Skeleton based Human Action Recognition using a Structured-Tree Neural Network

Muhammad Sajid Khan, Andrew Ware, Misha Karim, Nisar Bahoo, and Muhammad Junaid Khalid

***Abstract*—The ability for automated technologies to correctly identify a human's actions provides considerable scope for systems that make use of human-machine interaction. Thus, automatic 3D Human Action Recognition is an area that has seen significant research effort. In work described here, a human's everyday 3D actions recorded in the NTU RGB+D dataset are identified using a novel structured-tree neural network. The nodes of the tree represent the skeleton joints, with the spine being represented by the root. The connection between a child node and its parent is known as the incoming edge while the reciprocal connection is known as the outgoing edge. The uses of tree structure lead to a system that intuitively maps to human movements. The classifier uses the change in displacement of joints and change in the angles between incoming and outgoing edges as features for classification of the actions performed.**

***Index Terms*—Structure-Tree Neural Network (STNN), Skeleton, Human Action Recognition (HAR).**

I. Introduction

Human Action Recognition (HAR) is a dynamic and challenging task in which an individual's physical actions are identified. Generally, the HAR process involves several steps starting from harvesting human motion information from raw sensor data, through to correctly identifying the actions performed. MIT researchers were one of the earliest to develop such a technique, using a bottom-up approach to extrapolate 3D models of point clouds enabling computer vision systems capable of automating the capture and processing of images [1] in applications such as surveillance. Nowadays, Point Cloud Systems are the most common HAR technique, on account of their ability to facilitate fast detection and identification of actions recorded using 3D video. A point cloud is a collection of 3D positions (x, y, z) that represents the surface of an object or plane. Point clouds can be used for calculations directly in or on the object, e.g. distances, diameters, curvatures or cubature. While a specific cloud point represents a position on an object, it does not hold any detail regarding its internal features such as its colour or material.

Some HAR frameworks[2] have used Kinect datasets and a deep Convolution Neural Network (CNN) architecture that makes use of the spatial arrangement of pixels to classify objects correctly. However, 3D structure posture can also be determined using point clouds in multi-camera scenarios by employing a skeletal model of human joints [3], where each frame from the video is compared with the stored standard dataset.

To demonstrate the potency of the novel HAR system described here, the UTKinect-Action3D Dataset was used to identify the actions of walking, running, laying, sitting and waving etc. The system was trained using NTU-RGB+D 120[3, 4], which consists of 56,880 video recordings of 3D skeleton-based actions. It includes sixty action classes of both single-person and multi-person activities recorded using RGB videos and 3D skeleton information. The videos of 40 human subjects aged between 10 and 35 were captured using a Microsoft Kinect V2 at 30 fps. Three cameras set at the same height but different horizontal angles (−45◦, 0◦, 45◦) were used to record every action. Each camera recorded the human action from which the movement of the twenty-five primary joints can be determined. The methodology was tested using two benchmarks, Cross-subject (CS) and Cross-view (CV), to validate the decision regarding the use of the three angles.

II. Related Work

HAR systems play a role in a wide range of different applications such as Video Surveillance (used to record activity within an area covered by CCTV cameras), Video Retrieval (used to locate particular activities within a video recording), Healthcare Monitoring (enabling the constant observation of an individual's physiological parameters), Robot Vision (allowing for the automated response to human actions). Though these applications highlight the potential for HAR, there are still significant issues to be addressed with regards to improving response times and increasing the range of actions that can be identified.

A significant driver behind ongoing HAR development is to improve their accuracy, efficiency and response time. One such development has been the use of intelligent environments where a spherical coordinate system is used to calculate the difference between a silhouetted image sequences frame-by-frame [5]. The 2D silhouette sequences supports space and time dimensions only. In any case, the method actualises a cubic transformation from 2D input to 3D by including the spatial measurement. For obtaining the

Published on August 13, 2020.

M. S. Khan, Computer Science Department, Army Public College of Management and Sciences, Khadim Hussain Road, Rawalpindi, Pakistan.
(e-mail: sajidpk48@yahoo.com)
A. Ware, Faculty of Computing, Engineering and Science, University of South Wales, United Kingdom.
(e-mail: andrew.ware@southwales.ac.uk)
M. Karim, Computer Science Department, Army Public College of Management and Sciences, Khadim Hussain Road, Rawalpindi, Pakistan
(e-mail: misha.karim276@gmail.com)
N. Bahoo, Computer Science Department, Army Public College of Management and Sciences, Khadim Hussain Road, Rawalpindi, Pakistan
(e-mail: nisar.bahoo123456@gmail.com)
M. J. Khalid, Computer Science Department, Army Public College of Management and Sciences, Khadim Hussain Road, Rawalpindi, Pakistan.
(e-mail: mjk22071998@gmail.com)

DOI: http://dx.doi.org/10.24018/ejers.2020.5.8.2004

images of the silhouettes, techniques such as background subtraction and Gaussian spatial filter were applied to segment the data, and for differentiating actions over gestures, K-D trees were used.

