiently. Because of the weight sharing mechanism in spatial dimensions, CNN models can not learn free pa-

rameters for each joint. This motivates us to design a model which is able to get global response from all joints to exploit the correlations between different joints.

We propose an end-to-end co-occurrence feature learning framework, which uses CNN to learn hierarchical co-occurrence features from skeleton sequences automatically. We find the output of a convolution layer is global response from all input channels. If each joint of a skeleton is treated as a channel, then the convolution layer can learn the co-occurrences from all joints easily. More specifically, we represent a skeleton sequence as a tensor of shape $frames \times joints \times 3$ (the last dimension as channel). We first learn point-level features for each joint independently using convolution layers with kernel size $n \times 1$. And then we transpose the output of the convolution layers to make the dimension of $joints$ as channel. After the transpose operation, the subsequent layers aggregate global features from all joints hierarchically. Furthermore, the two-stream framework [Simonyan and Zisserman, 2014] is introduced to fuse the skeleton motion feature explicitly.

The main contributions of this work are summarized as follows:

- We propose to employ the CNN model for learning *global* co-occurrences from skeleton data, which is shown superior over local co-occurrences.
- We design a novel end-to-end hierarchical feature learning network, where features are aggregated gradually from point-level features to global co-occurrence features.
- We comprehensively exploit multi-person feature fusion strategies, which makes our network scale well to variable number of persons.
- The proposed framework outperforms all existing state-of-the-art methods on benchmarks for both action recognition and detection tasks.

2 Related Work

The LSTM networks were designed to model long-term temporal dependency problems. Thus they are a natural choice and have been well exploited for feature learning from skeleton sequences. However, more and more literatures adopted CNN to learn skeleton features and achieved impressive performance in recent years. For example, [Du *et al.*, 2016] proposed to cast the frame, joint and coordinate dimensions of a skeleton sequence into width, height and channel of an image respectively. Then they applied CNN for skeleton-based action recognition in the same way as CNN-based image classification. [Ke *et al.*, 2017] proposed an improved representation of skeleton sequences where the 3D coordinates are separated into three gray-scale images. In [Li *et al.*, 2017b], a skeleton transformer module was introduced to learn a new representation of skeleton joints. And a two-stream convolutional network was proposed to incorporate skeleton motion features.

Similar to the above works, we also adopt CNN to learn features from skeleton data. However, we attempt to model global co-occurrence patterns with CNN. And we explicitly

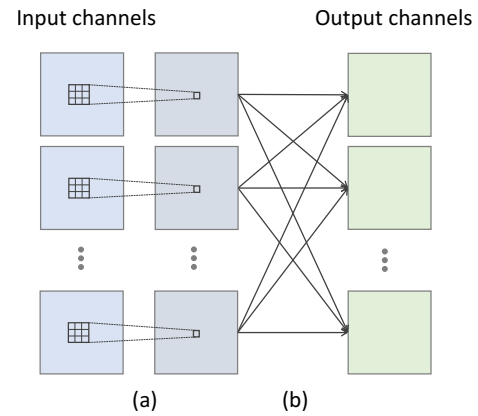


Figure 2: Decomposition of a 3×3 convolution into two steps. (a) Independent 2D convolution in the spatial domain for each input channel, where features are aggregated locally from 3×3 neighborhoods. (b) Element-wise summation across channels, where features are aggregated globally from all input channels.

formulate the problem into two levels of feature learning sub-problems, i.e. independent point-level feature learning and cross-joint co-occurrence feature learning.

3 Methods

3.1 Co-occurrence Feature Learning with CNN

CNN is one of the most powerful and successful neural network models, which has been widely applied in image classification, object detection, video classification etc. Compared with sequential structures such as RNNs, it is capable of encoding spatial and temporal contextual information simultaneously. By investigating the convolution operation, we may decompose it into two steps (Figure 2), i.e. local feature aggregation across the spatial domain (width and height) and global feature aggregation across channels. Then one can conclude a simple yet very practical way to regulate the aggregation extent on demand. Denote T as a $d_1 \times d_2 \times d_3$ 3D tensor. We can assign different context by reorganizing (transposing) the tensor. Any information from dimension d_i can be aggregated *globally* if it is specified as channels while the other two encode *local* context.

