Learning Action Recognition Model From Depth and Skeleton Videos

Hossein Rahmani, and Mohammed Bennamoun
School of Computer Science and Software Engineering,
The University of Western Australia
hossein.rahmani@uwa.edu.au, mohammed.bennamoun@uwa.edu.au

Abstract

Depth sensors open up possibilities of dealing with the human action recognition problem by providing 3D human skeleton data and depth images of the scene. Analysis of human actions based on 3D skeleton data has become popular recently, due to its robustness and view-invariant representation. However, the skeleton alone is insufficient to distinguish actions which involve human-object interactions. In this paper, we propose a deep model which efficiently models human-object interactions and intra-class variations under viewpoint changes. First, a human body-part model is introduced to transfer the depth appearances of body-parts to a shared view-invariant space. Second, an end-to-end learning framework is proposed which is able to effectively combine the view-invariant body-part representation from skeletal and depth images, and learn the relations between the human body-parts and the environmental objects, the interactions between different human body-parts, and the temporal structure of human actions. We have evaluated the performance of our proposed model against 15 *existing techniques on two large benchmark human action recognition datasets including NTU RGB+D and UWA3DII. The Experimental results show that our technique provides a significant improvement over state-of-the-art methods.*

1. Introduction

Automatic human action recognition from videos is a significant research problem with wide applications in various fields such as smart surveillance, health and medicine, sports and recreation. Depth cameras, such as Microsoft Kinect, have become popular for this task because, depth images are robust to variations in illumination, clothing color and texture, and 3D human joint positions can be extracted from a single depth image due to the development of a real-time human skeleton tracking framework [28].

The depth sensor based human action recognition research can be broadly divided into three categories including skeleton data [6,12,25,33,34,38,40], depth images [15, 17,18,20,21,23,37,39] and depth+skeleton [19,26,27,36] based methods. Although depth based approaches achieve impressive results on most RGB-Depth action recognition datasets, their performance drops sharply in scenarios where the humans significantly change their spatial locations, and the temporal extents of the activities significantly vary [15]. On the other hand, limiting the learning into skeleton based features cannot deliver high recognition accuracy in action recognition, because depth visual appearances of human body-parts provide discriminative information, and most of the usual human actions are defined based on the interaction of the body with other objects. For example, *drinking* and *eating snacks* actions have a very similar skeleton motion. Thus, additional information, such as depth images, is required to distinguish such actions. The straightforward method for combining depth and skeleton data (feature fusion) is to concatenate these different types of features [19, 35, 36]. In this fashion, achieving the optimal combination of features for an accurate classification cannot be guaranteed.

Moreover, a practical system should be able to recognize human actions from novel and unseen viewpoints (generalization). However, unlike 3D skeletal data, view-invariant representation of depth images is a challenging task [20–22, 24]. This is because the depth images of a human performing an action appear quite different when observed from different viewpoints. Thus, how to effectively represent these depth images in a view-invariant space and combine them with estimated 3D skeleton data is a significant research problem which remains under explored.

Furthermore, in human actions, body joints move together in groups. Each group can be considered as a set of body-parts, and actions can be interpreted as interactions of different body-parts. Thus, the most discriminative interactions corresponding to different actions need to be exploited for better recognition. Moreover, human actions may have a specific temporal structure. Modeling the temporal structure of action videos is also crucial for the action recognition problem. Most existing depth sensor based methods [23,33–36] model the temporal variations of videos us-

ing Fourier Temporal Pyramid (FTP) and/or Dynamic Time Warping (DTW) which results in a two-step system that typically performs worse than an end-to-end system [9]. Some other methods [6, 12, 25], use Recurrent Neural Networks (RNNs) or extensions, such as Long Short Term Memory (LSTM) networks, to model the temporal variations of action videos. However, CNN+RNN/LSTM models introduce a large number of additional parameters, and thus need much more training videos which are expensive to label [7].

This paper proposes a deep model for human action recognition from depth and skeleton data to deal with the above mentioned challenges in an end-to-end learning framework. First, we propose a deep CNN model which transfers the depth appearance of human body-parts to a shared view-invariant space. Learning this deep CNN requires a large dataset containing a variety of human body-parts in different actions observed from many viewpoints. Rahmani *et al*. [23] showed that a model learned on synthetic depth human body images can be generalized to real depth images without the need of fine-tuning. Thus, we generate a large training dataset by synthesizing human body-parts from different views. More importantly, we propose a framework which is able to **1)** effectively combine information from depth and skeletal data, **2)** capture the relations between the human body-parts and the environmental objects, **3)** model the interactions between different human body-parts, and **4)** learn the temporal structure of human actions in an end-to-end learning framework.

Our main contributions include the following three aspects. **First**, this paper proposes a model for view-invariant appearance representation of human body-parts. **Second**, we propose an end-to-end learning human action recognition model, which is shown through our experiments to be well suitable for the depth sensor based human action recognition task. **Third**, the proposed method simultaneously learns to combine features from different modalities, *i.e*. depth and skeleton, capture the interactions between different human body-parts for different actions, and model the temporal structure of different human actions.

The proposed method is evaluated on two large benchmark datasets including NTU RGB+D [25] and UWA3D Multiview Activity II [20] datasets. The first dataset contains more than $56K$ sequences captured simultaneously by three Kinect cameras from three different views, and the second dataset consists of 30 human actions captured from four different viewpoints. This dataset is challenging because the videos were captured by a Kinect camera at four different times from four different viewpoints. Our extensive experimental results show that the proposed method is able to achieve a significantly better recognition accuracy compared with the state-of-the-art methods.

