# NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis

Amir Shahroudy[†,‡] amir3@ntu.edu.sg
Jun Liu[†] jliu029@ntu.edu.sg
Tian-Tsong Ng[‡] ttng@i2r.a-star.edu.sg
Gang Wang[†,*] wanggang@ntu.edu.sg

† School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
‡ Institute for Infocomm Research, Singapore

## Abstract

*Recent approaches in depth-based human activity analysis achieved outstanding performance and proved the effectiveness of 3D representation for classification of action classes. Currently available depth-based and RGB+D-based action recognition benchmarks have a number of limitations, including the lack of training samples, distinct class labels, camera views and variety of subjects. In this paper we introduce a large-scale dataset for RGB+D human action recognition with more than 56 thousand video samples and 4 million frames, collected from 40 distinct subjects. Our dataset contains 60 different action classes including daily, mutual, and health-related actions. In addition, we propose a new recurrent neural network structure to model the long-term temporal correlation of the features for each body part, and utilize them for better action classification. Experimental results show the advantages of applying deep learning methods over state-of-the-art hand-crafted features on the suggested cross-subject and cross-view evaluation criteria for our dataset. The introduction of this large scale dataset will enable the community to apply, develop and adapt various data-hungry learning techniques for the task of depth-based and RGB+D-based human activity analysis.*

## 1. Introduction

Recent development of depth sensors enabled us to obtain effective 3D structures of the scenes and objects [13]. This empowers the vision solutions to move one important step towards 3D vision, *e.g.* 3D object recognition, 3D scene understanding, and 3D action recognition [1].

Unlike the RGB-based counterpart, 3D video analysis suffers from the lack of large-sized benchmark datasets. Yet there are no any sources of publicly shared 3D videos such as YouTube to supply "in-the-wild" samples. This limits our ability to build large-sized benchmarks to evaluate and compare the strengths of different methods, especially the recent data-hungry techniques like deep learning approaches. To the best of our knowledge, all the current 3D action recognition benchmarks have limitations in various aspects.

First is the small number of subjects and very narrow range of performers' ages, which makes the intra-class variation of the actions very limited. The constitution of human activities depends on the age, gender, culture and even physical conditions of the subjects. Therefore, variation of human subjects is crucial for an action recognition benchmark.

Second factor is the number of the action classes. When only a very small number of classes are available, each action class can be easily distinguishable by finding a simple motion pattern or even the appearance of an interacted object. But when the number of classes grows, the motion patterns and interacting objects will be shared between classes and the classification task will be more challenging.

Third is the highly restricted camera views. For most of the datasets, all the samples are captured from a front view with a fixed camera viewpoint. For some others, views are bounded to fixed front and side views, using multiple cameras at the same time.

Finally and most importantly, the highly limited number of video samples prevents us from applying the most advanced data-driven learning methods to this problem. Although some attempts have been done [9, 42], they suffered from overfitting and had to scale down the size of learning parameters; as a result, they clearly need many more samples to generalize and perform better on testing data.

To overcome these limitations, we develop a new large-scale benchmark dataset for 3D human activity analysis. The proposed dataset consists of $56,880$ RGB+D video samples, captured from 40 different human subjects, using Microsoft Kinect v2. We have collected RGB videos, depth sequences, skeleton data (3D locations of 25 major body joints), and infrared frames. Samples are captured in 80 distinct camera viewpoints. The age range of the subjects in our dataset is from 10 to 35 years which brings more realis-

