Structured Images for RGB-D Action Recognition

Pichao Wang[1]*, Shuang Wang[2]*, Zhimin Gao[1], Yonghong Hou[2]† and Wanqing Li[1]
[1]Advanced Multimedia Research Lab, University of Wollongong, Australia
[2]School of Electronic Information Engineering, Tianjin University, China
pw212@uowmail.edu.au,wangshuang1993@tju.edu.cn,zg126@uowmail.edu.au,houroy@tju.edu.cn,wanqing@uow.edu.au

Abstract

This paper presents an effective yet simple video representation for RGB-D based action recognition. It proposes to represent a depth map sequence into three pairs of structured dynamic images at body, part and joint levels respectively through bidirectional rank pooling. Different from previous works that applied one Convolutional Neural Network (ConvNet) for each part/joint separately, one pair of structured dynamic images is constructed from depth maps at each granularity level and serves as the input of a ConvNet. The structured dynamic image not only preserves the spatial-temporal information but also enhances the structure information across both body parts/joints and different temporal scales. In addition, it requires low computational cost and memory to construct. This new representation, referred to as Spatially Structured Dynamic Depth Images (S^{2}DDI), aggregates from global to fine-grained levels motion and structure information in a depth sequence, and enables us to fine-tune the existing ConvNet models trained on image data for classification of depth sequences, without a need for training the models afresh. The proposed representation is evaluated on five benchmark datasets, namely, MSRAction3D, G3D, MSRDailyActivity3D, SYSU 3D HOI and UTD-MHAD datasets and achieves the state-of-the-art results on all five datasets.

1. Introduction

Human action recognition from RGB-D (Red, Green, Blue and Depth) data has attracted increasing attention in computer vision in recent years due to the advantages of depth information over conventional RGB video, typically, being insensitive to illumination changes and reliable to estimate body silhouette and skeleton [26]. Since the first work of such a type [18] reported in 2010, many methods [1, 23, 41] have been proposed based on specific hand-crafted feature descriptors extracted from depth and/or skeleton data. Most of previous methods employ aggregation of local video descriptors to provide invariance to variations in the video. However, these methods may fail to capture important spatio-temporal and structure information at the same time.

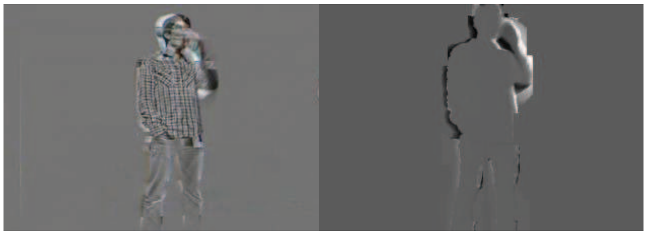

Figure 1: The differences between DI (left) and DDI (right) for action "drink" from MSRDailyActivity3D Dataset [34]. The DI has much interference of texture information on the body compared with DDI.

With the recent development of deep learning, a few methods [35, 36, 8, 38, 45, 24] have been developed based on Convolutional Neural Network (ConvNet) or Recurrent Neural Network (RNN). However, it remains unclear how video could be effectively represented and fed to deep neural networks for classification. For example, one can conventionally consider a video as a sequence of still images [40] with some form of temporal smoothness [27, 15] and feed them into a ConvNet, or extend ConvNet to a third, temporal dimension [16, 30] by replacing 2D filters with 3D ones, or regard the video as the output of a neural network encoder [29], or treat the video as a sequence of images and feed the sequence to a RNN [7, 8, 31, 45, 25, 24], or encode the video into motion images [35, 36, 38, 2, 37]. Which one among these and other possibilities would result in the best representation is not well understood.

Inspired by the promising performance of recently introduced rank pooling machine [10, 2, 9] on RGB video, this paper proposes to adopt rank pooling method to encode depth map sequences into dynamic images. Given a

*Both authors contributed equally to this work
†Corresponding author

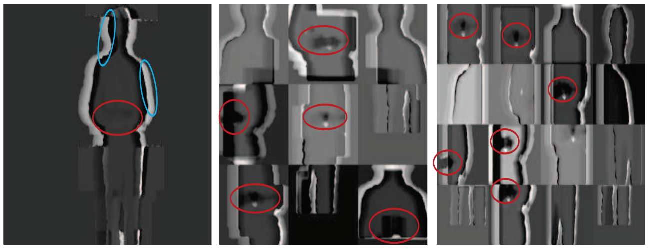

Figure 2: The three hierarchical structured DDIs for action "play game" from the MSRDailyActivity3D Dataset [34]. From left to right: structured body DDI, structured part DDI and structured joint DDI. The red circle denotes the hand motion need to be recognized while the blue one represents the large body swaying motion.

sequence of video frames, the rank pooling method returns a vector of parameters that aggregates spatio-temporal information contained in that video sequence. This vector of parameters is obtained by solving an unsupervised learning problem using RankSVM [28], where the order of frames in the video is considered as weak labels. Our empirical study has demonstrated that the rank pooling method works more effectively on depth map sequences than RGB ones. As shown in Figure 1, for the action "drink" from MSRDailyActivity3D dataset, the Dynamic Depth Images (DDI) generated by rank pooling is more informative (without interference of texture information on the body) than Dynamic Images (DI) proposed in [2] with respect to classification of actions, due to the fact of depth being insensitive to illumination and object appearance variations.

