**ISBN** 978-981-14-4787-7
**Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)**
*Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 79-83*
doi: 10.18178/wcse.2020.02.014

# Effective Multi-View for Human Activity Recognition on Skeletal Model

Sandar Win [1], Thin Lai Lai Thein[2+]

[1 2]University of Computer Studies, Yangon, Myanmar

**Abstract.** The recognition of 3D human pose from 2D joint location is fundamental to numerous vision issues in analysis of video sequences. Various methods using with skeletal model have been described in past decades, but there is required a powerful system with stable and reliable manner in activity recognition because video sequences can contain different people that may be any position or scale and complex spatial interference. With the development of deep learning, skeleton-based human representation is more reliable to motion speed and appearance of human body scale. Skeleton data contains compact information of the major body joints and that support multi-view to human activity recognition. To satisfy our aim, the proposed system is developed by using OpenPose detector that achieve effective results for 2D pose and Deep Learning based approach. Our goal is to extract valuable information between human joints and to recognize correct activity from human representation in video sequences.

**Keywords:** OpenPose, Human Activity Recognition, Deep Learning

## 1. Introduction

The recognition of perception visual data in human activity is generally analysed into five categories that are body part locations or skeletal joints, 3D silhouettes, local space-time information, features of scene flow, and local residence features [1]. Skeleton-based human activity recognition has affected in a great deal of interest to a lot of researchers and available in real-world applications such as intelligent surveillance, activity recognition, sport video analytics, and autonomous monitoring system, etc. Currently, several approaches were developed by using 3D skeletal joint positions that is directly taken from sensors. Many of researchers used depth images to recognize human activity in their system. Depth image provides geometric feature of pixel information that supports the scene in 3D space and estimate 3D pose based on depth information, but computing of depth maps is slow and prone to errors when searching the correspondence map in noisy depth information [2]. On the other hand, by using depth images and reconstructing of 3D point clouds are robust to scale, rotation and illumination changes [3]. This approach is very constraint in application area. Another approach is automatic acquisition of accurate 3D pose from an image requires a very sophisticated setup. And, some approach calculates relative joint orientations and utilizes order of joint to connect adjacent vectors [4], but occurs ambiguity in recognition system. So, effective multi-view for activity recognition from video sequence have been required. To solve these problems, the system integrates OpenPose with Joint Correlation Distance and skeleton visualization method to estimate human pose for activity recognition. The purpose of our system is to propose skeleton based robust method on the Deep Learning Unified framework. The remained portions of the paper are organized as follows: Related works are described in Section 2. Methodology of proposed approach is presented in Section 3. Experimental results are shown in Section 4, and conclusion and future works are expressed in Section 5.

## 2. Related Works

[+] Corresponding author. Tel.: + 95-09-793757403; fax: + 95-013-610-633.
*E-mail address*: sandarwin@ucsy.ed.mm

The skeleton data have been widely utilized for activity recognition system because it can provide dynamic conditions and complex circumstances. Since human joint information in skeleton data have been proved great success for action recognition tasks. Wu and Shao [5] proposed to recognize human action by using deep neural networks with hierarchical dynamic framework on extracted 3D skeleton feature and estimate the probability of action sequences. Luvizon et al. [6] proposed a multitask framework for 2D joint and 3D pose estimation from video sequences in recognition of human action. They trained multi type of dataset to generate 3D predictions from 2D annotated data and proved an efficient way for action recognition based on skeleton information. Iqbal et al. [7] developed dual source network on 3D human pose estimation. They collected large amount of unconstrained data in 2D and 3D pose. And by taking nearest 3D pose and reconstructed the 3D pose for single image estimation. Du et al. [8] proposed Recurrent Pose Attention Network (RPAN) that predicts the related features in human pose. The system used end to end recurrent network layers for temporal action modelling to construct a skeleton-based human representation. As the number of layer increase, the representations extracted by the subnets are hierarchically fused to figure a high-level feature to represent human in 3D space. Yang et al. [9] proposed Double-feature Double-motion Network (DD-Net) to achieve lightweight network structure by employing skeleton sequence attributes and motion scale variation. Li et al. [10] described translation scale invariant method that work well on 2D skeleton video. They used Convolutional Neural Network (CNN) architecture and result is compared on benchmark dataset. In activity recognition system, most of system used depth image. Although these are very constraint in outdoor applications. To eliminate this constraint, the proposed system is developed with efficient multi-view for human activity recognition from video sequences which contains forward, frontal, lateral and backward motion.