四种限制:
1. 同一动作不同年龄、性别、文化对象的差异
2. 动作类别越多、越难
3.摄像机角度和数量
4.视频样本数量不足

一个动作内部

深度图像 = 普通的RGB三通道彩色图像 + Depth Map
一幅画的尺寸是1024*768，深度为16，则它的数据量为1.5M。

红外帧

*Corresponding author

| Datasets | | Samples | Classes | Subjects | Views | Sensor | Modalities | Year |
|---|---|---|---|---|---|---|---|---|
| MSR-Action3D | [19] | 567 | 20 | 10 | 1 | N/A | D+3DJoints | 2010 |
| CAD-60 | [34] | 60 | 12 | 4 | - | Kinect v1 | RGB+D+3DJoints | 2011 |
| RGBD-HuDaAct | [23] | 1189 | 13 | 30 | 1 | Kinect v1 | RGB+D | 2011 |
| MSRDailyActivity3D | [38] | 320 | 16 | 10 | 1 | Kinect v1 | RGB+D+3DJoints | 2012 |
| Act4$^2$ | [6] | 6844 | 14 | 24 | 4 | Kinect v1 | RGB+D | 2012 |
| CAD-120 | [18] | 120 | 10+10 | 4 | - | Kinect v1 | RGB+D+3DJoints | 2013 |
| 3D Action Pairs | [25] | 360 | 12 | 10 | 1 | Kinect v1 | RGB+D+3DJoints | 2013 |
| Multiview 3D Event | [43] | 3815 | 8 | 8 | 3 | Kinect v1 | RGB+D+3DJoints | 2013 |
| Online RGB+D Action | [46] | 336 | 7 | 24 | 1 | Kinect v1 | RGB+D+3DJoints | 2014 |
| Northwestern-UCLA | [40] | 1475 | 10 | 10 | 3 | Kinect v1 | RGB+D+3DJoints | 2014 |
| UWA3D Multiview | [28] | ∼900 | 30 | 10 | 1 | Kinect v1 | RGB+D+3DJoints | 2014 |
| Office Activity | [41] | 1180 | 20 | 10 | 3 | Kinect v1 | RGB+D | 2014 |
| UTD-MHAD | [4] | 861 | 27 | 8 | 1 | Kinect v1+WIS | RGB+D+3DJoints+ID | 2015 |
| UWA3D Multiview II | [26] | 1075 | 30 | 10 | 5 | Kinect v1 | RGB+D+3DJoints | 2015 |
| **NTU RGB+D** | | **56880** | **60** | **40** | **80** | **Kinect v2** | **RGB+D+IR+3DJoints** | **2016** |

Table 1. Comparison between NTU RGB+D dataset and some of the other publicly available datasets for 3D action recognition. Our dataset provides many more samples, action classes, human subjects, and camera views in comparison with other available datasets for RGB+D action recogniton.

tic variation to the quality of actions. Although our dataset is limited to indoor scenes, due to the operational limitation of the acquisition sensor, we provide the ambiance (易变的氛围) inconstancy by capturing in various background conditions. This large amount of variation in subjects and views makes it possible to have more accurate cross-subject and cross-view evaluations for various 3D-based action analysis methods.

The proposed dataset can help the community to move steps forward in 3D human activity analysis and makes it possible to apply data-hungry methods such as deep learning techniques for this task.

As another contribution, inspired by the physical characteristics of human body motion, we propose a novel part-aware extension of the long short-term memory (LSTM) model [14]. Human actions can be interpreted as interactions of different parts of the body. In this way, the joints of each body part always move together and the combination of their 3D trajectories form more complex motion patterns. By splitting the memory cell of the LSTM into part-based sub-cells, the recurrent network will learn the long-term patterns specifically for each body part and the output of the unit will be learned from the combination of all the sub-cells.

Our experimental results on the proposed dataset shows the clear advantages of data-driven learning methods over state-of-the-art hand-crafted features.

The rest of this paper is organized as follows: Section 2 explores the current 3D-based human action recognition methods and benchmarks. Section 3 introduces the proposed dataset, its structure, and defined evaluation criteria. Section 4 presents our new part-aware long short-term memory network for action analysis in a recurrent neural network fashion. Section 5 shows the experimental evaluations of state-of-the-art hand-crafted features alongside the proposed recurrent learning method on our benchmark, and section 6 concludes the paper.

## 2. Related work

In this section we briefly review publicly available 3D activity analysis benchmark datasets and recent methods in this domain. Here we introduce a limited number of the most famous ones. For a more extensive list of current 3D activity analysis datasets and methods, readers are referred to these survey papers [47, 1, 5, 12, 21, 45, 3].

### 2.1. 3D activity analysis datasets

After the release of Microsoft Kinect [48], several datasets are collected by different groups to perform research on 3D action recognition and to evaluate different methods in this field.

MSR-Action3D dataset [19] was one of the earliest ones which opened up the research in depth-based action analysis. The samples of this dataset were limited to depth sequences of gaming actions *e.g. forward punch, side-boxing, forward kick, side kick, tennis swing, tennis serve, golf swing, etc.* Later the body joint data was added to the dataset. Joint information includes the 3D locations of 20 different body joints in each frame. A decent number of methods are evaluated on this benchmark and recent ones reported close to saturation accuracies [22, 20, 32].

CAD-60 [34] and CAD-120 [18] contain RGB, depth, and skeleton data of human actions. The special character-

istic of these datasets is the variety of camera views. Unlike most of the other datasets, camera is not bound to front-view or side-views. However, the limited number of video samples (60 and 120) is the downside of them.

RGBD-HuDaAct [23] was one of the largest datasets. It contains RGB and depth sequences of 1189 videos of 12 human daily actions (plus one background class), with high variation in time lengths. The special characteristic of this dataset was the synced and aligned RGB and depth channels which enabled local multimodal analysis of RBGD signals[1].

MSR-DailyActivity [38] was among the most challenging benchmarks in this field. It contains 320 samples of 16 daily activities with higher intra-class variation. Small number of samples and the fixed viewpoint of the camera are the limitations of this dataset. Recently reported results on this dataset also achieved very high accuracies [20, 15, 22, 31].