In work [9], the authors indicated that the rank pooling method employed by Fernado et al. [10, 11] is restricted to exploit long term dynamics. This paper further argues that the rank pooling method is also limited in the spatial domain. Due to the unsupervised learning process, the rank pooling method mainly encodes the salient global features in the temporal domain, without mining the discriminative motion patterns in both spatial and temporal domains simultaneously. It is also found that by applying the rank pooling method directly on the full body sequences, the small but discriminative motion information to recognize actions is usually suppressed by large motion, especially for these fine-grained actions where the local spatio-temporal subvolume motion is more important compared with the global motion of the whole sequences. As shown in Figure 2, the action "play game" from the MSRDailyActivity3D dataset, the large interference of body swaying motion occupies the motion in structured body DDI, and hands motion which is essential for recognition is not well highlighted in the DDI.

To address this problem, this paper proposes to apply rank pooling method on depth map sequences at three hierarchical spatial levels, namely, body level, part level and joint level based on our proposed non-scaling method. Different from previous method [6] that adopted one ConvNet for each human body part, it is proposed to construct one structured dynamic depth image as the input of a ConvNet for each level such that the structured dynamic images not only preserve the spatial-temporal information but also enhance the structure information, i.e. the coordination and synchronization of body parts over the period of the action. Such construction requires low computational cost and memory requirement. This representation, referred to as Spatially Structured Dynamic Depth Images (S^{2}DDI), aggregates motion and structure information from global to fine-grained levels for action recognition. In this way, the interference of large motion with small motion can be minimized. As shown in Figure 2, for action "play game", in the structured part DDI and structured joint DDI, the small hand motion is easy to recognize compared with that in structured body DDI. Moreover, the three structured dynamic images are complementary to each other, and an effective multiply score fusion method is adopted to improve the final recognition accuracy. The proposed image-based representation can take advantage of the available pre-trained models for standard ConvNet architectures without training millions of parameters afresh. It is evaluated on five benchmark datasets, namely, MSRAction3D [18], G3D [3], MSRDailyActivity3D [34], SYSU 3D HOI [13] and UTD-MHAD [4], and achieves the state-of-the-art results.

The key highlights of this paper are four folds. (1) A simple yet effective video representation, S^{2}DDI, is proposed for RGB-D video based action recognition by constructing three level structured dynamic depth images through bidirectional rank pooling. (2) An efficient non-scaling method is proposed to construct the S^{2}DDI. (3) The three level structured dynamic images aggregate motion and structure information from global to fine-grained levels for action recognition. A multiply score fusion method is adopted to improve the final action recognition accuracy. (4) The proposed method achieves state-of-the-art results on five benchmark datasets.

2. Spatially Structured Dynamic Depth Images

The proposed method mainly consists of three phases, as illustrated in Figure 3, the constructions of S^{2}DDI guided by skeletons, three weights-shared ConvNets training and multiply-score fusion for final action recognition. The first phase is an unsupervised learning process. It applies bidirectional rank pooling method to three hierarchical levels of a depth sequence to generate the structured DDIs, with each level of DDIs being represented by two motion images, forward (DDIF) and backward (DDIB). In the following sections, the three phases will be described in detail. The rank pooling method [2], that aggregates spatio-temporal infor-