Analysing the movement of a human's 3D joints can likewise help in identifying the actions being carried out [6]. A Temporal Pyramid covariance descriptor allows features such as relative joint distance to be extracted for representation as a Joint Spatial Graph (JSG). The JSG Kernel is then used to perform edge attribute similarity checking and vertex attribute similarity checking, a process which helps facilitate action identification. Three different datasets were used for training, testing and evaluation: MSR-3D-action dataset; UTKinect-3D-action dataset; and Florence-3D action dataset.

Yang and Tian [7] proposed an effective method to recognise human actions using 3D skeleton joints recovered from 3D depth data of RGBD cameras. They used EigenJoints, as an action feature descriptor for action recognition, which combine joint action information, including static posture, motion property, and overall dynamics. They then used Accumulated Motion Energy (AME) to remove noisy frames and reduce computational cost, before employing non-parametric Naïve-Bayes-Nearest-Neighbour (NBNN) to classify multiple actions. Their experimental results demonstrate that the approach was significantly better than comparable methods.

An autonomous system developed by Munaro et al. [8], uses 3D motion flow for real-time recognition of online human actions. A Microsoft Kinect is used to determine the RGB values of the scene before a grid-based descriptor connects multiple point clouds to form a recognisable shape. A K-nearest neighbor (KNN) algorithm then classifies the storedactions. During testing, 90% of actions were accurately recognised.

Yang et al. [9] developed an advanced depth sensor, suitable for identifying human actions, that makes use of the Microsoft Kinect. The approach makes use of body posture and motion information to determine the action recorded within a video clip. The sensor provides depth maps that are projected onto three orthogonal Cartesian planes facilitating identification of the global activities within the video clip. Following identification of activities, a histogram of oriented gradients is computed from the depth motion map, enabling the recorded actions to be identified.

A method proposed by Wu et al. [9], uses a KNN classifier that makes use of human kinematic similarity in real-time. The method is based on action descriptors using angular velocity, angular acceleration, along with joint positions. The action descriptors first remove noisy frames containing segmented and irregular data by combining all the geometric parameters of the human skeleton to achieve higher accuracy.

Lei Shi et al. [10], proposed a Directed Graph Neural Network (DGNN) which represents a skeleton as a directed acyclic graph. Their algorithm represents an improvement as it extracts information (attributes) between edges and vertices in the form of vectors working in different layers. The approach is iterative, where each layer receives updated attributes from the previous layer. The bottom layer is responsible for the manipulation of nearer/adjacent vertices, while the top layer is responsible for the manipulation of farther vertices. The DGNN extracts information about the bone'sand joint's coordinates; and provides the spatial information to represent their vertexes and edges within a directed acyclic graph (DAG). The direction of the graph can be determined using the root vertex and its distance from the current vertex. Here, the neck joint is mapped to the root joint [10].

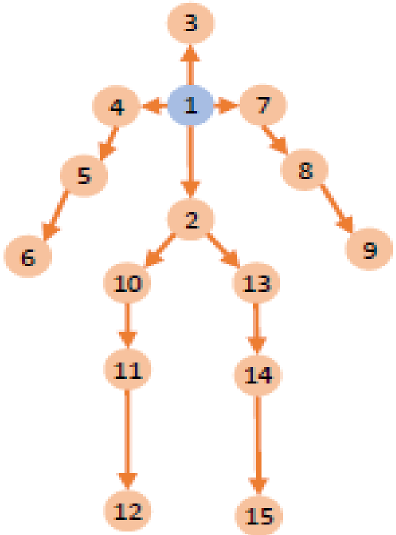

Fig. 1. Directed Acyclic Graph used in Directed Graph Neural Network [10].

III. Proposed Method

A structured-tree neural network is an approach in which the joints of the skeleton are represented as nodes in a tree. The root node represents the spine of the skeleton. Each node contains static information about its parent node, its position in the tree, its incoming and outgoing edges, and the displacement of vertices and edges. This static information is used for obtaining the static spatial information (such as relative position and angle) and caulculating the dynamic temporal information (movements such as change in positions and angle) of each node. The movements of the skeleton are determined using the differences between consecutive static frames. A convolutional layer is applied to each frame to extract its features, which are aggregated using global pooling before using a softmax layer to facilitate classification.