3D Action Pairs [25] was proposed to provide multiple pairs of action classes. Each pair contains very closely related actions with differences along temporal axis *e.g. pick up/put down a box, push/pull a chair, wear/take off a hat, etc.* State-of-the-art methods [17, 32, 31] achieved perfect accuracy on this benchmark.

Multiview 3D event [43] and Northwestern-UCLA [40] datasets used more than one Kincect cameras at the same time to collect multi-view representations of the same action, and scale up the number of samples.

It is worth mentioning, there are more than 40 datasets specifically for 3D human action recognition [47]. Although each of them provided important challenges of human activity analysis, they have limitations in some aspects. Table 1 shows the comparison between some of the current datasets with our large-scale RGB+D action recognition dataset.

To summarize the advantages of our dataset over the existing ones, NTU RGB+D has: 1- many more action classes, 2- many more samples for each action class, 3- much more intra-class variations (poses, environmental conditions, interacted objects, age of actors, ...), 4- more camera views, 5- more camera-to-subject distances, and 6- used Kinect v.2 which provides more accurate depth-maps and 3D joints, especially in a multi-camera setup compared to the previous version of Kinect.

## 2.2. 3D action recognition methods

After the introduction of first few benchmarks, a decent number of methods were proposed and evaluated on them.

Oreifej *et al*. [25] calculated the four-dimensional normals (X-Y-depth-time) from depth sequences and accumulates them on spatio-temporal cubes as quantized histograms over 120 vertices of a regular polychoron. The work of [26] proposed histograms of oriented principle components of depth cloud points, in order to extract robust features against viewpoint variations. Lu *et al*. [20] applied $\tau$ test based binary range-sample features on depth maps and achieved robust representation against noise, scaling, camera views, and background clutter. Yang and Tian [44] proposed supernormal vectors as aggregated dictionary-based codewords of four-dimensional normals over space-time grids.

To have a view-invariant representation of the actions, features can be extracted from the 3D body joint positions which are available for each frame. Evangelidis *et al*. [10] divided the body into part-based joint quadruples and encodes the configuration of each part with a succinct 6D feature vector, so called skeletal quads. To aggregate the skeletal quads, they applied Fisher vectors and classified the samples by a linear SVM. In [37] different skeleton configurations were represented as points on a Lie group. Actions as time-series of skeletal configurations, were encoded as curves on this manifold. The work of [22] utilized group sparsity based class-specific dictionary coding with geometric constraints to extract skeleton-based features. Rahmani and Mian [29] introduced a nonlinear knowledge transfer model to transform different views of human actions to a canonical view. To apply ConvNet-based learning to this domain, [30] used synthetically generated data and fitted them to real mocap data. Their learning method was able to recognize actions from novel poses and viewpoints.

In most of 3D action recognition scenarios, there are more than one modality of information and combining them helps to improve the classification accuracy. Ohn-Bar and Trivedi [24] combined second order joint-angle similarity representations of skeletons with a modified two step HOG feature on spatio-temporal depth maps to build global representation of each video sample and utilized a linear SVM to classify the actions. Wang *et al*. [39], combined Fourier temporal pyramids of skeletal information with local occupancy pattern features extracted from depth maps and applied a data mining framework to discover the most discriminative combinations of body joints. A structured sparsity based multimodal feature fusion technique was introduced by [33] for action recognition in RGB+D domain. In [27] random decision forests were utilized for learning and feature pruning over a combination of depth and skeleton-based features. The work of [32] proposed hierarchical mixed norms to fuse different features and select most informative body parts in a joint learning framework. Hu *et al*. [15] proposed dynamic skeletons as Fourier temporal pyramids of spline-based interpolated skeleton points and their gradients, and HOG-based dynamic color and depth patterns to be used in a RGB+D joint-learning model for action classification.

[1] We emphasize the difference between RGBD and RGB+D terms. We suggest to use RGBD when the two modalities are aligned pixel-wise, and RGB+D when the resolutions of the two are different and frames are not aligned.

我们强调RGBD和RGB + D术语之间的区别。 当两个模态按像素对齐时，我们建议使用RGBD;当两个模式的分辨率不同且帧不对齐时，我们建议使用RGB + D.