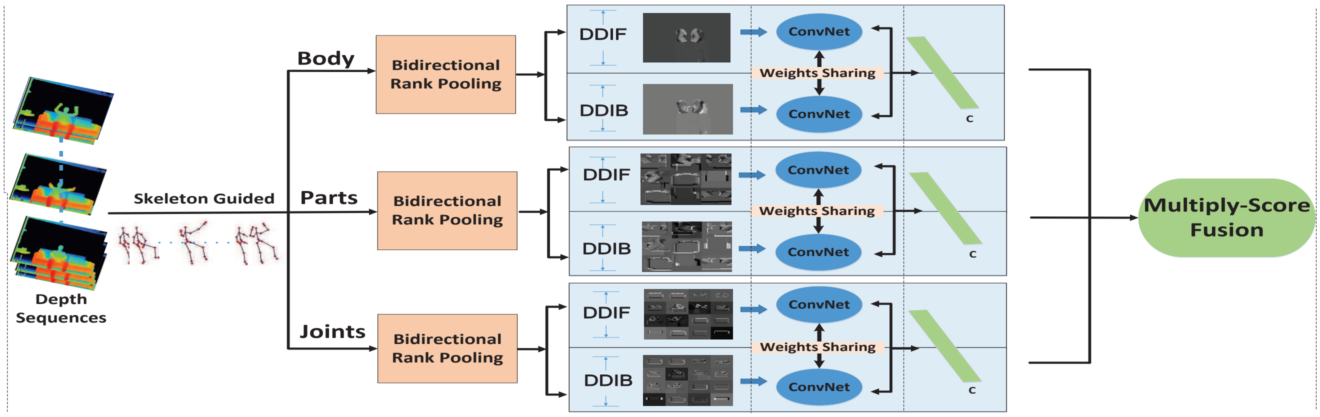

Figure 3: The framework of proposed method.

mation from one video sequence into one dynamic image, is also briefly summarized.

2.1. Bidirectional Rank Pooling

Rank pooling defines a function that maps a video clip into one feature vector [2]. A *rank pooling function* is formally defined as follows.

Rank Pooling Let a depth map sequence with k frames be represented as $< d_1, d_2, ..., d_t, ..., d_k >$, where d_t is the average of depth features over the frames up to t-timestamp. At each time t, a score $r_t = \omega^T \cdot d_t$ is assigned. The score satisfies $r_i > r_j \iff i > j$. In general, more recent frames are associated with larger scores. The process of rank pooling is to find ω^* that satisfies the following objective function:

$$\arg\min_{\omega} \frac{1}{2} \parallel \omega \parallel^2 + \lambda \sum_{i>j} \xi_{ij} \quad , \qquad (1)$$
$$s.t.\ \ \omega^T \cdot (d_i - d_j) \geq 1 - \xi_{ij}, \xi_{ij} \geq 0$$

where ξ_{ij} is a slack variable. Since the score r_i assigned to frame i is often defined as the order of the frame in the sequence, ω^* aggregates information from all of the frames in the sequence and can be used as a descriptor of the sequence. In this paper, the rank pooling is directly applied on the pixels of depth maps and the ω^* is of the same size as depth maps and forms a dynamic depth image (DDI).

Bidirectional Rank Pooling Different from work [2], this paper proposes to apply rank pooling bidirectionally, i.e. to apply the rank pooling forward and backward, to a sequence of depth maps. In the forward rank pooling, the r_i is defined in the same order as the time-stamps of the frames. In the backward rank pooling, r_i is defined in the reverse order of the time-stamps of the frames. When bidirectional rank pooling is applied to a sequence of depth maps, two DDIs, DDIF and DDIB, are generated. Since in rank pooling the averaged feature up to time t is used to classify frame t, the pooled feature is biased towards beginning frames of the depth sequence, hence, frames at the beginning has more influence to ω^*. This is not justifiable in action recognition as there is no prior knowledge on which frames are more important than other frames. The proposed bidirectional rank pooling is to reduce such bias.

2.2. Construction of S^{2}DDI

In the construction of S^{2}DDI, a human body is processed hierarchically at three spatial levels, namely, joint level, part level and body level. At each level, the body is divided into several components, and each component is composed of several joints. Specifically in this paper, there are 16 components at the joint level, each component containing 1 joint; at body part level, there are 9 components, each component consisting of 3 joints as defined below; at body level, the entire body is treated as a single component consisting of 16 joints. For each component, a Dynamic Depth Images (DDI) is generated by applying the rank pooling forward or backward to a sequence of depth patches that encloses the component. Two DDIs, i.e. DDIF and DDIB, at each level are constructed by simply stitching their component DDIs in a predefined arrangement. The three DDIFs and three DDIBs at body, part and joint levels together are referred to as S^{2}DDI. Note the rank pooling requires that the frames in a depth patch sequences be of same size.

Let $C = \{j_1, j_2, \ldots, j_n\}$ be a component consisting of n joints. Centered at each joint in the image plane, a depth patch, referred to as a joint patch, of size $p \times q$ pixels is cropped. A patch for the component C at frame t is extracted from the depth map based on the bounding box of C by keeping the depth values inside the joint patches and setting depth values outside of the joint patches but within the bounding box to zero. Notice that size of the component bounding box varies from frame to frame due to movement

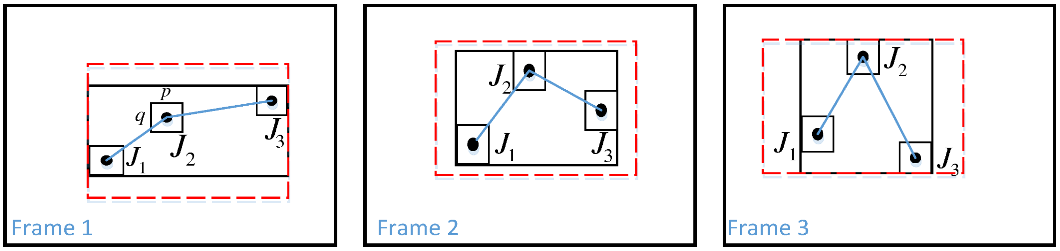

Figure 4: Illustration of non-scaled component patches of a component consisted of three joints $\{J_1, J_2, J_3\}$ from three frames. The solid black boxes are the bounding boxes of the component in each frame, while the dashed red box is the sequence-based bounding box of the component.