2. Related Work

Human action recognition has been explored from different aspects during the recent years. In this section, we limit our review to the most recent related approaches, which can be divided into three different categories, namely depth, skeleton and skeleton+depth video based methods.

Depth Videos: Most existing depth video based action recognition methods use global features such as silhouettes and space-time volume information. For instance, Oreifej and Liu [15] proposed a spatio-temporal depth video representation by extending histogram of oriented 3D normals [30] to 4D by adding the time derivative. Yang and Tian [39] extended HON4D by concatenating the 4D normals in the local neighbourhood of each pixel as its descriptor. However, these methods are not view-invariant. Recently, Rahmani *et al*. [23] proposed to generate synthetic depth training data of human poses in order to train a deep CNN model that transfers human poses, acquired from different views, to a view-invariant space. They used group sparse Fourier Temporal Pyramid to encode the temporal variations of human actions. Such holistic methods may fail in scenarios where the human significantly changes her/his spatial position, or the temporal extent of the activities significantly vary [15]. Some other methods [20,21,37] use local features where a set of interest points are detected, and then, the depth features are extracted from the local neighbourhood of each interest point. For example, DSTIP [37] localizes activity related interest points from depth videos by suppressing flip noise. This approach may fail when the action execution speed is faster than the flip of the signal caused by sensor noise. Recently, Rahmani *et al*. [20, 21] introduced to directly process the sequence of pointclouds corresponding to an action video. They proposed a view-invariant interest point detector and descriptor by calculating Principle Component Analysis (PCA) at every point which is a computationally expensive process [23]. However, these methods use hand-crafted features and implicitly assume that the viewpoint does not change significantly [23].

Skeleton Videos: Due to the development of real-time human skeleton tracker from a single depth image [28], motion patterns can be effectively encoded using the positional dynamics of joints [38] or body-parts [33]. For example, Yang and Tian [38] used pairwise 3D joint position differences in each frame and temporal differences across frames to represent an action. Zanfir *et al*. [40] proposed a moving pose descriptor for capturing postures and skeleton joints. Vemulapalli *et al*. [33, 34] utilized rotations and translations to represent the 3D geometric relationships of body-parts in Lie group, and then employed Dynamic Time Warping (DTW) and Fourier Temporal Pyramid (FTP) to model the temporal dynamics. To avoid using hand-engineered features, deep learning based methods have also been proposed. For instance, HBRNN [6] divided the entire skeleton

to five major groups of joints which were passed through hierarchical RNNs. The hidden representation of the final RNN was fed to a softmax classifier layer for action classification. Differential LSTM [32] introduced a new gating inside LSTM to discover patterns within salient motion patterns. Shahroudy *et al*. [25] proposed a part-aware extension of LSTM by splitting the memory cell of the LSTM into part-based sub-cells and pushing the network towards learning the long-term context representations individually for each part.The output of the network was learned over the concatenated part-based memory cells followed by the common output gate. More recently, Liu *et al*. [12] introduced a spatial-temporal LSTM to jointly learn both spatial and temporal relationships among joints.

Skeleton+Depth Videos: Although skeleton based methods achieve impressive action recognition accuracies on human action datasets, it is not sufficient to only use the skeletal data to model actions, especially when the actions have very similar skeleton motion and include the interactions between the subject and other objects [36]. Therefore, skeleton+depth based approaches are becoming an effective way to describe activities of interactions. Rahmani *et al*. [19] proposed a set of random forests to fuse the spatio-temporal depth and joints features. Wang *et al*. [36] proposed to compute the histogram of occupancy patterns of a fixed region around each joint in each frame of action videos. In the temporal dimension, low frequency Fourier components were used as features for classification. Recently, Shahroudy *et al*. [27] proposed hierarchical mixed norms to fuse different features and select the most informative body joints. Our proposed approach also falls in this category. However, unlike the above mentioned works, the framework proposed in this paper is robust to significant changes in viewpoints. Moreover, the proposed model learns the relations between the human body-parts and the environmental objects, along with the temporal structure of actions in an end-to-end learning framework.

3. Proposed Approach

This section presents a view-invariant human action recognition model using a depth camera. The proposed architecture is illustrated in Fig. 1. Based on the depth images and the estimated 3D joint positions, we propose a new model which is able to **1)** transfer the human body-parts to a shared view-invariant space, **2)** capture the relations between the human body-parts and environmental objects for human-object interaction, and **3)** learn the temporal structure of actions in an end-to-end learning framework. The details of the proposed architecture are given below.

3.1. Human Body-Part Representation

In this section, we introduce two approaches to represent skeletal information of a body-part (subsection 3.1.1) and to transfer the depth appearances of human body-parts (subsection 3.1.2) to a shared view-invariant space.

3.1.1 Body-Part Skeletal Representation

Given a skeletal data of a human pose performing an action, all 3D joint coordinates are transformed from the real-world coordinate system to a person-centric coordinate system by placing the hip center at the origin. This operation makes the skeletons invariant to absolute location of the human in the scene. Given a skeleton as reference, all the other skeletons are normalized without changing their joint angles such that their body-part lengths are equal to the corresponding lengths of the reference skeleton. This normalization makes the skeletons scale-invariant. The skeletons are also rotated such that the ground plane projection of the vector, from the left hip to the right hip, is parallel to the global x-axis. This rotation makes the skeletons view-invariant.