In 2014 [7], researchers developed a technique, using EigenJoints, for recognising single actions recorded as 3D videos. However, it was not reliable in identifying multiple actions within a scene. Research in 2017, making use of angular spatio-temporal descriptors, provided a quick response for a limited number of actions stored in both segmented and unsegmented datasets. Later, in 2019, the Directed Graph Neural Network was proposed, which offered excellent performance and accuracy, but consumed significant storage space. The timeline shows that with the advancement of technology, and discovery of new algorithms, the task of action recognition turned out to be progressively more accurate, achieving a precision of 90%. Yet, at the same time, it was inefficient in terms of storage.

Of the recent approaches, the Directed Graph Neural

DOI: http://dx.doi.org/10.24018/ejers.2020.5.8.2004

Network (DGNN) seems to hold the most potential for the classification of different actions. Even though its complex algorithms affect performance, it achieved more than 95% accuracy. During testing of DGNN three 3D image videos, from the NTU-RGB+D, for each of a set of individuals are presented as input to the system. Each video having been recorded using a separate view with Kinect V2 camera. Various pre-processing techniques split the video into multiple frames before detailed information about the subject depicted is extracted. To ensure a high degree of quality, part of the pre-processing involves noise removal and image sharpening. The noise removal was accomplished using the Median Filter, and image sharpening by the application of a Laplacian filter. The standard Point Cloud Library can then be used to create a skeletal model of the humans in the scene.

Images can contain redundant artefacts which serve to hamper the recognition. Thus, to facilitate the removal of such objects, the skeletal model is copied to a new empty 3D plane, providing a clear visualisation of movements. This enables skeletal features to be extracted using a Body Pose Evolution Map. Once this has been achieved, the Late Fusion and Cross Setup Protocol can be used to combine the features before classification. The features combined in this way are: the movement of nodes (frame-by-frame difference between node positions); and, the deformation of bones (frame-by-frame difference between angles and difference between incoming and outgoing edges). These movement and deformation features are passed through a Batch Normalisation and Rectified Linear Unit. The processed feature data are then fed through Global-average poolingand softmax layers to determine the class-category of a given action.

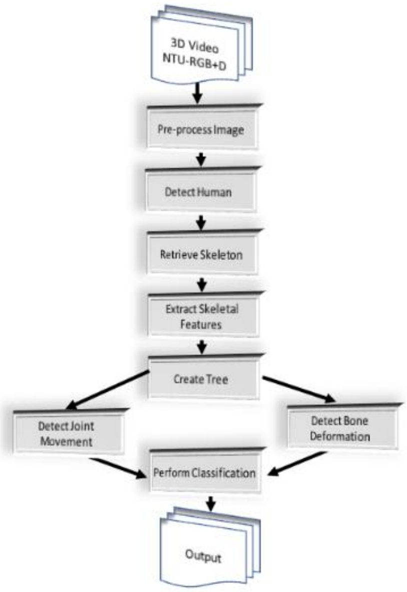

Fig. 2. Block Diagram of the system.

A. Pre-processing

Video clips of human actions from the 3D-NTU RGB+D dataset [4] are used for training. Frames within the clips typically contain a level of noise that must be removed to enhance the clarity of objects depicted and sharpen their edges. In terms of clarification, the median filter is effective at noise removal, and when applied to each frame replaces the grey level of each pixel with the median of the grey levels in its neighbourhood. With regards sharpening, a Laplacian filter is used to enhance the clarity of edges, which helps in feature-extraction and skeleton-shaping.

$$L(x, y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \tag{1}$$

B. Human Body detection

PCL (Point Cloud Library) [11], an open-source and a cross-platform library, enables the location of humans within a scene to be determined.

C. Skeleton Extraction

Points provided by the PCL are used to extract skeletal joint position information from the frame, enabling it to be plotted and visualised. The whole image can be displayed using just 25 skeleton focus points, resulting in improved performance and reduced computational overhead.

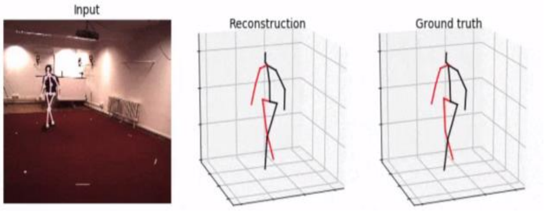

Fig. 3. Extracted skeleton of walk action.