In all previous CNN-based methods [Du *et al.*, 2016; Ke *et al.*, 2017; Li *et al.*, 2017b], joint coordinates are specified as channels. It causes a problem: the co-occurrence features are aggregated locally, which may not be able to capture long-range joint interactions involved in actions like wearing a shoe. To this end, we argue that aggregating co-occurrence features *globally* is of great importance and leads to better action recognition performance. It can be easily implemented by putting the joint dimension into channels of CNN’s input.

3.2 Explicit Skeleton Motion

Besides joint co-occurrences, temporal movements of joints are crucial cues to recognize the underlying action. Although the temporal evolution pattern can be learned implicitly with CNN, we argue that an explicit modeling is preferable. Thus we introduce a representation of skeleton motion and feed it explicitly into the network.

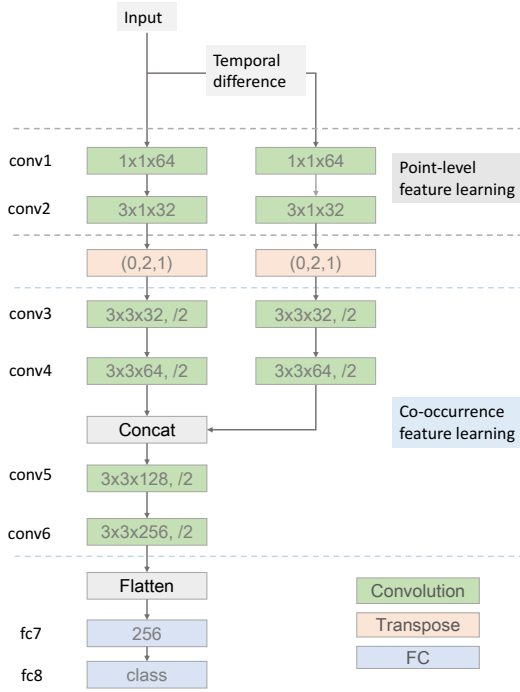


Figure 3: Overview of the proposed hierarchical co-occurrence network. Green blocks are convolution layers, where the last dimension denotes the number of output channels. A trailing “/2” means an appended MaxPooling layer with stride 2 after convolution. A Transpose layer permutes the dimensions of the input tensor according to the order parameter. ReLU activation function is appended after *conv1*, *conv5*, *conv6* and *fc7* to introduce non-linearity.

For the skeleton of a person in frame t , we formulate it as $\mathbf{S}^t = \{\mathbf{J}_1^t, \mathbf{J}_2^t, \dots, \mathbf{J}_N^t\}$ where N is the number of joint and $\mathbf{J} = (x, y, z)$ is a 3D joint coordinate. The skeleton motion is defined as the temporal difference of each joint between two consecutive frames:

$$\begin{aligned} \mathbf{M}^t &= \mathbf{S}^{t+1} - \mathbf{S}^t \\ &= \{\mathbf{J}_1^{t+1} - \mathbf{J}_1^t, \mathbf{J}_2^{t+1} - \mathbf{J}_2^t, \dots, \mathbf{J}_N^{t+1} - \mathbf{J}_N^t\}. \end{aligned}$$

Raw skeleton coordinates \mathbf{S} and the skeleton motion \mathbf{M} are fed into the network independently with a two-stream paradigm. To fuse information from the two sources, we concatenate their feature maps across channels in subsequent layers of the network (see Figure 3).

3.3 Hierarchical Co-occurrence Network

In this section, we will elaborately describe the proposed Hierarchical Co-occurrence Network (HCN) framework, which is designed to learn the joint co-occurrences and the temporal evolutions jointly in an end-to-end manner.