As shown in Fig. 2, let $S = (V, E))$ be a human body skeleton, where $V = \{v_1, v_2, \cdots, v_N\}$ represents the set of body joints, and $E = \{e_1, e_2, \cdots, e_M\}$ denotes the set of body-parts. Vemulapalli *et al*. [33] showed that the relative geometry of a pair of human body-parts e_n and e_m can be represented in a local coordinate system attached to the other. The local coordinate system of body-part e_n is calculated by minimum rotation so that its stating joint becomes the origin and it coincides with the x-axis. Then, we can compute the translation vector, $\vec{d}_{m,n}(t)$, and the rotation matrix, $R_{m,n}$, from e_m to the local coordinate system of e_n. Thus, the relative geometry between e_m and e_n at time instance t can be described using

$$P_{m,n}(t) = \begin{bmatrix} R_{m,n}(t) & \vec{d}_{m,n}(t) \\ 0 & 1 \end{bmatrix}. \quad (1)$$

Using the relative geometry between pairs of body-parts, we represent a body-part e_i at time instance t using $C_i(t) = (P_{1,i}(t), P_{2,i}(t), \cdots, P_{i-1,i}(t), P_{i+1,i}(t), \cdots, P_{M,i}(t)) \in SE(3) \times \cdots \times SE(3)$, where $SE(3)$ denotes Special Euclidean group and M is the number of body-parts. Then, the representation of the body-part e_i at time instance t, $C_i(t)$, is mapped from Special Euclidean group to its Lie algebra in vector representation using

$$\begin{aligned} \zeta_i(t) = [&vec(\mathbf{log}(P_{1,i}(t))), vec(\mathbf{log}(P_{2,i}(t))), \\ &\cdots vec(\mathbf{log}(P_{i-1,i}(t))), vec(\mathbf{log}(P_{i+1,i}(t))), \\ &\cdots vec(\mathbf{log}(P_{M,i}(t)))], \end{aligned} \quad (2)$$

where **log** denotes the usual matrix logarithm.

For a body-part i at a time instance t, $\zeta_i(t)$ is a vector of dimension $6(M-1)$. Hence, we represent each body-part at time t as a $6(M-1)$ dimensional vector (as shown in Fig. 1).

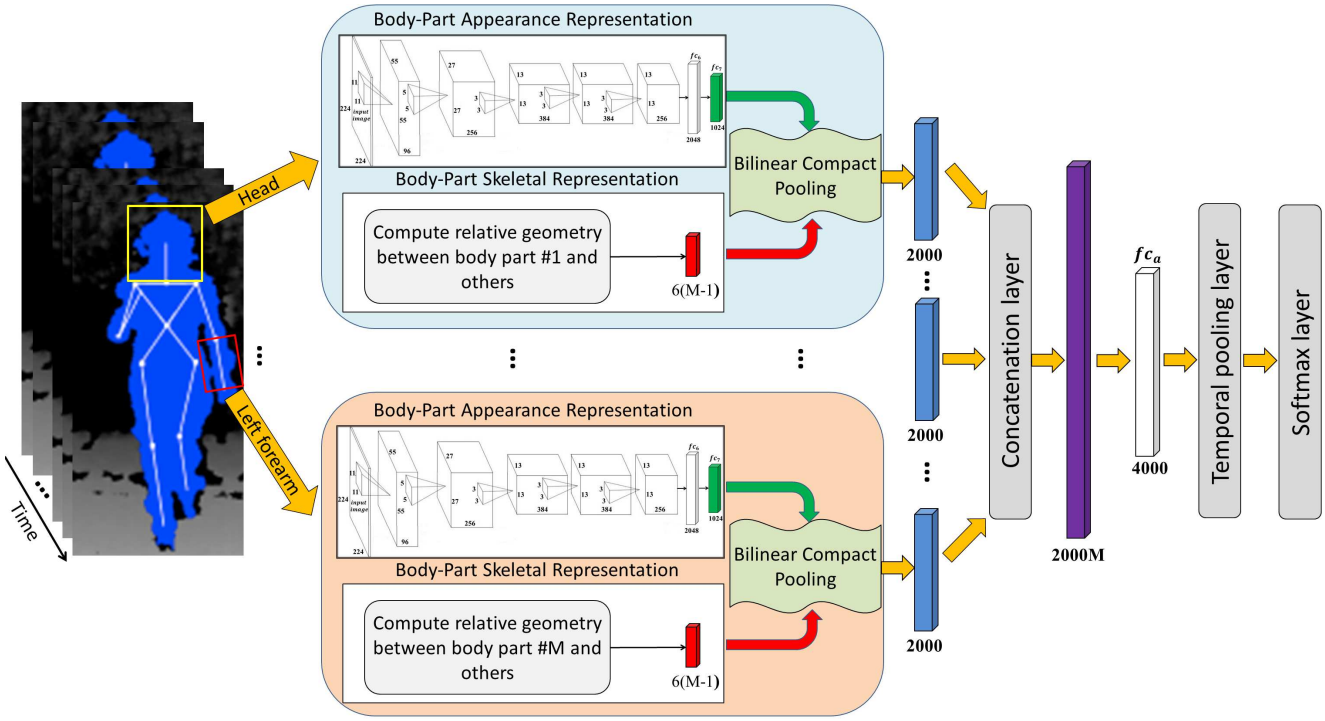

Figure 1: Architecture of the proposed model. Given a sequence of depth images and their corresponding skeleton data, the relative geometry between every body-part and others are calculated (see Section 3.1.1). The bounding boxes containing human body-parts are passed through the same body-part appearance representation model (see Section 3.1.2) to extract view-invariant depth information (the outputs of fc_7 layer). Bilinear compact pooling is then applied to the skeletal and appearance representations of each body-part (see Section 3.2.1) which results in a compact 2000-dimensional feature vector. The compact feature vectors from all M body-parts are concatenated to from a $2000 \times M$-dimensional vector and then passed through a fully-connected layer, fc_a, which encodes the interactions between different human body-parts (see Section 3.2.2). Finally, a sequence of 4000-dimensional feature vectors corresponding to the sequence of depth frames are passed through the temporal pooling layer (see Section 3.2.3) to extract a fixed-length feature vector for classification.