D. Neuron Tree Model

A structured tree is used to represent the 25 main joints of the skeletal structure. Of these 25 joints, four (4, 21, 2 and 1) are the most significant when mapping from the skeletal to the tree structure. Fig. 4 shows the tree formed by applying the Global Long-sequence Attention Network and Sub Short–sequence Attention Network to the skeleton point [12]. These networks form the tree without losing information stored in leaf nodes or spatio-temporal information.

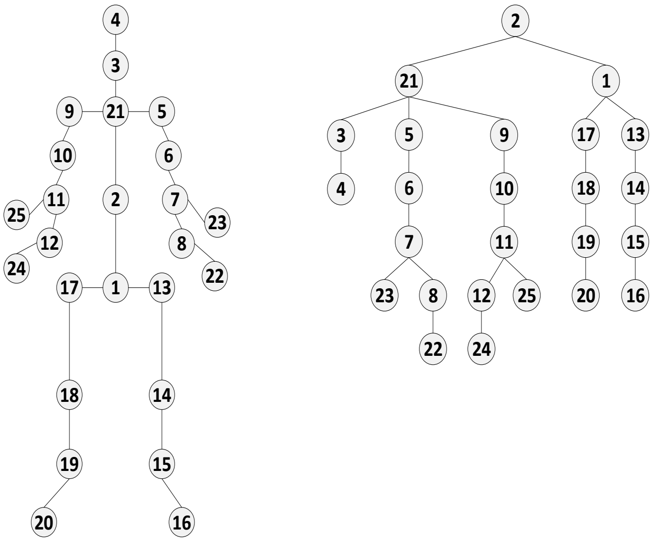

Fig. 4. Skeleton in NTU RGBD dataset (left) tree formed by the skeleton (right)

E. Movements of Joints

The position of each joint is determined for each frame in the video clip. This enables the movement of each joint between frames to be calculated. The equation $\mathbf{M}_{vt} = \mathbf{V}_{t+1} - \mathbf{V}_{t}$is used to calculate the movement of joints between each frame. Where $\mathbf{V}_{t+1}$ is the next position of the joint in the sequence after $\mathbf{V}_{t}$.Knowing this movement is vitalfor identifying the performed action correctly. For example, when waving "hello" to someone, the three main joints moved are hand, elbow and shoulder (joints number 9, 10 and 11 or joints number 5, 6 and 7 from Fig. 4).

F. Edge deformation

The change of angle between incoming and outgoing edges, in consecutive video frames, is referred to as an edge deformation, calculated using $\mathbf{M}_{et} = \mathbf{M}_{e+1} - \mathbf{M}_{e}$. To help clarify, considera person waving "hello", an action that involves the hand, elbow and shoulder. Suppose traversal is at elbow node, so the current node is elbow node. The x y coordinates of elbow nodes are represented as X_{curr} andY_{curr}. Similarly, the x y coordinates of hand and shoulder nodes are represented as X_{child}, Y_{child} and X_{parent}, Y_{parent} respectively (see Fig. 5). The slopes of incoming edge (connecting with parent) and outgoing edge (connection with child) are calculated using the following formulas:

$$Slope_{incoming} = \frac{Y_{parent} - Y_{curr}}{X_{parent} - X_{curr}}$$

$$Slope_{outgoing} = \frac{Y_{curr} - Y_{child}}{X_{curr} - X_{child}}$$

The angle between these slopes is given by:

$$\theta = \tan^{-1} \frac{Slope_{outgoing} - Slope_{incoming}}{1 + Slope_{incoming} Slope_{outgoing}}$$

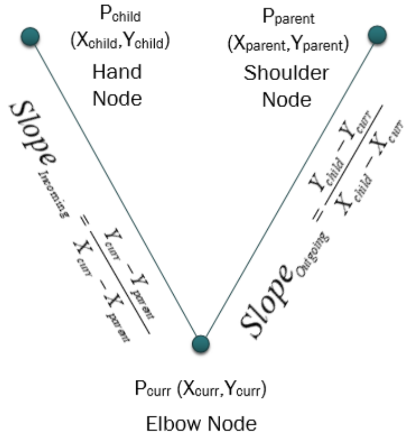
P_{child}
(X_{child}, Y_{child})
Hand
Node
P_{parent}
(X_{parent}, Y_{parent})
Shoulder
Node
$Slope_{Incoming} = \frac{Y_{curr} - Y_{parent}}{X_{curr} - X_{parent}}$
$Slope_{Outgoing} = \frac{Y_{child} - Y_{curr}}{X_{child} - X_{curr}}$
P_{curr} (X_{curr}, Y_{curr})
Elbow Node

Fig. 5. Slope calculation for the angle.