# 3. Methodology

The proposed system consists of three parts. There are body pose detection from video sequences and extract 2D joint locations, 3D pose estimation and Activity Recognition on Deep Neural Network. An overview of proposed system is shown in Fig.1.

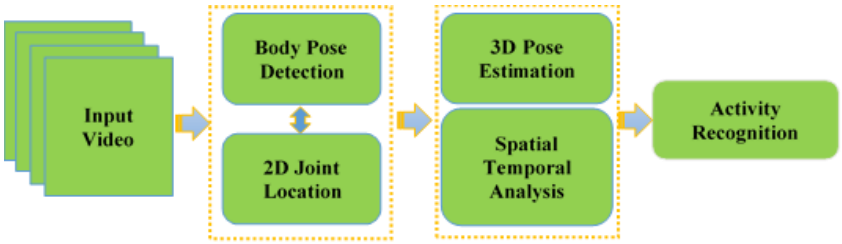


Fig. 1: An overview of proposed system. Input video is first mapped with OpenPose detector via activity recognition on Deep Neural Network

## 3.1. Body Pose Detection and Extract 2D Joint Locations

The system detects human body by OpenPose detector that search key points of human body parts and extract 2D joint locations. To perform better result, a non-maximum suppression method is used to select the interest points in the representation structure and that returns the top scoring result as defined the estimated pose. This approach also supports partial occlusions and improves motion detection [11].

### 3.1.1. Scale Invariant Motion Modeling

For skeleton-based action recognition, geometric feature and Cartesian coordinate feature are used as an input feature. The geometric feature such as joint indices and distances are location viewpoint invariant, but not stable from one data to another. The indices of joint (i.e., the IDs of the head, left and right shoulders, etc.) could be dynamically changed in different actions. Hence, the difficulty arises and requires to predefine the correlation of joints by ordering of their indices. By using Joint Collection Distances (JCD) can solve for problems. As another benefit, the embedding process can reduce the effect of skeleton noise. Joint collection as defined by $S_k = \{J_1^k, J_2^k, \ldots, J_N^k\}$, where $k$ is number of frame and $JCD$ feature of $S_k$is computed:

$$JCD^k = \begin{bmatrix} \left\|\overrightarrow{J_2^k J_1^k}\right\| & \cdots & \left\|\overrightarrow{J_2^k J_{N-1}^k}\right\| \\ \vdots & \ddots & \vdots \\ \left\|\overrightarrow{J_N^k J_1^k}\right\| & \cdots & \left\|\overrightarrow{J_N^k J_{N-1}^k}\right\| \end{bmatrix} \quad (1)$$

where $\left\|J_i^k J_j^k\right\| (i \neq j)$ denotes Euclidean distance between $J_i^k and J_j^k$

## 3.2. 3D Pose Estimation

Pose estimation is an important class in action recognition. 3D pose estimation from 2D joint locations with skeleton data has advantage to view point invariant and greatly effects on performance. We also consider changes between 2D and 3D pose structure to reduce ambiguity physical constraint on limb lengths. Deep fully convolutional network is trained to predict the uncertainty maps of the 2D joint locations. Then, 3D pose is estimates via an Expectation-Maximization (EM) algorithm over the entire sequence. EM algorithm search probability distribution of 2D joint location (heat map value) in frame t and compute mean and normalized nearest 3D poses, the steps continue iteratively until convergence. Then final 3D pose is retrieved by minimizing the projection error in solution. The relative positions of the limbs and generate pose model that can be used to control 3D motion model. Finally, activity recognition is deeper and more reliable approach with deep neural network alongside human poses for understanding of human activity.

### 3.2.1. Spatial-Temporal Information Analysis

Human action recognition remains a problem to efficiently represent spatiotemporal skeleton sequences. To produce efficient sequences for each human body, the main concept is to measure the minimum distance between the detected pose and OpenPose library poses across the frame.