Figure 3 shows the network architecture of the proposed framework. A skeleton sequence \mathbf{X} can be represented with a $T \times N \times D$ tensor, where T is the number of frames in the sequence, N is the number of joints in the skeleton and D is the coordinate dimension (e.g. 3 for 3D skeleton). The skeleton motion described above is of the same shape as \mathbf{X} . They

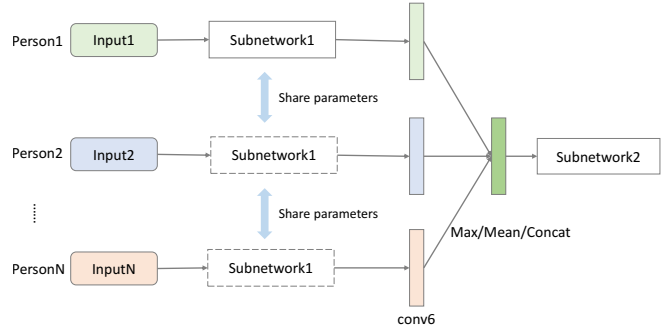


Figure 4: Late fusion diagram for multi-person feature fusion. Maximum, mean and concatenation operations are evaluated in terms of performance and generalization.

are fed into the network directly as two streams of inputs. The two network branches share the same architecture. However, their parameters are not shared and learned separately. Their feature maps are fused by concatenation along channels after *conv4*.

Given skeleton sequence and motion inputs, the features are learned hierarchically. In stage 1, point-level features are encoded with a 1×1 (*conv1*) and $n \times 1$ (*conv2*) convolution layers. Since the kernel sizes along the joint dimension are kept 1, they are forced to learn point-level representation from 3D coordinates for each joint independently. After that, we transpose the feature maps with parameter (0, 2, 1) so that the joint dimension is moved to channels of the tensor. Then in stage 2, all subsequent convolution layers extract global co-occurrence features from all joints of a person as described in Section 3.1. Afterwards, the feature maps are flattened into a vector and go through two fully connected layers for final classification. Note that the skeleton’s temporal evolutions are encoded all through the network with convolutions of kernel size equal to 3 along the frame dimension.

Our network contains about 0.8 million parameters, just two-thirds of the model in [Li *et al.*, 2017b]. While [Ke *et al.*, 2017] used an ImageNet-pretrained VGG19 model. The extremely small model size allows us to easily train the network from scratch without the needs of pretraining.

3.4 Scalability to Multiple Persons

In activities like hugging and shaking hands, multiple persons are involved. To make our framework scalable to multi-person scenarios, we perform a comprehensive evaluation on different feature fusion strategies.

Early fusion. All joints from multiple persons are stacked as input of the network. For variable number of persons, zero padding is applied if the number of persons is less than the pre-defined maximal number.

Late fusion. As illustrated in Figure 4, inputs of multiple persons go through the same subnetwork and their *conv6* feature maps are merged with either concatenation along channels or element-wise maximum / mean operation.

Note that element-wise late fusion can generalize well to variable number of persons, while the others require a pre-defined maximal number. Besides, compared with single person, no extra parameter is introduced.

3.5 Action Recognition and Detection

For the recognition task, a softmax function is used to normalize the output of the network. The probability that a sequence \mathbf{X} belongs to the i^{th} class is

$$P(C_i|\mathbf{X}) = \frac{e^{o_i}}{\sum_{j=1}^C e^{o_j}}, \quad i = 1, 2, \dots, C, \quad (1)$$

where $\mathbf{o} = (o_1, o_2, \dots, o_C)^T$ is the output of the network, C is the number of classes.