3.1.2 Body-Part Appearance Representation

Unlike 3D skeletons, view-invariant depth appearance representation of human body-parts is challenging. This is because the depth images of human body-parts appear quite different when observed from different viewpoints. To overcome this problem, we learn a single deep CNN model for all human body-parts to transfer them to a shared view-invariant space. However, learning such a deep CNN requires a large training dataset containing a large number of human body-parts, performing a variety of actions observed from many viewpoints. Our solution is to generate synthetic training data since Rahmani and Mian [23] showed that a model trained on synthetic depth images is able to generalize real depth images without the need for retraining or fine tuning the model.

The set of all possible human actions, and thus, body-part appearances is extremely large. Therefore, we propose a method to select the most representative human body-parts. We use the CMU mocap database [2] containing over $200K$ poses of subjects performing a variety of actions. However, many body-parts are quite similar even for different actions. In order to find the most representative body-parts, we first normalize the mocap skeleton data using the approach given in the previous section. We consider the 3D vector connecting the hip centre to spine as a reference. The rotation required to take a body-part to the reference is used as the body-part features. Using Euclidean Distance between the Euler Angles [10], we apply k-means clustering to each body-part features independently. Given the fact that human body-parts have different degrees of movement, we extracted different numbers of clusters for different body-parts. This algorithm results in 480 representative body-parts.

To generate depth images of the selected body-parts, we first use the open source MakeHuman software [3] to synthesize different realistic 3D human shapes and the open source Blender package [1] to fit 3D human shapes to mocap data. Then, the full human bodies corresponding to 480 representative body-parts are rendered from 108 differ-

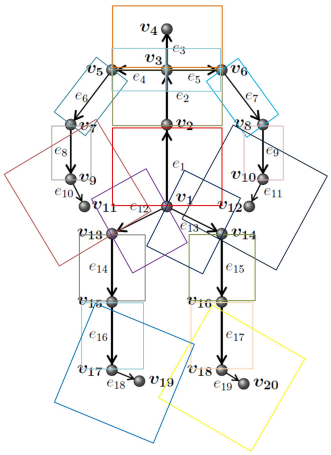

Figure 2: An example skeleton consisting of 20 joints and 19 body-parts. The bounding boxes show regions of interest corresponding to different body-parts.

ent viewpoints and the bounding boxes containing the representative body-parts are used as training data. In total, 480×108 depth images corresponding to 480 representative body-parts viewed from 108 different views are generated. Note that the length and width of a bounding box containing a body-part are appropriately chosen to cover the whole body-part. For example, we set the length and width of a bounding box surrounding the *left hand* to twice of the body-part's length. This is because humans usually interact with their environment, *e.g.* objects and other human, by hands and feet. Thus, a larger bounding box is able to cover both the hand and the object. For instance, Fig. 1 shows two bounding boxes corresponding to *head* and *left forearm* of a human performing an action, and Fig. 2 illustrates the bounding boxes corresponding to 19 body-parts of an example skeleton.

We propose a view-invariant human body-part representation model that learns to transfer body-parts from any view to a shared view-invariant space. The model is a deep convolutional neural network (CNN) whose architecture is as follows: $C(11,96,4) \rightarrow RL \rightarrow P(3,2) \rightarrow N \rightarrow C(5,256,1) \rightarrow RL \rightarrow P(3,2) \rightarrow N \rightarrow C(3,384,1) \rightarrow RL \rightarrow C(3,384,1) \rightarrow RL \rightarrow C(3,256,1) \rightarrow RL \rightarrow P(3,2) \rightarrow FC(2048) \rightarrow RL \rightarrow D(0.5) \rightarrow FC(1024) \rightarrow RL \rightarrow D(0.5) \rightarrow FC(480)$, where $C(k,n,s)$ denotes a convolutional layer with kernel size $k \times k$, n filters and a stride of s, $P(k,s)$ is a max pooling layer of kernel size $k \times k$ and stride s, N is a normalization layer, RL a rectified linear unit, $FC(n)$ a fully connected layer with n filters and $D(r)$ a dropout layer with dropout ratio r. We refer to the fully-connected layers as fc_6, fc_7, and fc_8, respectively. During learning, a softmax loss layer is added at the end of the network.

The generated synthetic depth images corresponding to each representative body-part i, where $i = 1, \cdots, 480$, from all 108 viewpoints are assigned the same class label i. Thus, our training dataset consists of 480 human body-part classes. We initialize the convolution layers of the proposed CNN with the model trained on depth images of full human body from [23]. We fine-tune the proposed CNN model with back-propagation and use an initial learning rate of 0.001 for the convolution layers and 0.01 for the fully-connected layers. We use a momentum of 0.9 and a weight decay of 0.0005. We train the network for $30K$ iterations. After training, the last fully-connected (fc_8) and softmax layers are removed and the remaining layers, as shown in Fig. 1, are used as view-invariant body-part appearance representation model.