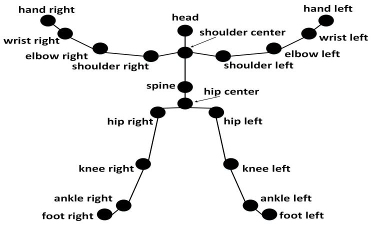

Figure 5: The joint configuration for Kinect V1 skeleton. The total number of joints is 20.

of the joints on one hand and, on the other hand, rank pooling requires the same size of the component patches over a sequence. Conventionally, the component patches would be scaled to a same size, referring to as *scaled patches*. The obvious disadvantage of such scaling is the distortion of the spatial information within a frame and, hence, motion information over the sequence. It is proposed in this paper to define a sequence-based component bounding box that is able to enclose the instances of the component over the sequence instead of using the bounding box at each frame. A component patch at each frame is then extracted by centering the sequence-based bounding box onto the component in the frame, referring to as *non-scaled patches*. In this way, the spatial and temporal distortion due to scaling can be eliminated. Figure 4 illustrates the extraction of non-scaling patches of a component consisting of three joints $\{J_1, J_2, J_3\}$ from three frames. In the figure, the solid black boxes are the bounding boxes of the component in each frame, while the dashed red box is the sequence-based bounding box of the component.

For the structured body DDIs, all the 20 joints are included in a single component. For the structured part DDIs, 9 components are defined according to the joint configuration in Figure 5 as follows.

C1	head,shoulder center,shoulder left
C2	head,shoulder center,shoulder right
C3	elbow left,wrist left,hand left
C4	elbow right,wrist right,hand right
C5	spine,hip center,hip right
C6	spine,hip center,hip left
C7	knee left,ankle left,foot left
C8	knee right,ankle right,foot right
C9	shoulder left,shoulder center,shoulder right

For the structured joint DDIs, the following 16 out of the 20 joints which usually bear relatively small noise are used and each joint forms a component.

hip center	spine	shoulder center	head
shoulder left	elbow left	hand left	shoulder right
elbow right	hand right	hand left	knee left
foot left	hip right	knee right	foot right

Different from the work in [6] that adopted one ConvNet for each component, all component DDIs at the part level are stitched together to form a structured part DDI and the component DDIs at the joint level are stitched together to form a structured joint DDI as shown in Figure 6. Such arrangement of component DDIs into a single structured DDI at each spatial level enables ConvNets to explore more effectively the structured information of an action than any late fusion approach.

2.3. Network Training

After the construction of structured DDIs at three levels, there are six dynamic images for each depth map sequence, as illustrated in Figure 3. Three ConvNets are trained on the three kinds of DDIs individually. The AlexNet [17] is adopted in this paper. The network weights are learned using the mini-batch stochastic gradient descent with the momentum being set to 0.9 and weight decay being set

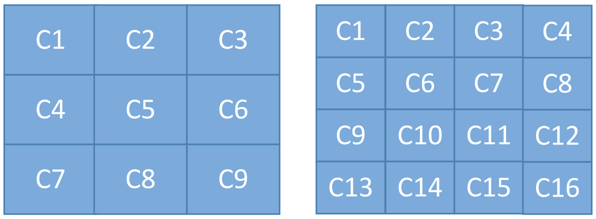

Figure 6: Stitching of component DDIs to a structured part DDI (left) and structured joint DDI (right).

to 0.0005. All hidden weight layers use the rectification (RELU) activation function. At each iteration, a mini-batch of 256 samples is constructed by sampling 256 shuffled training samples. All the images are resized to 256 × 256. The learning rate is set to 10^{-3} for fine-tuning the pre-trained models on ILSVRC-2012, and then it is decreased according to a fixed schedule, which is kept the same for all training sets. For each ConvNet, the training undergoes 3K iterations and the learning rate decreases every 1K iterations. For all experiments, the dropout regularization ratio is set to 0.5 in order to reduce complex co-adaptations of neurons in the nets.