We also extend the proposed network for temporal action detection. Previously the Faster R-CNN framework [Ren *et al.*, 2015] has been adapted for the task of temporal action detection [Xu *et al.*, 2017; Li *et al.*, 2017b]. Following these works, we briefly introduce our implementation and later show that with the proposed hierarchical co-occurrence features, detection performance also gets significantly improved. The detection framework is shown in Figure 5. Specifically, based on the backbone feature learning network in Figure 3, two subnetworks are appended after *conv5*, i.e. the temporal proposal subnetwork and the action classification subnetwork. The temporal proposal subnetwork predicts variable-length temporal segments that potentially contain an action. The corresponding feature maps of each proposal are extracted using a crop-and-resize operation. After that, the classification subnetwork predicts their action categories. In both subnetworks, window regression is performed to obtain more accurate localization. We use softmax loss for classification and smooth L_1 loss for window regression. The two tasks are optimized jointly. The objective function is given by:

$$L = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i L_{reg}(t_i, t_i^*). \quad (2)$$

N_{cls} and N_{reg} are the number of samples. λ is the weight applied on the regression loss (empirically set to 1 in our experiments). i is the index of an anchor or proposal in a batch. p is the predicted probability and p^* is the groundtruth label. $t = \{t_x, t_w\}$ is the predicted regression target and $t^* = \{t_x^*, t_w^*\}$ is the groundtruth target. Window regression is essentially an adaptation of bounding box regression [Girshick *et al.*, 2014], where targets along one dimension rather than two are predicted. The target transformations are computed as follows:

$$t_x = (x - x_a)/w_a, \quad t_w = \log(w/w_a), \quad (3)$$

$$t_x^* = (x^* - x_a)/w_a, \quad t_w^* = \log(w^*/w_a), \quad (4)$$

where x and w denote the center and length of the temporal window. And x , x_a and x^* represent the predicted window, anchor window and groundtruth window respectively. The same rule applies for w .

4 Experiments

We evaluate the proposed method on three common benchmark datasets, i.e. the NTU RGB+D [Shahroudy *et al.*, 2016] and SBU Kinect Interaction [Yun *et al.*, 2012] datasets for action recognition, and the PKU-MMD [Liu *et al.*, 2017] dataset for temporal action detection. Besides, an ablation study is performed to show the importance of global co-occurrence feature aggregation.

4.1 Datasets and Implementation Details

NTU RGB+D

The NTU RGB+D dataset is so far the largest skeleton-based human action recognition dataset. It contains 56880 skeleton sequences, which are annotated as one of 60 action classes. There are two recommended evaluation protocols, i.e. Cross-Subject (CS) and Cross-View (CV). In the cross-subject setting, sequences of 20 subjects are used for training and sequences of the rest 20 subjects are used for validation. In the cross-view setting, samples are split by camera views. Samples from two camera views are used for training and the rest are used for testing.

During training, we randomly crop a sub-sequence from the entire sequence. The cropping ratio is drawn from uniform distribution between [0.5, 1]. During inference, we center crop a sub-sequence with a ratio of 0.9. Since different actions last for various durations, the input sequences are normalized to a fixed length (32 in our experiments) with bilinear interpolation along the frame dimension. To alleviate the problem of overfitting, we append dropout after *conv4*, *conv5*, *conv6* and *fc7* with a dropout ratio of 0.5. A weight decay of 0.001 is applied on the weights of *fc7*. We train the model for 300k iterations in total, with a mini-batch size of 64. The Adam [Kingma and Ba, 2015] optimizer is utilized. The learning rate is initialized to 0.001 and exponentially decayed every 1k steps with a rate of 0.99.

SBU Kinect Interaction

The SBU Kinect Interaction dataset [Yun *et al.*, 2012] is a Kinect captured human activity recognition dataset depicting two person interaction. It contains 282 skeleton sequences and 6822 frames of 8 classes. There are 15 joints for each skeleton. For evaluation we perform subject-independent 5-fold cross validation as suggested in [Yun *et al.*, 2012].