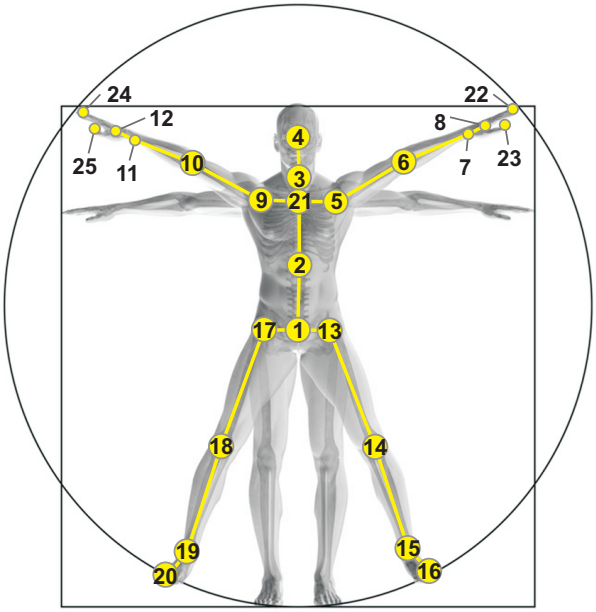


Figure 1. Configuration of 25 body joints in our dataset. The labels of the joints are: 1-base of the spine 2-middle of the spine 3-neck 4-head 5-left shoulder 6-left elbow 7-left wrist 8-left hand 9-right shoulder 10-right elbow 11-right wrist 12-right hand 13-left hip 14-left knee 15-left ankle 16-left foot 17-right hip 18-right knee 19-right ankle 20-right foot 21-spine 22-tip of the left hand 23-left thumb 24-tip of the right hand 25-right thumb

**RNN based 3D action recognition**: The applications of recurrent neural networks for 3D human action recognition were explored very recently [36, 9, 49].

Differential RNN [36] added a new gating mechanism to the traditional LSTM to extract the derivatives of internal state (DoS). The derived DoS was fed to the LSTM gates to learn salient dynamic patterns in 3D skeleton data.

HBRNN-L [9] proposed a multilayer RNN framework for action recognition on a hierarchy of skeleton-based inputs. At the first layer, each subnetwork received the inputs from one body part. On next layers, the combined hidden representation of previous layers were fed as inputs in a hierarchical combination of body parts.

The work of [49] introduced an internal dropout mechanism applied to LSTM gates for stronger regularization in the RNN-based 3D action learning network. To further regularize the learning, a co-occurrence inducing norm was added to the network's cost function which enforced the learning to discover the groups of co-occurring and discriminative joints for better action recognition.

Different from these, our Part-aware LSTM (section 4) is a new RNN-based learning framework which has internal part-based memory sub-cells with a novel gating mechanism.

## 3. The Dataset

This section introduces the details and the evaluation criteria of NTU RGB+D action recognition dataset.[2]

### 3.1. The RGB+D Action Dataset

**Data Modalities:** To collect this dataset, we utilized Microsoft Kinect v2 sensors. We collected four major data modalities provided by this sensor: depth maps, 3D joint information, RGB frames, and IR sequences.

Depth maps are sequences of two dimensional depth values in millimeters. To maintain all the information, we applied lossless compression for each individual frame. The resolution of each depth frame is $512 \times 424$.

Joint information consists of 3-dimensional locations of 25 major body joints for detected and tracked human bodies in the scene. The corresponding pixels on RGB frames and depth maps are also provided for each joint and every frame. The configuration of body joints is illustrated in Figure 1.

RGB videos are recorded in the provided resolution of $1920 \times 1080$.

Infrared sequences are also collected and stored frame by frame in $512 \times 424$.

**Action Classes:** We have 60 action classes in total, which are divided into three major groups: 40 daily actions (drinking, eating, reading, etc.), 9 health-related actions (sneezing, staggering, falling down, etc.), and 11 mutual actions (punching, kicking, hugging, etc.).

**Subjects:** We invited 40 distinct subjects for our data collection. The ages of the subjects are between 10 and 35. Figure 4 shows the variety of the subjects in age, gender, and height. Each subject is assigned a consistent ID number over the entire dataset.

**Views:** We used three cameras at the same time to capture three different horizontal views from the same action. For each setup, the three cameras were located at the same height but from three different horizontal angles: $-45°, 0°, +45°$. Each subject was asked to perform each action twice, once towards the left camera and once towards the right camera. In this way, we capture two front views, one left side view, one right side view, one left side 45 degrees view, and one right side 45 degrees view. The three cameras are assigned consistent camera numbers. Camera 1 always observes the 45 degrees views, while camera 2 and 3 observe front and side views.

To further increase the camera views, on each setup we changed the height and distances of the cameras to the subjects, as reported in Table 2. All the camera and setup numbers are provided for each video sample.