2.4. Multiply-Score Fusion for Classification

Given a test depth video sequence (sample), three pairs of dynamic images (structured body DDIs, structured part DDIs and structured joint DDIs) are generated and fed into three different trained ConvNets. For each image pair, multiply-score fusion is used. The score vectors outputted by the weight sharing ConvNets are multiplied in an element-wise way, and then the resultant score vectors are normalized using L_1 norm. The three normalized score vectors are then multiplied in an element-wise fashion and the max score in the resultant vector is assigned as the probability of the test sequence. The index of this max score corresponds to the recognized class label.

3. Experiments

The proposed method is evaluated on five widely used benchmark RGB-D datasets [41], namely, MSRAction3D [18], G3D [3], MSRDailyActivity3D [34], SYSU 3D HOI [13] and UTD-MHAD [4] datasets. These five datasets cover a wide range of different types of actions including simple actions, actions for gaming, daily activities, human-object interactions and fine-grained activities. For the experiments on all datasets, the offset parameters (p, q) are empirically set. Specifically, for the construction of structured body DDI, they are $(80, 30)$ for head, two feet and two hands, and $(80, 50)$ for other joints. For structured part DDI, they are fixed to $(30, 30)$ for all joints. For structured joint DDI, they are set to be $(20, 30)$. In the following, ablations studies are conducted, and the results on the five datasets are presented and the detailed analysis on MSRDailyActivity3D Dataset are described. The detailed analysis based on the confusion matrices for the other four datasets are described in the supplementary material.

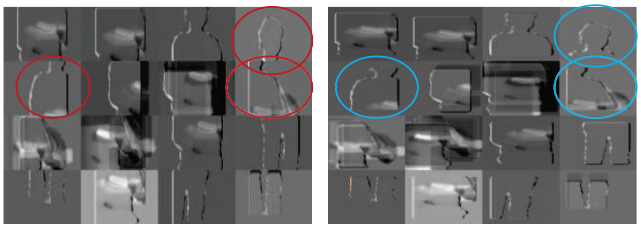

Figure 7: Illustration of using scaled component patches (left) and non-scaled component patches (right) for action "write on a paper" from MSRDailyActivity3D Dataset [34] for construction of structured joint DDI. The red circle denotes spatial distortion among human body while the blue one represents the preservation of aspect ratio among the parts and joints.

3.1. Effects of Design Choices

3.1.1 DDI vs. DI

Table 1 compares the performance of body DDI from depth and DI [2] from RGB for action recognition on the MSRDailyActivity3D dataset. Three DDIs are generated, one without foreground extraction, one using bounding box as foreground extraction, and the last one using the proposed method. From the results it can be seen that the DDI, especially the proposed structured body DDI, achieves much better results than DI. This verifies that the proposed method is robust to the noise in skeleton data.

Method	Accuracy
DI [2]	52.13%
DDI (without foreground extraction)	53.01%
DDI (with foreground bounding box)	58.75%
Structured body DDI (proposed)	61.00%

Table 1: Comparison of DDI and DI on the MSRDailyActivity3D dataset.

3.1.2 Scaled vs. Non-Scaled Component Patches in Constructing DDI

Experiments are conducted to evaluate on the S^{2}DDI constructed using scaled and non-scaled component patches. Table 2 shows the comparisons of these two methods in terms of recognition accuracy. It can be seen that using non-scaled patches greatly outperforms using scaled-patches mainly due to the elimination of distortion induced by the scaling.

Method	Accuracy
Structured part DDI (scaled)	67.88%
Structured joint DDI (scaled)	85.15%
S^2DDI (scaled)	87.04%
Structured part DDI (non-scaled)	81.88%
Structured joint DDI (non-scaled)	93.13%
S^2DDI (non-scaled)	97.50%

Table 2: Comparison of Construction of S^2DDI using scaled and non-scaled component patches on the MSRDailyActivity3D dataset.

3.1.3 Structured Images vs. Channel Fusion

To verify the effectiveness of proposed structured images, taking part level from MSRDailyActivity3D dataset for example, we compared the structured images with channel fusion using ConvNets and SIFT+FV+SVM [12], as in Table 3. It can be seen that the proposed structured part DDI not only outperforms the fusion of 9 separate DDIs, but also has computational advantage (1 channel vs. 9 channels). This is probably because the structural information is explored by the ConvNet from the structured part DDI. But such structural information can hardly be explored if each DDI is input to separate ConvNets and fused at the score level. From the comparisons we can also see that the proposed method can take advantages of the pre-trained models over ImagesNet for recognition compared with the traditional classifiers (e.g. SVM).

Method	Acc
Structured part DDI (ConvNet)	81.88%
Structured part DDI (SIFT+FV+SVM)	76.25%
9 channel part DDIs fusion (ConvNet)	72.81%
9 channel part DDIs fusion (SIFT+FV+SVM)	71.88%

Table 3: Comparison of structured images and channel fusion on the MSRDailyActivity3D dataset.