Considering the small size of the dataset, we simplify the network architecture in Figure 3 accordingly. Specifically, the output channels of *conv1*, *conv2*, *conv3*, *conv5*, *conv6* and *fc7* are reduced to 32, 16, 16, 32, 64 and 64 respectively. And the *conv4* layer is removed. Besides, all the input sequences are normalized to a length of 16 frames rather than 32.

PKU-MMD

The PKU-MMD dataset [Liu *et al.*, 2017] is currently the largest skeleton-based action detection dataset. It contains 1076 long untrimmed video sequences performed by 66 subjects in three camera views. 51 action categories are annotated, resulting almost 20,000 action instances and 5.4 million frames in total. Similar to NTU RGB+D, there are also two recommended evaluate protocols, i.e. cross-subject and cross-view. For detection-specific hyper-parameters, we basically follow the settings in [Ren *et al.*, 2015]. In particular, we use anchor scales of {50, 100, 200, 400} in the temporal proposal network.

4.2 Multi-person Feature Fusion

To evaluate different ways of multi-person feature fusion described in Section 3.4, we perform an ablation study on the NTU RGB+D dataset. As shown in Table 1, all late fusion methods outperform the early fusion method. The reason

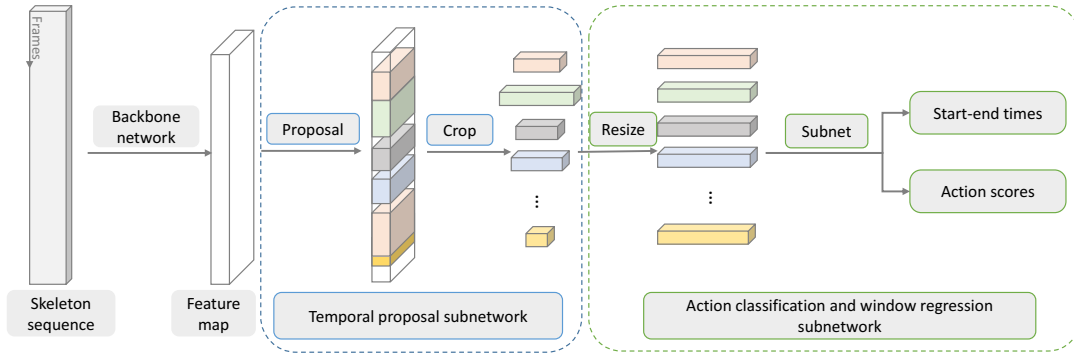


Figure 5: The temporal action detection framework. The backbone network is described in Figure 3. Two subnetworks are designed for temporal proposal segmentation and action classification respectively.

| Method | | Accuracy (%) |
|--------------|------------|--------------|
| Early fusion | | 85.2 |
| Late fusion | Mean | 85.8 |
| | Concat | 85.9 |
| | Max | 86.5 |

Table 1: Performance of different fusion methods for multi-person feature on the NTU RGB+D dataset in the cross-subject setting.

might be that features of different persons are better aligned and thus more compatible in high-level semantic space than in raw input space. Among the three late fusion implementations, the element-wise maximum operation achieves the best accuracy. This is due to the side effect of zero padding for single-person actions. That is, compared with multi-person samples, features of single-person samples get weakened with the padded zeros in both cases of concatenation and element-wise mean. While element-wise maximum does not suffer from this issue. In the following experiments the late fusion with element-wise maximum strategy is adopted.

4.3 Comparison to Other State-of-the-arts

A systematic evaluation of the proposed HCN framework is performed on the three datasets mentioned above. As shown in Table 2, Table 3 and Table 4, our approach consistently outperforms the current state-of-the-arts in terms of both action recognition accuracy and action detection mAP.