3.2. Learning End-to-End Human Action Recognition Model

So far, we have represented the skeletal data and depth images corresponding to human body-parts in two different view-invariant spaces, Lie algebra and fc_7, respectively. However, a human action is carried out by moving body-parts and interacting between them along time. To take this information into account, we propose an end-to-end learning model, which is shown in Fig. 1. The details of each step are given below.

3.2.1 Bilinear Compact Pooling Layer

Bilinear classifiers take the outer product of two vectors $x_1 \in \mathbb{R}^{n_1}$ and $x_2 \in \mathbb{R}^{n_2}$ and learn a model W, *i.e.* $z = W[x_1 \otimes x_2]$, where $\otimes$ denotes the outer product $x_1 {x_2}^T$ and $[.]$ denotes linearizing the matrix in a vector. Thus, all elements of both vectors x_1 and x_2 are interacted with each other in a multiplicative way. However, the outer product of vectors x_1 and x_2, when n_1 and n_2 are large, results in an infeasible number of parameters to learn in W. To overcome this problem, Gao *et al.* [8] proposed a bilinear compact pooling for a single modality which projects the outer product to a lower dimensional space and also avoids computing the outer product directly. This idea is based on the Tensor Sketch algorithm of [16].

Let $x_s^{(i)}(t) \in \mathbb{R}^{6(M-1)}$ and $x_d^{(i)}(t) \in \mathbb{R}^{1024}$ denote the feature vectors obtained from the body-part skeletal (Section 3.1.1) and appearance (Section 3.1.2) representation models for the i-th body-part at time t, respectively. We apply bilinear compact pooling on $x_s^{(i)}(t)$ and $x_d^{(i)}(t)$ to combine them efficiently and expressively. We set the projection dimension, d to 2000. This process results in a 2000-dimensional feature vector for every body-part.

3.2.2 Fully-Connected Layer

In order to encode the interactions between different human body-parts, we propose to first concatenate the com-

pact feature vectors of all M body-parts to form a $2000M$-dimensional feature vector and then pass them through a fully-connected layer consisting of 4000 units. Thus, the proposed model extracts a 4000-dimensional feature vector for every depth image in an action video.

3.2.3 Temporal Pooling Layer

The straightforward CNN-based method for encoding a video is to apply temporal max pooling or temporal average pooling over the frames. However, such temporal pooling methods are not able to capture the time varying information of the video. To overcome this problem, we use the recently proposed rank-pool operator [4]:

$$\underset{\mathbf{u}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{u}||^2 + \frac{C}{2}\sum_{t=1}^{T} max(0, |t - \mathbf{u}^T\mathbf{v_t}| - \epsilon)^2, \quad (3)$$

where $\mathbf{v_t} \in \mathbb{R}^{4000}$ denotes the outputs of the fc_a layer corresponding to the frame at time t, $\vec{\mathbf{v}} = (\mathbf{v_1}, \cdots, \mathbf{v_T})$ denotes a sequence of feature vectors and $\mathbf{u}$ is a fixed-length representation of the video which is used for classification. This temporal rank pooling layer attempts to capture the order of elements in the sequence by finding a vector $\mathbf{u}$ such that $\mathbf{u}^T\mathbf{v_i} < \mathbf{u}^T\mathbf{v_j}$ for all $i < j$, which can be solved using regularized support vector regression (SVR) in Eq. (3).

3.2.4 Classification Layer and Learning

The aim of the classification layer is to assign one of the action class labels to the sequence descriptor $\mathbf{u}$. In this work, we use soft-max classifier. Given a dataset of video-label pairs,$\{(\vec{\mathbf{x}}^{(i)}, y^{(i)})\}_{i=1}^{n}$, we jointly estimate the parameters of the temporal pooling layer Eq. (3) and the soft-max classification function as follows:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{n} -\log P(y^{(i)}|\vec{\mathbf{x}}^{(i)}) + R(\theta)$$
$$\text{subject to } \mathbf{u}^{(i)} \in Eq.\ (3), \quad (4)$$

where $-\log P(y^{(i)}|\vec{\mathbf{x}}^{(i)})$ is the cross-entropy loss for the soft-max classifier and $R(\theta)$ is $\ell 2$-norm regularization of the model parameters. Eq. (4) is a bilevel optimization problem and its derivative with respect to any parameters in the model can be computed by applying the chain rule [7]. We use stochastic gradient descent (SGD) to learn all parameters jointly.

4. Experiments

We evaluate our proposed model on two large benchmark datasets including NTU RGB+D [25] and UWA3D Multiview Activity II [20]. We compare our performance to the state-of-the-art action recognition methods including Comparative Coding Descriptor (CCD) [5], Histogram of Oriented Gradients (HOG^2) [14], Discriminative Virtual Views (DVV) [11], Continuous Virtual Path (CVP) [41], Histogram of Oriented 4D Normals (HON4D) [15], Actionlet Ensemble (AE) [36], Lie Algebra Relative Pairs (LARP) [33, 34], Super Normal Vector (SNV) [39], Histogram of Oriented 3D Pointcloud (HOPC) [20], Hierarchical recurrent neural network (HBRNN) [6], STA-LSTM [29], Spatio-Temporal LSTM with Trust Gates (ST-LSTM+TG) [12], Human Pose Model with Temporal Modelling (HPM+TM) [23], Long-Term Motion Dynamics (LTMD) [13], and Deep Learning on Lie Groups (LieNet) [9]. The baseline results are reported from their original papers or [12, 23].