3.1.4 Traditional Rank pooling vs. Bidirectional Rank Pooling

Traditional pooling emphasizes the earlier frames in the pooling segment more than later frames. One of the key motivations of bidirectional rank pooling is to overcome this so that reversing cyclic movement patterns can be well distinguished. In addition, it effectively arguments the training data. The effectiveness of bidirectional rank pooling is shown in Table 4, taking MSRDailyActivity3D dataset for example.

3.1.5 Multiply vs. Average vs. Max Score Fusion

This paper adopts multiply score fusion method to improve the final accuracy on the three structured DDIs. The other two commonly used late score fusion methods are average and maximum score fusion. The comparisons among the three late score fusion methods are shown in Table 5. It can be seen that the multiply score fusion method achieves the best results on all the five datasets. This verifies that the three structured DDIs are likely to be statistically independent and carry complementary information.

Dataset	Score Fusion Method		
	Max	Average	Multiply
MSRAction3D	93.67%	97.56%	**100%**
G3D	94.83%	94.83%	**96.05%**
MSRDailyActivity3D	93.75%	95.00%	**97.50%**
SYSU 3D HOI	91.25%	94.17%	**95.42%**
UTD-MHAD	87.44%	88.54%	**89.04%**

Table 5: Comparison of three different late score fusion methods on the five datasets.

3.2. MSRAction3D Dataset

The MSRAction3D Dataset [18] contains 20 simple actions performed by 10 subjects facing the camera, with each subject performing each action 2 or 3 times. The same experimental setting adopted in [34] is followed, namely, the cross-subjects settings: subjects 1, 3, 5, 7, 9 for training and subjects 2, 4, 6, 8, 10 for testing. Table 6 lists the performance of the proposed method, as well as the results of several methods reported in recent three years. From the results, we can see that the proposed method can well recognize the simple actions, because the three hierarchical spatial dynamic image patches generated via bidirectional rank pooling can aggregate rich spatio-temporal information in each level, and the structure information of human body is explicitly exploited by the proposed non-scaled component patches and structured motion images.

3.3. G3D Dataset

Gaming 3D Dataset (G3D) [3] focuses on real-time action recognition in gaming scenario. It contains 10 subjects performing 20 gaming actions. For this dataset, the first 4 subjects are used for training, the fifth for validation and the remaining 5 subjects for testing, following the configuration in [21]. Table 7 compares the performance of the proposed method with that reported in [21, 38]. It can been seen that S^2DDI achieves better results.

Method	body DDI	part DDI	joint DDI	fusion
Traditional rank pooling(SIFT+FV+SVM)	42.50%	68.75%	80.00%	86.25%
Traditional rank pooling(ConvNets)	59.38%	80.00%	89.37%	95.63%
Bidirectional rank pooling(SIFT+FV+SVM)	49.69%	76.25%	81.25%	88.75%
Bidirectional rank pooling(ConvNets)	61.00%	81.88%	93.130%	97.50%

Table 4: Comparison of traditional rank pooling and bidirectional rank pooling on the MSRDailyActivity3D dataset.

Method	Accuracy
Lie Group [32]	89.48%
HCM [19]	93.00%
SNV [39]	93.09%
Range Sample [20]	95.62%
MTDMM + FV [5]	95.97%
Structured body DDI	79.18%
Structured part DDI	83.83%
Structured joint DDI	95.40%
S^2DDI	**100%**

Table 6: Comparison of the proposed method with existing methods on the MSRAction3D dataset.

Method	Accuracy
LRBM [21]	90.50%
JTM [38]	94.24%
Structured body DDI	74.81%
Structured part DDI	89.97%
Structured joint DDI	93.62%
S^2DDI	**96.05%**

Table 7: Comparison of the proposed method with previous methods on the G3D dataset.

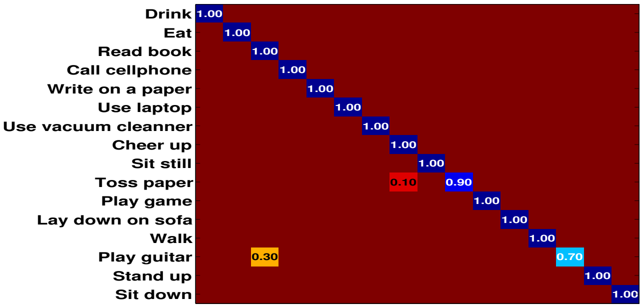

Figure 8: Confusion matrix for S^2DDI on the MSRDailyActivity3D dataset.

3.4. MSRDailyActivity3D Dataset

The MSRDailyActivity3D Dataset [34] has 16 activities and there are 10 subjects and each subject performed each activity twice, one in standing position and the other in sitting position. Most activities in this dataset involve human-object interactions. The same cross-subject experimental setting as in [34] is adopted. Compared with existing methods on this dataset, the results in Table 8 show that the proposed method is superior for dataset having fine-grained human-object interaction actions.