On the large-scale NTU RGB+D dataset (Table 2), our approach achieves the best action recognition accuracy. Compared with the state-of-the-art LSTM-based method [Zhang *et al.*, 2017], the accuracy is improved by 7.3% in the cross-subject setting and 3.4% in the cross-view setting. Compared with the most recent two-stream CNN method [Li *et al.*, 2017b], the accuracy is improved by 3.3% in the cross-subject setting and 1.8% in the cross-view setting.

On the small SBU Kinect Interaction dataset (Table 3), the proposed method also outperforms other methods by a large margin. Compared with the LSTM-based co-occurrence learning baseline [Zhu *et al.*, 2016], the accuracy is improved by 8.2%, which proves the superiority of our CNN-based global co-occurrence feature learning framework. Since the recognition task on SBU is much easier than

| Methods | Accuracy (%) | |
|--|--------------|-------------|
| | CS | CV |
| Deep LSTM [Shahroudy <i>et al.</i> , 2016] | 60.7 | 67.3 |
| Part-aware LSTM [Shahroudy <i>et al.</i> , 2016] | 62.9 | 70.3 |
| ST-LSTM+Trust Gate [Liu <i>et al.</i> , 2016] | 69.2 | 77.7 |
| STA-LSTM [Song <i>et al.</i> , 2017] | 73.4 | 81.2 |
| Clips + CNN + MTLN [Ke <i>et al.</i> , 2017] | 79.6 | 84.8 |
| VA-LSTM [Zhang <i>et al.</i> , 2017] | 79.2 | 87.7 |
| Two-stream CNN [Li <i>et al.</i> , 2017b] | 83.2 | 89.3 |
| Proposed HCN | 86.5 | 91.1 |

Table 2: Action classification performance on the NTU RGB+D dataset. CS and CV mean the cross-subject and cross-view settings respectively.

| Methods | Accuracy (%) |
|---|--------------|
| Raw skeleton [Ji <i>et al.</i> , 2014] | 79.4 |
| Joint feature [Ji <i>et al.</i> , 2014] | 86.9 |
| ST-LSTM [Liu <i>et al.</i> , 2016] | 88.6 |
| Co-occurrence RNN [Zhu <i>et al.</i> , 2016] | 90.4 |
| STA-LSTM [Song <i>et al.</i> , 2017] | 91.5 |
| ST-LSTM+Trust Gate [Liu <i>et al.</i> , 2016] | 93.3 |
| VA-LSTM [Zhang <i>et al.</i> , 2017] | 97.6 |
| Proposed HCN | 98.6 |

Table 3: Action classification performance on the SBU dataset.

on NTU RGB+D, the best accuracy reported has achieved 97.6% [Zhang *et al.*, 2017]. Yet we push the frontier further and achieve an accuracy of 98.6%.

On the PKU-MMD action detection dataset (Table 4), we also achieve state-of-the-art performance. Compared with the Skeleton boxes method [Li *et al.*, 2017a], in which an adapted VGGNet is employed for temporal action detection, our method improve the mAP by 38% in the cross-subject setting. Compared with the recent work in [Li *et al.*, 2017b], where a similar detection framework is utilized, our method improves the mAP by 2.2% and 0.5% in the cross-subject and cross-view settings respectively. Note that our improvement over it is purely owing to better features learned with the proposed framework.

From the results, we can conclude that our proposed hier-

| Methods | mAP (%) | |
|---|-------------|-------------|
| | CS | CV |
| STA-LSTM [Song <i>et al.</i> , 2017] | 44.4 | 13.1 |
| JCRRNN [Li <i>et al.</i> , 2016] | 32.5 | 53.3 |
| Skeleton boxes [Li <i>et al.</i> , 2017a] | 54.8 | 94.2 |
| Li <i>et al.</i> [Li <i>et al.</i> , 2017b] | 90.4 | 93.7 |
| Proposed HCN | 92.6 | 94.2 |

Table 4: Action detection performance on the PKU-MMD dataset. mAP is measured at an IoU threshold of 0.5.