Let $J_a, J_b, J_c, J_d, J_e$ be subset of body parts $J = \{1, \ldots, 17\}$ such as $J_a = \{3,4,5\} \subset J, J_b = \{6,7,8\} \subset J, \ldots, J_e = \{15,16,17\} \subset J$

The distance of generic pose $p_i(t)$ and generic prototypes $v \in V_l$ is obtained by

$$d_{\bar{J}_*}(p_i(t), v) = \frac{1}{|\bar{J}_*|} \sum_{j \in \bar{J}_*} \left\|(x_j, y_j)_i - (x_j^\dagger, y_j^\dagger)\right\|_2 \quad (2)$$

where $(x_j^\dagger, y_j^\dagger)$ are coordinates of $j^{th}$ landmarks of $v$ and $\bar{J}_*$ denote without missing landmark.

$V_l$ be library of prototypes for action $l$ . The embedding sequence is taking as:

$$D_{V_l, J_*}(t) = \min_{v \in V_l} d_{\bar{J}_*}(p_i(t), v) \quad (3)$$

Given a set of action $\mathcal{L}$, the meaningful sequences extracted from $p_i(t)$ is defined as :

$$Seq_i(V_l) = \left\{ D_{V_l, J}(t), D_{V_l, J_a}(t), \ldots, D_{V_l, J_e}(t) \right\} \forall l \in \mathcal{L} \quad (4)$$

## 3.3. Activity Recognition on Deep Neural Network

The system is developed process 2D data to 3D poses for human activity recognition. Human pose is basically represented as a graph where the joints are the nodes and the bones are the edges. Since every joint is connected to another joints in its neighbourhood and it has distribution power. The encoded structure matrix and model representation ability of weight matrix which become the dependent features in graph modelling.

### 3.3.1. Graph Modeling

To get better results with deeper network, VGG net is used in this system. VGG net is a Deep Neural Network architecture for object recognition. VGG net is designed as the simplest with 3x3 convolution and 2x2 max pooling layers are used throughout the whole network. Each layer along with pre-trained set of weights. It performs better on the training set because smaller and smaller features processing on additional layers. The training error actually decreases as the network gets deeper and can gain accuracy from considerably increased depth. By developing The novel adaptive dependency matrix and learn it through node embedding, our model can precisely capture the hidden spatial dependency and can generate a heat-map to encode a per-pixel likelihood for human joint localization. These heat maps are combined with temporal information alongside spatial information. Then, full-body information of 3D skeleton human pose is produced by simultaneously taking joint distribution in fully connected layers of Deep Neural Network.

### 3.3.2. Training and Testing

All through training, by taking the input from each sequence with rectangular patch around it and is resized to $224 \times 224$ RGB pixels. For data normalization, channel-wise RGB mean values are computed and subtracted from the images. The training labels to be regressed are multi-channel heat maps with each channel corresponding to the image location for each joint and is defined by using associated probability. The network consists of $3 \times 3$ convolutional layers with filters followed by ReLU activition function to provide dense prediction for all joints. A $2 \times 2$ max pooling layer is inserted after each of the first three convolutional layers and returns a set of corresponding probability to the class labels as output. The network is trained by minimizing the $l_2$ loss between the prediction and open source Caffe framework.

During testing, consistent with previous 3D pose methods, the subject with bounding box is assumed and the image patch in the bounding box is cropped in frame $t$ and fed forward through the network to predict the probability of the joint occurring at each pixel. Finally, recognize the efficient understanding of human activity.

## 4. Experimental Result

### 4.1. Prepare Implementation

The deep neural network is trained by using Adam optimization algorithm and it takes advantage of momentum with moving part, the pre-trained weight model from Deep Learning framework and the start learning rate of 0.001 using 128 mini batch-size.


Fig. 2: Example result of our test on HMDB51 dataset in outdoor area

### 4.2. Result Analysis and Discussion

To describe the efficient results, we experiment on HMDB51dataset which contains the full complexity of video clips commonly found in YouTube, Google videos, movies and public database. As a result of our test in Fig.2. and the system shows well-defined to unseen activity in human movement and outperform good recognition. The proposed approach obtained high accuracy rates and the confusion matrix with colormap also states that the result is robust to multi-view as expressed in Fig.3. The accuracy result of training and testing as described in Fig.4.