[2] http://rose1.ntu.edu.sg/datasets/actionrecognition.asp

| Setup No. | Height (m) | Distance (m) | Setup No. | Height (m) | Distance (m) |
|---|---|---|---|---|---|
| 1 | 1.7 | 3.5 | 2 | 1.7 | 2.5 |
| 3 | 1.4 | 2.5 | 4 | 1.2 | 3.0 |
| 5 | 1.2 | 3.0 | 6 | 0.8 | 3.5 |
| 7 | 0.5 | 4.5 | 8 | 1.4 | 3.5 |
| 9 | 0.8 | 2.0 | 10 | 1.8 | 3.0 |
| 11 | 1.9 | 3.0 | 12 | 2.0 | 3.0 |
| 13 | 2.1 | 3.0 | 14 | 2.2 | 3.0 |
| 15 | 2.3 | 3.5 | 16 | 2.7 | 3.5 |
| 17 | 2.5 | 3.0 | | | |

Table 2. Height and distance of the three cameras for each collection setup. All height and distance values are in meters.

### 3.2. Benchmark Evaluations

To have standard evaluations for all the reported results on this benchmark, we define precise criteria for two types of action classification evaluation, as described in this section. For each of these two, we report the classification accuracy in percentage.

#### 3.2.1 Cross-Subject Evaluation

In cross-subject evaluation, we split the 40 subjects into training and testing groups. Each group consists of 20 subjects. For this evaluation, the training and testing sets have $40,320$ and $16,560$ samples, respectively. The IDs of training subjects in this evaluation are: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38; remaining subjects are reserved for testing.

#### 3.2.2 Cross-View Evaluation

For cross-view evaluation, we pick all the samples of camera 1 for testing and samples of cameras 2 and 3 for training. In other words, the training set consists of front and two side views of the actions, while testing set includes left and right 45 degree views of the action performances. For this evaluation, the training and testing sets have $37,920$ and $18,960$ samples, respectively.

## 4. Part-Aware LSTM Network

In this section, we introduce a new data-driven learning method to model the human actions using our collected 3D action sequences.

Human actions can be interpreted as time series of body configurations. These body configurations can be effectively and succinctly represented by the 3D locations of major joints of the body. In this fashion, each video sample can be modeled as a sequential representation of configurations.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) [14] have been shown to be among the most successful deep learning models to encode and learn sequential data in various applications [35, 8, 2, 16].

In this section, we introduce the traditional recurrent neural networks and then propose our part-aware LSTM model.

### 4.1. Traditional RNN and LSTM

A recurrent neural network transforms an input sequence $(\mathbf{X})$ to another sequence $(\mathbf{Y})$ by updating its internal state representation $(\mathbf{h}_t)$ at each time step $(t)$ as a linear function of the last step's state and the input at the current step, followed by a nonlinear scaling function. Mathematically:

$$\mathbf{h}_t = \sigma\left(\mathbf{W}\begin{pmatrix}\mathbf{x}_t\\ \mathbf{h}_{t-1}\end{pmatrix}\right) \quad (1)$$

$$\mathbf{y}_t = \sigma\left(\mathbf{V}\mathbf{h}_t\right) \quad (2)$$

where $t \in \{1, .., T\}$ represents time steps, and $\sigma \in \{Sigm, Tanh\}$ is a nonlinear scaling function.

Layers of RNNs can be stacked to build a deep recurrent network:

$$\mathbf{h}_t^l = \sigma\left(\mathbf{W}^l\begin{pmatrix}\mathbf{h}_t^{l-1}\\ \mathbf{h}_{t-1}^l\end{pmatrix}\right) \quad (3)$$

$$\mathbf{h}_t^0 := \mathbf{x}_t \quad (4)$$

$$\mathbf{y}_t = \sigma\left(\mathbf{V}\mathbf{h}_t^L\right) \quad (5)$$

where $l \in \{1, ..., L\}$ represents layers.

Traditional RNNs have limited abilities to keep long-term representation of the sequences and were unable to discover relations among long-ranges of inputs. To alleviate this drawback, Long Short-Term Memory Network [14] was introduced to keep a long term memory inside each RNN unit and learn when to remember or forget information stored inside its internal memory cell $(c^t)$:

$$\begin{pmatrix}i\\ f\\ o\\ g\end{pmatrix} = \begin{pmatrix}Sigm\\ Sigm\\ Sigm\\ Tanh\end{pmatrix}\left(\mathbf{W}\begin{pmatrix}\mathbf{x}_t\\ \mathbf{h}_{t-1}\end{pmatrix}\right) \quad (6)$$

$$c_t = f \odot c_{t-1} + i \odot g \quad (7)$$

$$h_t = o \odot Tanh(c_t) \quad (8)$$

In this model, $i, f, o$, and $g$ denote input gate, forget gate, output gate, and input modulation gate respectively. Operator $\odot$ denotes element-wise multiplication. Figure 2 shows the schema of this recurrent unit.

The output $\mathbf{y}_t$ is fed to a softmax layer to transform the output codes to probability values of class labels. To train such networks for action recognition, we fix the training output label for each input sample over time.