| Methods | NTU RGB+D | | SBU - | PKU-MMD | |
|-----------|-----------|------|----------|---------|------|
| | CS | CV | | CS | CV |
| HCN-local | 83.9 | 89.7 | 96.8 | 91.1 | 93.9 |
| HCN | 86.5 | 91.1 | 98.6 | 92.6 | 94.2 |

Table 5: Comparison of HCN-local and HCN in terms of classification accuracy on the NTU RGB+D and SBU datasets and detection mAP on the PKU-MMD dataset.

archical co-occurrence feature learning framework is superior over other state-of-the-art LSTM and CNN-based methods. It is scalable to datasets of different sizes. And the learned features generalize well across various tasks, from action classification to action detection.

4.4 Importance of Global Co-occurrence Feature Learning

To further understand the behavior of global co-occurrence feature learning, we perform an ablation study. Specifically, we deliberately prepend a Transpose layer with parameter (0, 2, 1) before *conv1* so that the joint and coordinate dimensions are swapped. Then in subsequent layers after *conv2*, co-occurrence features are aggregated locally rather than globally, as in previous CNN-based methods [Du *et al.*, 2016; Ke *et al.*, 2017; Li *et al.*, 2017b]. The modified network is referred to as HCN-local. We train HCN-local models with the same hyper-parameters as HCN. The results are listed in Table 5. We can see that HCN consistently outperforms HCN-local on all of the three datasets. In particular, performance gain of HCN over HCN-local is more significant in the case of cross-subject than cross-view. This observation indicates that with global co-occurrence feature, the variation across different persons can be well addressed.

For a detailed comparison, we further investigate the per-category change in accuracy. Figure 6 shows the results, where the categories are sorted by accuracy gain. We can see that most actions get improved, in particular those involving long-range joint interaction. For example, over 10% absolute improvement is observed for *wear a shoe*, *clapping*, *wipe face* and *take-off a shoe*. For actions with no obvious joint interaction such as *nausea* and *typing on a keyboard*, global co-occurrence feature is not critical.

5 Discussion

What motivates us to explore the different levels of context aggregation is to figure out the importance of modeling interactions among joint points in action recognition.

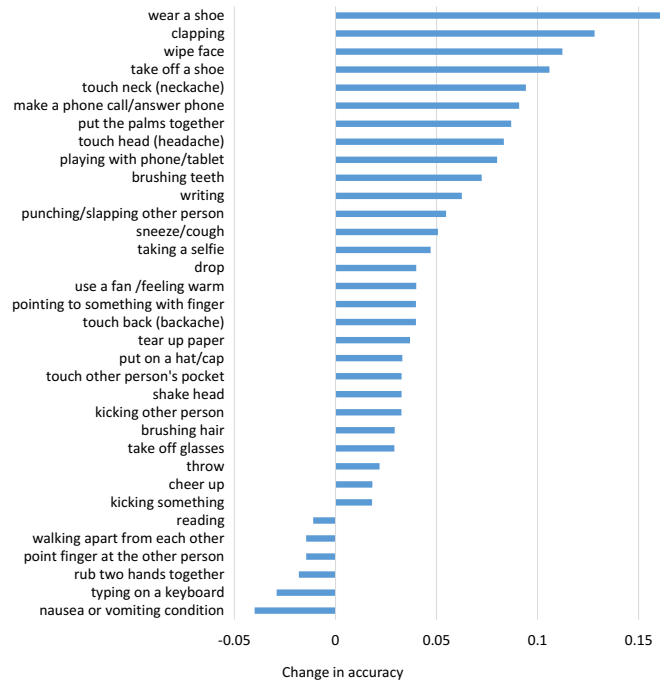


Figure 6: Per-category change in accuracy of HCN over HCN-local on the NTU RGB+D dataset in the cross-subject setting. For clarity only categories with change greater than 1% are shown.