Method	Accuracy
IPM [44]	83.30%
WHDMMs+ConvNets [36]	85.00%
SNV [39]	86.25%
DS+DCP+DDP+JOULE-SVM [13]	95.00%
Range Sample [20]	95.63%
MFSK+BoVW [33]	95.70%
Structured body DDI	61.00%
Structured part DDI	81.88%
Structured joint DDI	93.13%
S^2DDI	**97.50%**

Table 8: Comparison of the proposed method with previous methods on the MSRDailyActivity3D dataset.

The confusion matrices for structured body DDI, structured part DDI and structured joint DDI are shown in Figure 9 and S^2DDI in Figure 8. From the confusion matrix, we can see that the structured body DDI confuses most activities, especially "Eat", "Read book", "Write on a paper" and "play game". This is because the structured body DDIs of these activities have similar shapes, and the motion to be recognized is very small compared with the interference of large body swaying motion, as illustrated in Figure 2. But as the granularity increases, most of the activities can be well recognized, because the fine-grained small motion is enhanced in the patches of parts and joints. By fusion of the three levels, most of the activities are better recognized, which reflects that the three structured motion images are complementary to each other. Compared with the method proposed in [13], ours can better recognize "Drink", "Read book","Write on a paper" and "Play game" activities, due to the capability of both global to fine-grained motion and structure information aggregation of our method. These activities are very easily confused by global motion information aggregation method. However, the skeleton guided decomposition can not work well for human-large object interaction. For example, due to the large size of guitar, the pro-

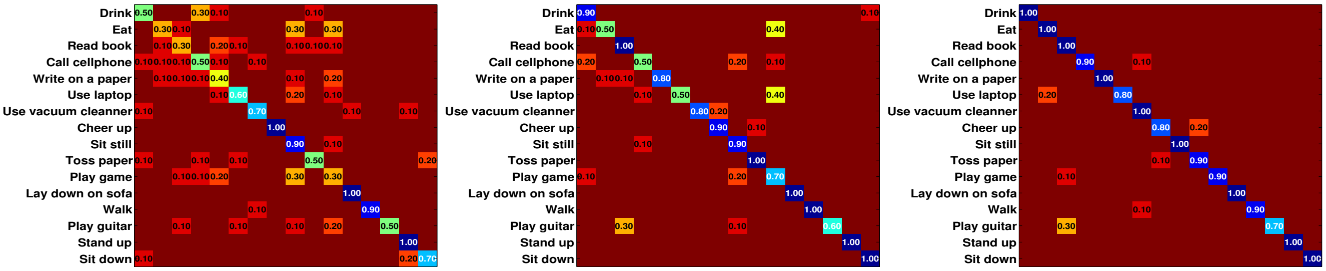

Figure 9: Confusion matrix for structured body DDI (left), structured part DDI (middle) and structured joint DDI (right) on MSRDailyActivity3D Dataset.

posed method loses much object information and confused "play guitar" with "Read book". This can be improved by setting larger extension around the joints.

3.5. SYSU 3D HOI Dataset

The SYSU 3D Human-Object Interaction Dataset (SYSU 3D HOI Dataset) [13] was collected to focus on human-object interactions. There are 40 subjects performing 12 different activities. For each activity, each participants manipulate one of the six different objects: phone, chair, bag, wallet, mop and besom. Table 9 compares the performances of the proposed method and that of existing methods on this dataset using cross-subject settings as in [13]. It can bee seen that, our proposed method outperforms previous methods largely. It should be noticed that on this dataset, the structured joint DDI achieves the best performance. From the confusion matrices in the supplementary material we can see that the "Taking from wallet" action is greatly confused in structured body and part DDIs, that affects the final performance of S^{2}DDI.

Method	Accuracy
HON4D [22]	79.22%
DS+DCP+DDP+MTDA [42]	84.21%
DS+DCP+DDP+JOULE-SVM [13]	84.89%
structured body DDI	65.00%
structured part DDI	85.83%
structured joint DDI	**95.83%**
S^{2}DDI	95.42%

Table 9: Comparison of the proposed method with previous approaches on SYSU 3D HOI Dataset.

3.6. UTD-MHAD Dataset

UTD-MHAD [4] contains 27 actions performed by 8 subjects (4 females and 4 males) with each subject perform each action 4 times. For this dataset, cross-subjects protocol is adopted as in [4], namely, the data from the subject numbers 1, 3, 5, 7 used for training while 2, 4, 6, 8 used for testing. The results are shown in Table 10. It can be seen that even the structured joint DDI itself can achieve better result than previous methods. From the performances on the five datasets, we can conclude that as the granularity increases, the proposed method achieves higher accuracy.