In addition to other compared methods, we report the accuracy of our defined baseline methods, including:

- Baseline 1 (appearance only): The vectors of size 2000 in our proposed architecture are replaced by their corresponding body-part appearance representations,
- Baseline 2 (skeletal only): The vectors of size 2000 in our proposed architecture are replaced by by their corresponding body-part skeletal representations,
- Baseline 3 (max-pooling for encoding temporal variation): Rank-pooling layer is replaced by max-pooling,
- Baseline 4 (replacing bilinear pooling by concatenation): The proposed view-invariant body-part appearance representation is combined with the corresponding skeletal representation by a concatenation layer, followed by a fully connected and rank pooling layers.

We used the MatConvNet toolbox [31] as the deep learning platform. We train the network using stochastic gradient descent, and set the learning rate, momentum and weight decay 10^{-3}, 0.9 and 5×10^{-4}, respectively. We also use a dropout rate of 0.5 for fully connected layers. It is important to note that we set the learning rate of the proposed view-invariant visual body-part representation model to zero. This is because we have already learned this model using synthetic training depth images to transfer human body-part images to the view-invariant space.

4.1. NTU RGB+D Action Recognition Dataset

This dataset [25] is currently the largest depth based action recognition dataset. It is collected by Microsoft Kincet v2 and contains 60 action classes including daily actions, pair actions, and medical conditions performed by 40 subjects from three different views. Figure 3 shows sample frames from this dataset. The dataset consists of more than 56000 sequences. The large intra-class and viewpoint variations make this dataset very challenging. This dataset

has two standard evaluation protocols [25] including cross-subject and cross-view. Following the cross-subject protocol, we split the 40 subjects into training and testing sets. Each set contains samples captured from different views performed by 20 subjects. Following the cross-view protocol in [25], we use all the samples of camera 1 and 2 for training and samples of the remaining camera for testing.

Table 1 shows the performances of various methods on this dataset. The recognition accuracy of our proposed method significantly outperforms our defined baselines and all existing methods. Our defined baseline methods achieve lower accuracy than our proposed method in both cross-subject and cross-view settings. This demonstrates the effectiveness of the feature fusion and end-to-end learning of the proposed method.

Since this dataset provides rich samples for training deep models, the RNN-based methods, *e.g.* ST-LSTM+TG [12], achieve high accuracy. However, our method achieves 6% and 5.4% higher accuracy than ST-LSTM+TG [12] in cross-subject and cross-view settings, respectively. This result shows the effectiveness of our method to tackle challenges such as viewpoint variations in large scale of data.

Table 1: Comparison of action recognition accuracy (%) on the the NTU RGB+D dataset [25].

Method	Cross-Subject	Cross-View
HON4D [15]	30.6	7.3
SNV [39]	31.8	13.6
HOG^2 []	32.24	22.27
LARP-SE [33]	50.1	52.8
LARP-SO [34]	52.1	53.4
HBRNN [6]	59.1	64.0
Part-aware LSTM [25]	62.9	70.3
STA-LSTM [29]	73.4	81.2
Deep RNN [25]	56.3	64.1
Deep LSTM [25]	60.7	67.3
ST-LSTM+TG [12]	69.2	77.7
LTMD [13]	66.2	–
LieNet [9]	61.4	67.0
Baseline 1	67.3	74.1
Baseline 2	58.8	62.7
Baseline 3	69.0	76.5
Baseline 4	68.1	75.7
Ours	**75.2**	**83.1**

It is important to emphasize that the proposed body-part appearance representation model (Section 3.1.2) was learned from synthetic depth images of human body-parts generated from a small number of human poses. A search for many human actions/poses such as *wear on glasses, taking a selfie, typing on a keyword,* and *tear up paper*, from the NTU RGB-D dataset [25] returns no results in the CMU mocap dataset which is used to train the body-part appearance representation model. However, the proposed method achieves a high classification accuracy. For instance, the accuracies obtained for *wear on glasses, taking a selfie, typing on a keyword*, and *tear up paper* are 90.1%, 85.6%, 80.0%, and 85.7%, respectively.

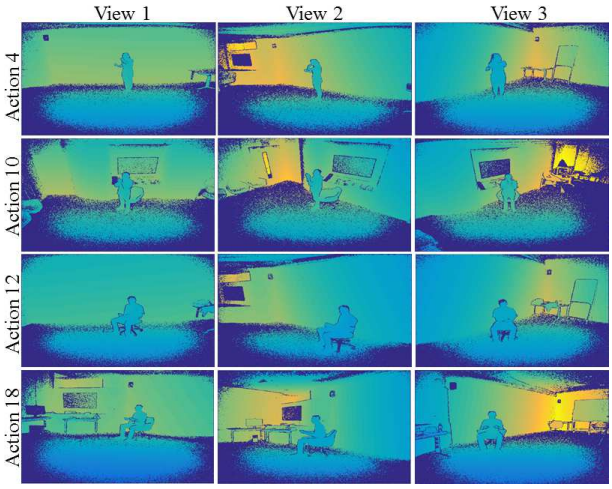

Figure 3: Sample frames from the NTU RGB+D [25] dataset.