Given a certain type of action of one or more subjects, do the whole or partial interactions of joints contribute to the recognition on earth? The answer from our experiments is initially counter-intuitive but makes sense, which has been proven by many analogous works [He *et al.*, 2016; Zhong *et al.*, 2017]. The so-called background context is an essential factor for boosting a task’s performance and it is the same case in action recognition. For recognition of a specific action, e.g. make a phone call, the joints of uninterest, say ankle, play a similar role as background context and the contribution is encoded implicitly with CNNs. This is just the insights which our method benefits from.

6 Conclusions

We present an end-to-end hierarchical co-occurrence feature learning framework for skeleton-based action recognition and detection. By exploiting the capability to global aggregation of CNN, we find that the joint co-occurrence features can be learned with CNN model simply and efficiently. In our method, we learn point-level features for each joint independently. Afterwards, we treat the feature of each joint as a channel of convolution layer to learn hierarchical co-occurrence features. And a two-stream framework is adopted to fuse the motion feature. Furthermore, we exploit the best way to deal with multi-person involved activities. Experiments on three benchmark datasets demonstrate that the proposed HCN model significantly improves the performance on both action recognition and detection tasks.

References

- [Du *et al.*, 2016] Yong Du, Yun Fu, and Liang Wang. Skeleton based action recognition with convolutional neural network. In *Pattern Recognition*, pages 579–583, 2016.
- [Girshick *et al.*, 2014] Ross B Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, pages 580–587, 2014.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.
- [Hussein *et al.*, 2013] Mohamed E. Hussein, Marwan Torki, Mohammad A. Gowayyed, and Motaz El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*, pages 639–44, 2013.
- [Ji *et al.*, 2014] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. In *ICMEW*, pages 1–6, 2014.
- [Jin and Choi, 2012] Sou Young Jin and Ho Jin Choi. Essential body-joint and atomic action detection for human activity recognition using longest common subsequence algorithm. In *ICCV*, pages 148–159, 2012.
- [Ke *et al.*, 2017] Qihong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A New Representation of Skeleton Sequences for 3D Action Recognition. In *CVPR*, July 2017.
- [Kingma and Ba, 2015] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.
- [Li *et al.*, 2016] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. Online human action detection using joint classification-regression recurrent neural networks. *ECCV*, pages 203–220, 2016.
- [Li *et al.*, 2017a] Bo Li, Huahui Chen, Yucheng Chen, Yuchao Dai, and Mingyi He. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In *ICMEW*, pages 613–616, July 2017.
- [Li *et al.*, 2017b] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Skeleton-based action recognition with convolutional neural networks. In *ICMEW*, pages 597–600, July 2017.
- [Liu *et al.*, 2016] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *ECCV*, pages 816–833, 2016.
- [Liu *et al.*, 2017] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *ACM Multimedia workshop*, 2017.
- [Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.
- [Sainath *et al.*, 2015] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584, April 2015.
- [Shahroudy *et al.*, 2016] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *CVPR*, June 2016.
- [Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS*, pages 568–576. Curran Associates, Inc., 2014.
- [Song *et al.*, 2017] 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 *AAAI*, pages 4263–4270, 2017.
- [Wang *et al.*, 2014] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3d human action recognition. *TPAMI*, 36(5):914, 2014.
- [Xia *et al.*, 2012] L. Xia, C. C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPRW*, pages 20–27, June 2012.
- [Xu *et al.*, 2017] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *ICCV*, Oct 2017.
- [Yun *et al.*, 2012] 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 *CVPRW*. IEEE, 2012.
- [Zhang *et al.*, 2017] 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. In *ICCV*, Oct 2017.
- [Zhong *et al.*, 2017] Qiaoyong Zhong, Chao Li, Yingying Zhang, Di Xie, Shicai Yang, and Shiliang Pu. Cascade region proposal and global context for deep object detection. *arXiv preprint arXiv:1710.10749*, 2017.
- [Zhu *et al.*, 2016] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, pages 3697–3703, 2016.