Method	Accuracy
WHDMMs+ConvNets [36]	73.95%
ELC-KSVD [43]	76.19%
Kinect & Inertial [4]	79.10%
Cov3DJ [14]	85.58%
JTM [38]	85.81%
structured body DDI	66.05%
structured part DDI	78.70%
structured joint DDI	86.81%
S^{2}DDI	**89.04%**

Table 10: Comparison of the proposed method with previous approaches on UTD-MHAD Dataset.

4. Conclusion and Future Work

In this paper, an effective yet simple video representation, S^{2}DDI, constructed using bidirectional rank pooling is presented for 3D action recognition. This representation not only preserves the motion information but also enhances the structure information, and aggregates motion and structure information from global to fine-grained for final action recognition. Such image-based representation takes advantage of the available trained deep ConvNets models to fine-tune on small training data. The proposed method has been evaluated on five popular datasets with large variations among the actions and achieves state-of-the-art results.The performance of the proposed method is expected to be further improved if large training data are available. Our future work is to extend the proposed representation to multiple persons for recognition of actions and interactions.

References

[1] J. K. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014. 1

[2] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *CVPR*, 2016. 1, 2, 3, 5

[3] V. Bloom, D. Makris, and V. Argyriou. G3D: A gaming action dataset and real time action recognition evaluation framework. In *CVPRW*, 2012. 2, 5, 6

[4] C. Chen, R. Jafari, and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *ICIP*, 2015. 2, 5, 8

[5] C. Chen, M. Liu, B. Zhang, J. Han, J. Jiang, and H. Liu. 3D action recognition using multi-temporal depth motion maps and fisher vector. In *IJCAI*, 2016. 7

[6] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In *ICCV*, 2015. 2, 4

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1

[8] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015. 1

[9] B. Fernando, P. Anderson, M. Hutter, and S. Gould. Discriminative hierarchical rank pooling for activity recognition. In *CVPR*, 2016. 1, 2

[10] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *CVPR*, 2015. 1, 2

[11] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Rank pooling for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 2

[12] Z. Gao, L. Wang, L. Zhou, and J. Zhang. Hep-2 cell image classification with deep convolutional neural networks. *IEEE journal of biomedical and health informatics*, 21(2):416–428, 2017. 6

[13] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 2015. 2, 5, 7, 8

[14] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. In *IJCAI*, 2013. 8

[15] D. Jayaraman and K. Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *CVPR*, 2016. 1

[16] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, 2013. 1

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4

[18] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *CVPRW*, 2010. 1, 2, 5, 6

[19] I. Lillo, J. Carlos Niebles, and A. Soto. A hierarchical pose-based approach to complex action understanding using dictionaries of actionlets and motion poselets. In *CVPR*, 2016. 7

[20] C. Lu, J. Jia, and C.-K. Tang. Range-sample depth feature for action recognition. In *CVPR*, 2014. 7

[21] S. Nie, Z. Wang, and Q. Ji. A generative restricted boltzmann machine based method for high-dimensional motion data modeling. *Computer Vision and Image Understanding*, pages 14–22, 2015. 6, 7

[22] O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 2013. 8

[23] L. L. Presti and M. La Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130–147, 2016. 1

[24] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*, 2016. 1

[25] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015. 1

[26] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 1

[27] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[28] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004. 2

[29] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015. 1

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. 1

[31] V. Veeriah, N. Zhuang, and G.-J. Qi. Differential recurrent neural networks for action recognition. In *ICCV*, 2015. 1

[32] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a lie group. In *CVPR*, 2014. 7

[33] J. Wan, G. Guo, and S. Z. Li. Explore efficient local features from RGB-D data for one-shot learning gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1626–1639, Aug 2016. 7

[34] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012. 1, 2, 5, 6, 7

[35] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. O. Ogunbona. Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring. In *ACM MM*, 2015. 1

[36] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona. Action recognition from depth maps using deep con-

volutional neural networks. *Human-Machine Systems, IEEE Transactions on*, 46(4):498–509, 2016. 1, 7, 8

[37] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *CVPR*, 2017. 1

[38] P. Wang, Z. Li, Y. Hou, and W. Li. Action recognition based on joint trajectory maps using convolutional neural networks. In *ACM MM*, 2016. 1, 6, 7, 8

[39] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014. 7

[40] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 1

[41] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang. RGB-D-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, 2016. 1, 5

[42] Y. Zhang and D. Y. Yeung. Multi-task learning in heterogeneous feature spaces. In *AAAI*, 2011. 8

[43] L. Zhou, W. Li, Y. Zhang, P. Ogunbona, D. T. Nguyen, and H. Zhang. Discriminative key pose extraction using extended lc-ksvd for action recognition. In *DICTA*, 2014. 8

[44] Y. Zhou, B. Ni, R. Hong, M. Wang, and Q. Tian. Interaction part mining: A mid-level approach for fine-grained action recognition. In *CVPR*, 2015. 7

[45] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *AAAI*, 2016. 1