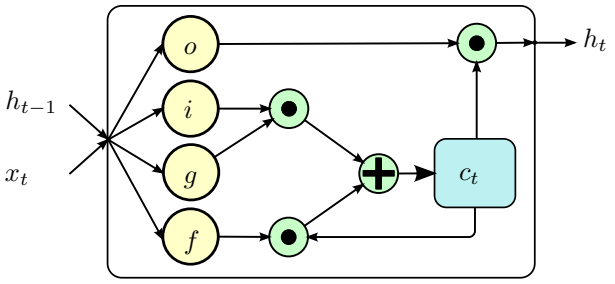


Figure 2. Schema of a long short-term memory (LSTM) unit. $o$ is the output gate, $i$ is the input gate, $g$ is the input modulation gate, and $f$ is the forget gate. $c$ is the memory cell to keep the long term context.

### 4.2. Proposed Part-Aware LSTM

In human actions, body joints move together in groups. Each group can be assigned to a major part of the body, and actions can be interpreted based on the interactions between body parts or with other objects. Based on this intuition, we propose a part-aware LSTM human action learning model. We dub the method P-LSTM.

Instead of keeping a long-term memory of the entire body's motion in the cell, we split it to part-based cells. It is intuitive and more efficient to keep the context of each body part independently and represent the output of the P-LSTM unit as a combination of independent body part context information. In this fashion, each part's cell has its individual input, forget, and modulation gates, but the output gate will be shared among the body parts. In our model, we group the body joints into five part groups: torso, two hands, and two legs.

At each frame $t$, we concatenate the 3D coordinates of the joints inside each part $p \in \{1, ..., P\}$ and consider them as the input representation of that part, denoted as $\mathbf{x}_t^p$.

Thusly, the proposed P-LSTM is modeled as:

$$\begin{pmatrix} i^p \\ f^p \\ g^p \end{pmatrix} = \begin{pmatrix} Sigm \\ Sigm \\ Tanh \end{pmatrix} \left( \mathbf{W}^p \begin{pmatrix} \mathbf{x}_t^p \\ \mathbf{h}_{t-1} \end{pmatrix} \right) \quad (9)$$

$$c_t^p = f^p \odot c_{t-1}^p + i^p \odot g^p \quad (10)$$

$$o = Sigm \left( \mathbf{W}_o \begin{pmatrix} \mathbf{x}_t^1 \\ \vdots \\ \mathbf{x}_t^P \\ \mathbf{h}_{t-1} \end{pmatrix} \right) \quad (11)$$

$$h_t = o \odot Tanh \begin{pmatrix} \mathbf{c}_t^1 \\ \vdots \\ \mathbf{c}_t^P \end{pmatrix} \quad (12)$$

A graphical representation of the propsed P-LSTM is illustrated in Figure 3.

The LSTM baseline has full connections between all the memory cells and all the input features via input modulation gate and the memory cell was supposed to represent the long-term dynamics of the entire skeleton over time. This leads to a very large size of training parameters which are prone to overfitting. We propose to regularize this by dropping unnecessary links. We divide the entire body's dynamics (represented in the memory cell) to the dynamics of body parts (part-based cells) and learn the final classifier over their concatenation. Our P-LSTM learns the common temporal patterns of the parts independently and combines them in the global level representation for action recognition.

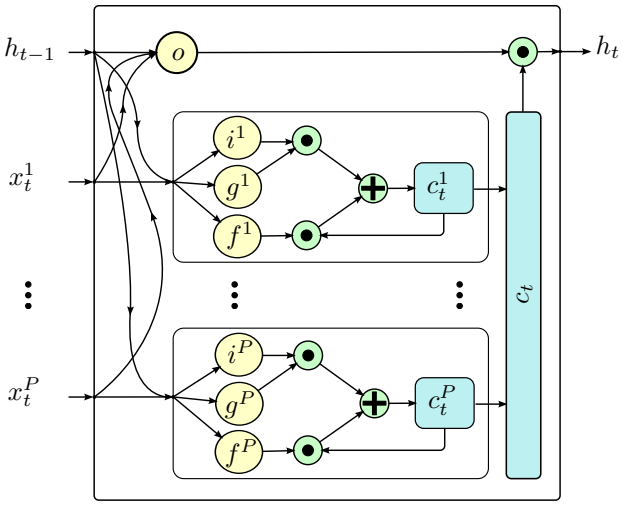


Figure 3. Illustration of the proposed part-aware long short-term memory (P-LSTM) unit.

## 5. Experiments

In our experiments, we evaluate state-of-the-art depth-based action recognition methods and compare them with RNN, LSTM, and the proposed P-LSTM based on the evaluation criteria of our dataset.