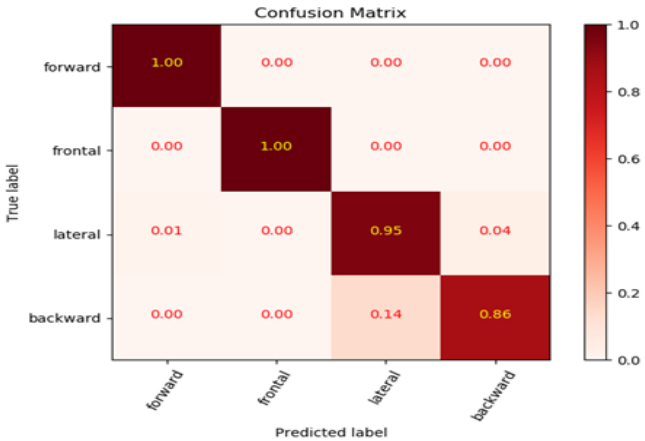


Fig. 3: The result of multi-view human recognition on confusion matrix

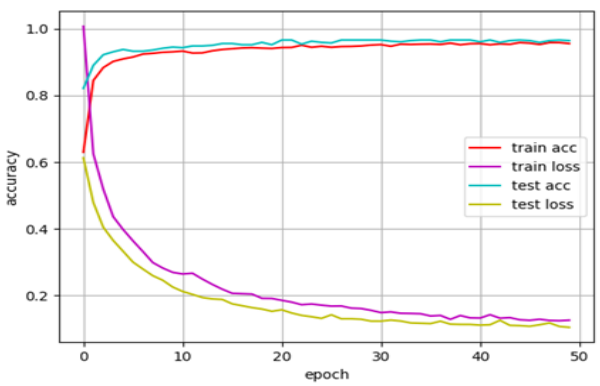


Fig. 4: The accuracy result of human recognition on training and testing

## 5. Conclusion and Future Work

There are various methods have been developed for human recognition by using skeleton sequence. But efficient multi-view human activity recognition systems have been required. That can record useful information and analyse the environment in many scopes. There are many challenges that are concerned with different variation in human pose, shape, illumination changes and background appearance. In this paper, the system is implemented by using deep neural network framework to get high accuracy recognition of human movement in indoor and outdoor areas. The experimental results have been concluded that recognition of moving object have a big dependency with different backgrounds, camera calibration and illumination changes. We trained and tested on HMDB51video dataset with different changes that are significantly increased recognition rate of our results.

Future research directions will continue 3D skeletal model for moving object with various dataset containing different activities and different views to describe the more accurate result of human activity recognition system on various data.

## 6. References


[1] J. Aggarwal, L. Xia, Human activity recognition from 3D data: A review, Pattern Recognition Letters 48 (0) (2014) 70–80.

[2] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with Microsoft Kinect sensor: A review, IEEE Transactions on Cybernetics 43 (5) (2013) 1318–1334.

[3] A. L. Brooks, A. Czarowicz, Markerless motion tracking: MS Kinect & Organic Motion OpenStage R, in: International Conference on Disability, Virtual Reality and Associated Technologies, 2012.

[4] S.Y. Jin, H.J. Choi, Essential body-joint and atomic action detection for human activity recognition using longest common subsequence algorithm, in: Workshops on Asian Conference on Computer Vision, 2013.

[5] D. Wu, L. Shao, Leveraging hierarchical parametric networks for skeletal joints action segmentation and recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[6] D. C. Luvizon et al., 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning, IEEE Conference on Computer Vision Foundation, 2014.

[7] U. Iqbal, A. Doering, H. Yasin, B. Krüger, A. Weber, and J. Gall, A Dual-Source Approach for 3D Human Pose Estimation from a Single Image, 2017.

[8] W. Du, Y. Wang, and Y. Qiao, RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos, in IEEE Int. Conf. on Computer Vision (ICCV), Oct. 2017, pp. 3745–3754.

[9] F. Yang, S. Sakti, Y. Wu, S. Nakamura, Make Skeleton-based Action Recognition Model Smaller, Faster and Better, in arXiv: 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, in CVPR, 2017, pp.7291–7299.