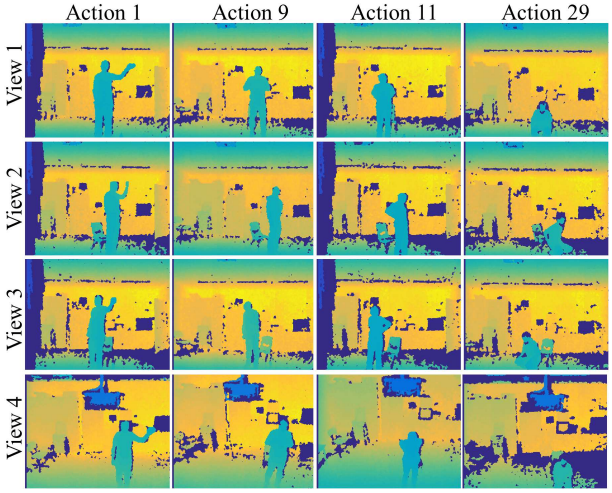

Figure 4: Sample frames from the UWA3DII [20] dataset.

4.2. UWA3D Multiview Activity II Dataset

This dataset [20] consists of a variety of daily-life human actions performed by 10 subjects with different scales. It includes 30 action classes, such as *two hand waving*, *holding chest*, *irregular walking*, and *coughing*. Each subject performed 30 actions 4 times. Each time the action was captured from a different viewpoint (front, top, left and right side views). Video acquisition from multiple views was not synchronous. There are, therefore, variations in the actions besides viewpoints. This dataset is challenging because of varying viewpoints, self-occlusion and high simi-

Table 2: Comparison of action recognition accuracy (%) on the UWA3D Multiview ActivityII dataset. Each time two views are used for training and the remaining two views are individually used for testing.

Training views	V_1 & V_2		V_1 & V_3		V_1 & V_4		V_2 & V_3		V_2 & V_4		V_3 & V_4		Mean
Test view	V_3	V_4	V_2	V_4	V_2	V_3	V_1	V_4	V_1	V_3	V_1	V_2	
CCD [5]	10.5	13.6	10.3	12.8	11.1	8.3	10.0	7.7	13.1	13.0	12.9	10.8	11.2
DVV [11]	23.5	25.9	23.6	26.9	22.3	20.2	22.1	24.5	24.9	23.1	28.3	23.8	24.1
CVP [41]	25.0	25.6	25.5	28.2	24.7	24.0	23.0	24.5	26.6	23.3	30.3	26.8	25.6
HON4D [15]	31.1	23.0	21.9	10.0	36.6	32.6	47.0	22.7	36.6	16.5	41.4	26.8	28.9
Actionlet [36]	45.0	40.4	35.1	36.9	34.7	36.0	49.5	29.3	57.1	35.4	49.0	29.3	39.8
LARP [33]	49.4	42.8	34.6	39.7	38.1	44.8	53.3	33.5	53.6	41.2	56.7	32.6	43.4
SNV [39]	31.9	25.7	23.0	13.1	38.4	34.0	43.3	24.2	36.9	20.3	38.6	29.0	29.9
HOPC [21]	52.7	51.8	59.0	57.5	42.8	44.2	58.1	38.4	63.2	43.8	66.3	48.0	52.2
HPM+TM [23]	80.6	80.5	75.2	82.0	65.4	72.0	77.3	67.0	83.6	81.0	83.6	74.1	76.9
Ours	**86.8**	**87.0**	**80.7**	**89.1**	**78.1**	**80.9**	**86.5**	**79.3**	**85.1**	**86.9**	**89.4**	**80.0**	**84.2**

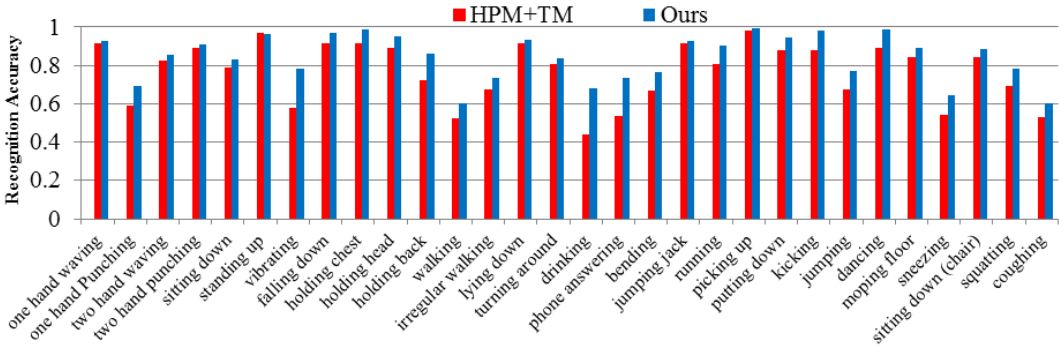

Figure 5: Per class recognition accuracy of our proposed method and HPM+TM [23] on the UWA3D Multiview ActivityII [20] dataset.

larity among actions. Moreover, in the top view, the lower part of the body was not properly captured because of occlusion. Figure 4 shows four sample actions observed from 4 viewpoints.

We follow [20] and use the samples from two views as training data, and the samples from the remaining views as test data. Table 2 summarizes our results. Our proposed model significantly outperforms the state-of-the-art methods on all view pairs. The overall accuracies of depth based methods, such as HOPC [20] and CCD [5], and Depth+Skeleton based methods, such as HON4D [15], SNV [39], and Actionlet [36], are low because depth appearances of many actions look very different across view changes. However, our method achieves 84.2% average recognition accuracy which is about 7.3% higher than than the nearest competitor, HPM+TM [23].