### 5.1. Experimental Setup

We use the publicly available implementation of six depth-based action recognition methods and apply them on our new dataset benchmark. Among them, HOG$^2$ [24], Super Normal Vector [44], and HON4D [25] extract features directly from depth maps without using the skeletal information. Lie group [37], Skeletal Quads [10], and FTP Dynamic Skeletons [15] are skeleton-based methods.

The other evaluated methods are RNN, LSTM, and the proposed P-LSTM method.

For skeletal representation, we apply a normalization preprocessing step. The original 3D locations of the body joints are provided in camera coordinate system. We translate them to the body coordinate system with its origin on the "middle of the spine" joint (number 2 in Figure 1), fol-

lowed by a 3D rotation to fix the $X$ axis parallel to the 3D vector from "right shoulder" to "left shoulder", and $Y$ axis towards the 3D vector from "spine base" to "spine". The $Z$ axis is fixed as the new $X \times Y$. In the last step of normalization, we scale all the 3D points based on the distance between "spine base" and "spine" joints.

In the cases of having more than one body in the scene, we transform all of them with regard to the main actor's skeleton. To choose the main actor among the available skeletons, we pick the one with the highest amount of 3D body motion.

Kinect's body tracker is prone to detecting some objects *e.g.* seats or tables as bodies. To filter out these noisy detections, for each tracked skeleton we calculate the spread of the joint locations towards image axis and filtered out the ones whose $X$ spread were more than $0.8$ of their $Y$ spread.

For our recurrent model evaluation, we reserve about five percent of the training data as validation set. The networks are trained on a large number of iterations and we pick the network with the least validation error among all the iterations and report its performance on testing data.

For each video sample at each training iteration, we split the video to $T = 8$ equal sized temporal segments and randomly pick one frame from each segment to feed the skeletal information of that frame as input to the recurrent leaning models in $t \in \{1, ..., T\}$ time steps.

For the baseline methods which use SVM as their classifier, to be able to manage the large scale of the data, we use Libliner SVM toolbox [11].

Our RNN, LSTM, and P-LSTM implementations are done on the Torch toolbox platform [7]. We use a Nvidia Tesla K40 GPU to run our experiments.

### 5.2. Experimental Evaluations

The results of our evaluations of the above-mentioned methods are reported in Table 3. First three rows show the accuracies of the evaluated depth-map features. They perform better in cross-subject evaluation compared to the cross-view one. The reason for this difference is that in the cross-view scenario, the depth appearance of the actions are different and these methods are more prone to learning the appearances or view-dependent motion patterns.

Skeletal-based features (Lie group [37], Skeletal Quads [10], and FTP Dynamic Skeletons [15]), perform better with a notable gap on both settings. They are stronger to generalize between the views because the 3D skeletal representation is view-invariant in essence, but it's prone to errors of the body tracker.

As the most relevant baseline, we implemented HBRNN-L [9] which achieved competitive results to the best hand-crafted methods. Although [9] reported the ineffectiveness of dropout on their experiments, we found it effective on all of our evaluations (including their method).

| Method | Cross Subject Accuracy | Cross View Accuracy |
|---|---|---|
| $HOG^2$ [24] | 32.24% | 22.27% |
| Super Normal Vector [44] | 31.82% | 13.61% |
| HON4D [25] | 30.56% | 7.26% |
| Lie Group [37] | 50.08% | 52.76% |
| Skeletal Quads [10] | 38.62% | 41.36% |
| FTP Dynamic Skeletons [15] | 60.23% | 65.22% |
| HBRNN-L [9] | 59.07% | 63.97% |
| 1 Layer RNN | 56.02% | 60.24% |
| 2 Layer RNN | 56.29% | 64.09% |
| 1 Layer LSTM | 59.14% | 66.81% |
| 2 Layer LSTM | 60.69% | 67.29% |
| 1 Layer P-LSTM | 62.05% | 69.40% |
| **2 Layer P-LSTM** | **62.93%** | **70.27%** |

Table 3. The results of the two evaluation settings of our benchmark using different methods. First three rows are depth-map based baseline methods. Rows 4, 5, and 6 are three skeleton-based baseline methods. Following rows report the performance of RNN, LSTM and the proposed P-LSTM model. Our P-LSTM learning model outperforms other methods on both of the evaluation settings.

This shows they have their model was prone to overfitting due to the lack of training data and proves the demand for a bigger dataset and approves our motivation for proposing NTU RGB+D dataset.

At the next step, we evaluate the discussed recurrent networks on this benchmark. Although RNN has the limitation in discovering long-term interdependency of inputs, they perform competitively with the hand-crafted methods. Stacking one more RNN layer improves the overall performance of the network, especially in cross-view scenario.

By utilizing long-term context in LSTM, the performances are improved significantly. LSTM's performance improves slightly by stacking one more layer.