Figure 5 compares the class specific action recognition accuracies of our proposed approach and the nearest competitor, HPM+TM [23]. The proposed method achieves better recognition accuracy on all action classes excluding *standing up*. For example, our method achieves 24% and 20% higher accuracies than HPM+TM [23] for *drinking* and *phone answering*, respectively. This is because our proposed model is able to capture the interactions between human body-parts and environmental objects.

Notice that for many actions in the UWA3D Multiview ActivityII dataset such as *holding chest, holding head, holding back, sneezing* and *coughing*, there are no similar actions in the CMU mocap dataset. However, our method still achieves high recognition accuracies for these actions. This demonstrates the effectiveness and generalization ability of our proposed model.

5. Conclusion

We proposed an end-to-end learning model for action recognition from depth and skeleton data. The proposed model learned to fuse features from depth and skeletal data, capture the interactions between body-parts and/or interactions with environmental objects, and model the temporal structure of human actions in an end-to-end learning framework. In order to make our method robust to viewpoint changes, we introduced a deep CNN which transfers visual appearance of human body-parts acquired from different unknown views to a view-invariant space. Experiments on two large benchmark datasets showed that the proposed approach outperforms existing state-of-the-art.

Acknowledgment We thank NVIDIA for their K40 GPU donation.

References

[1] Blender: a 3D modelling and rendering package. http://www.blender.org/. 4

[2] CMU motion capture database. http://mocap.cs.cmu.edu/. 4

[3] MakeHuman: an open source 3D computer graphics software. http://www.makehuman.org/. 4

[4] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould. Dynamic image networks for action recognition. In *CVPR*, 2016. 6

[5] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *ECCVW*, 2012. 6, 8

[6] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015. 1, 2, 6, 7

[7] B. Fernando and S. Gould. Learning end-to-end video classification with rank-pooling. In *ICML*, 2016. 2, 6

[8] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *CVPR*, 2016. 5

[9] Z. Huang, C. Wan, T. Probst, and L. V. Gool. Deep learning on Lie groups for skeleton-based action recognition. In *CVPR*, 2017. 2, 6, 7

[10] D. Q. Huynh. Metrics for 3D rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, 35(2):155–164, 2009. 4

[11] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In *CVPR*, 2012. 6, 8

[12] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal LSTM with trust gates for 3D human action recognition. In *ECCV*, 2016. 1, 2, 3, 6, 7

[13] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. In *CVPR*, 2017. 6, 7

[14] E. Ohn-Bar and M. Trivedi. Joint angles similarities and HOG^2 for action recognition. In *CVPRW*, 2013. 6

[15] O. Oreifej and Z. Liu. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences. In *CVPR*, 2013. 1, 2, 6, 7, 8

[16] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2013. 5

[17] H. Rahmani, D. Q. Huynh, A. Mahmood, and A. Mian. Discriminative human action classification using locality-constrained linear coding. *Pattern Recognition Letters*, 2016. 1

[18] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Action classification with locality-constrained linear coding. In *ICPR*, 2014. 1

[19] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Real time action recognition using histograms of depth gradients and random decision forests. In *WACV*, pages 626–633, 2014. 1, 3

[20] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. *TPAMI*, 2016. 1, 2, 6, 7, 8

[21] H. Rahmani, A. Mahmood, D. Q Huynh, and A. Mian. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In *ECCV*, 2014. 1, 2, 8

[22] H. Rahmani and A. Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *CVPR*, 2015. 1

[23] H. Rahmani and A. Mian. 3D action recognition from novel viewpoints. In *CVPR*, 2016. 1, 2, 4, 5, 6, 8

[24] H. Rahmani, A. Mian, and M. Shah. Learning a deep model for human action recognition from novel viewpoints. *TPAMI*, 2017. 1

[25] 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, 2, 3, 6, 7

[26] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *TPAMI*, 2017. 1

[27] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang. Multimodal multipart learning for action recognition in depth videos. *TPAMI*, 2016. 1, 3

[28] 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*, pages 1297–1304, 2011. 1, 2

[29] S. Song, C. Lan, J. Xing, W. Zeng, , and J. Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017. 6, 7

[30] S. Tang, X. Wang, X. Lv, T. Han, J. Keller, Z. He, M. Skubic, and S. Lao. Histogram of oriented normal vectors for object recognition with a depth sensor. In *ACCV*, 2012. 2

[31] A. Vedaldi and K. Lenc. MatConvNet - Convolutional Neural Networks for MATLAB. In *ACM International Conference on Multimedia*, 2015. 6

[32] V. Veeriah, N. Zhuang, and G. Qi. Differential recurrent neural networks for action recognition. In *ICCV*, 2015. 3

[33] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a Lie group. In *CVPR*, 2014. 1, 2, 3, 6, 7, 8

[34] R. Vemulapalli and R. Chellappa. Rolling rotations for recognizing human actions from 3d skeletal data. In *CVPR*, 2016. 1, 2, 6, 7

[35] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290 –1297, 2012. 1

[36] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3D human action recognition. *PAMI*, 2013. 1, 3, 6, 8

[37] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recongition using depth camera. In *CVPR*, 2013. 1, 2

[38] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive bayes nearest neighbor. In *CVPRW*, 2012. 1, 2

[39] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014. 1, 2, 6, 7, 8

[40] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013. 1, 2

[41] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi. Cross-view action recognition via a continuous virtual path. In *CVPR*, 2013. 6, 8