At the last step, we evaluate the proposed P-LSTM model. By isolating the context memory of each body part and training the classifier based on their combination, we model a new way of regularization in the learning process of LSTM parameters. It utilizes the high intra-part and low inter-part correlation of input features to improve the learning process of the LSTM network. As shown in Table 3 P-LSTM outperforms all other methods by achieving 62.93% in cross-subject, and 70.27% in cross-view evaluations.

## 6. Conclusion

A large-scale RGB+D action recognition dataset is introduced in this paper. Our dataset includes 56880 video samples collected from 60 action classes in highly variant

Figure 4. Sample frames of the NTU RGB+D dataset. First four rows show the variety in human subjects and camera views. Fifth row depicts the intra-class variation of the performances. The last row illustrates RGB, RGB+joints, depth, depth+joints, and IR modalities of a sample frame.

camera settings. Compared to the current datasets for this task, our dataset is larger in orders and contains much more variety in different aspects.

The large scale of the collected data enables us to apply data-driven learning methods like Long Short-Term Memory networks in this problem and achieve better performance accuracies compared to hand-crafted features.

We also propose a Part-aware LSTM model to utilize the physical structure of the human body to further improve the performance of the LSTM learning framework.

The provided experimental results show the availability of large-scale data enables the data-driven learning frameworks to outperform hand-crafted features. They also show the effectiveness of the proposed P-LSTM model over traditional recurrent models.

## 7. Acknowledgement

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

The research is in part supported by Singapore Ministry of Education (MOE) Tier 2 ARC28/14, and Singapore A*STAR Science and Engineering Research Council PSF1321202099.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

## References


[1] J. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *PR Letters*, 2014.

[2] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *CVPR*, 2015.

[3] Z. Cai, J. Han, L. Liu, and L. Shao. Rgb-d datasets using microsoft kinect or similar sensors: a survey. *Multimedia Tools and Applications*, 2016.

[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*, Sept 2015.

[5] L. Chen, H. Wei, and J. Ferryman. A survey of human motion analysis using depth imagery. *PR Letters*, 2013.

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

[7] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[8] 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.

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

[10] G. Evangelidis, G. Singh, and R. Horaud. Skeletal quads: Human action recognition using joint quadruples. In *ICPR*, 2014.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JMLR*, 2008.

[12] F. Han, B. Reily, W. Hoff, and H. Zhang. Space-Time Representation of People Based on 3D Skeletal Data: A Review. *arXiv*, 2016.

[13] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 2013.

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[15] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*, 2015.

[16] A. Karpathy, J. Johnson, and F. Li. Visualizing and understanding recurrent networks. *arXiv*, 2015.

[17] Y. Kong and Y. Fu. Bilinear heterogeneous information machine for rgb-d action recognition. In *CVPR*, 2015.

[18] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *IJRR*, 2013.

[19] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *CVPR Workshops*, 2010.

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

[21] R. Lun and W. Zhao. A survey of applications and human motion recognition with microsoft kinect. *IJPRAI*, 2015.

[22] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *ICCV*, 2013.

[23] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *ICCV Workshops*, 2011.

[24] E. Ohn-Bar and M. Trivedi. Joint angles similarities and $\text{hog}^2$ for action recognition. In *CVPR Workshops*, 2013.

[25] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013.

[26] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. *TPAMI*, 2016.

[27] 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*, 2014.

[28] 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.

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

[30] H. Rahmani and A. Mian. 3d action recognition from novel viewpoints. In *CVPR*, June 2016.

[31] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos. *arXiv*, 2016.

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

[33] A. Shahroudy, G. Wang, and T.-T. Ng. Multi-modal feature fusion for action recognition in rgb-d sequences. In *ISCCSP*, 2014.

[34] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. In *AAAI Workshops*, 2011.

[35] I. Sutskever, O. Vinyals, and Q. V. V. Le. Sequence to sequence learning with neural networks. In *NIPS*. 2014.

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

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

[38] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[39] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *TPAMI*, 2014.

[40] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning, and recognition. In *CVPR*, 2014.

[41] K. Wang, X. Wang, L. Lin, M. Wang, and W. Zuo. 3d human activity recognition with reconfigurable convolutional neural networks. In *ACM MM*, 2014.

[42] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona. Action recognition from depth maps using deep convolutional neural networks. In *THMS*, 2015.

[43] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. Modeling 4d human-object interactions for event and object recognition. In *ICCV*, 2013.

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

[45] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. A survey on human motion analysis from depth data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*. 2013.

[46] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *ACCV*, 2014.

[47] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang. Rgb-d-based action recognition datasets: A survey. *arXiv*, 2016.

[48] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE MultiMedia*, 2012.

[49] 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. *AAAI*, 2016.