IV. Structure Tree Neural Network

The spatio-temporal information is extracted from the tree using a Structured-Tree Neural Network (STNN). The STNN consists of multiple nodes (or vertices) the number of which is proportional to the number of frames in the video clip. In work reported here, the STNN has four layers; input, output,and upper and lower hidden layers. The input layer consists of 25 nodes, where each represents a skeletal joint. Features extracted by the first hidden layer get forwarded to the second hidden layer, which is responsible for filtering nodes with redundant features. The remaining nodes are then used by the output layer, for determiningthe actionperformed, asdepicted in Fig. 6.

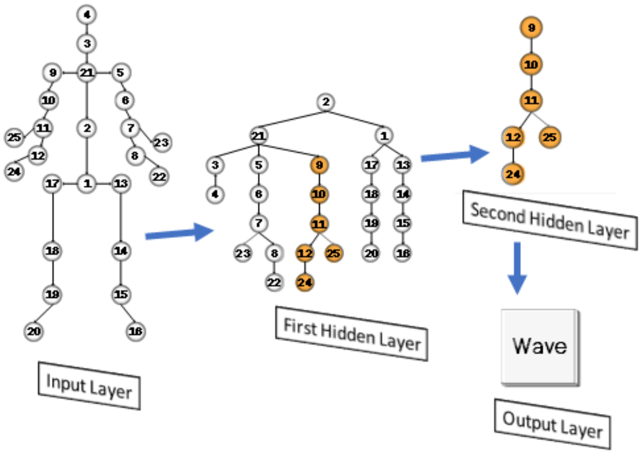

Fig. 6. Depiction of layers in classifier (on wave action)

Note that each node of the tree can manipulate its parent as well as its child node(s). The connection between parent and child provides information about the incoming and outgoing edges. As shown below, functions such as $h^{c}()$ and $h^{p}()$,act upon child and parent nodes respectively.

$$h^{c}(n_i) = n_{i+1}.\text{info} - n_i.\text{info}$$
$$h^{p}(n_i) = n_i.\text{info} - n_{i-1}.\text{info}$$

Info indicates the information above the features of the interest. However, it only provides information regarding adjacent nodes and not the whole tree. The adjacency matrix calculated is now mapped onto an attention map, where:

$$A_O = \text{originalMatrix}$$
$$A = PA_O$$

For stable training,

$$A = A_O + P$$

The input to the STNN consists of both spatial (such as the position of the node) and temporal (such as velocity) information. For extracting the temporal information, a temporal convolutional block containing a 1D convolution layer followed by a global-average pooling layer, which is further normalised through a ReLU layer, is used. Fig. 7 depicts the flow of the process; the output is an image like formation, generated using a softmax layer used for class prediction.

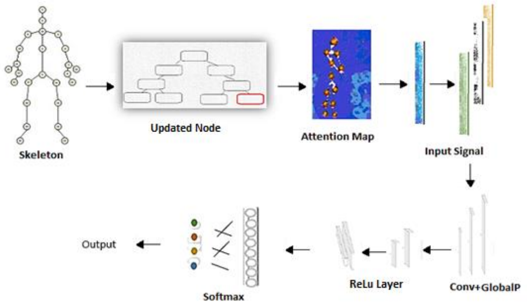

Fig. 7. The flow of the Structured-Tree Neural Network

DOI: http://dx.doi.org/10.24018/ejers.2020.5.8.2004

V. Test and Results

Initially, a skeletal model in the form of a tree was built using the PCL [11] to determine the location of each joint. Every node in the tree holds the relevant data and pointers to its offspring nodes. The data consists of the x, y, z location coordinates and the angle between the incoming and outgoing edges.

To identify the human action being performed, the difference in data values between consecutive frames is used to calculate the new position and angle for all nodes. The updated nodes are then mapped to the attention map. To form the input signal to the feature extraction algorithms, a1D convolution layer, followed by Global Pooling layer, is applied to extract the calculated differences, aAReLU layer removes redundant features and, finally, a softmax layer classifies the action.

The algorithm was executed on a Core i72nd generation, 16GB RAM and Nvidia Quadro 2000M GPU using the same implementation frameworks as for the DGNN. Comparing the performance, the DGNN approach [10] was less accurate than the new one. (See Table I, for a fuller comparison of the methodologies.)

TABLE I: A Comparison of the Methodologies

Algorithm	**Accuracy (CS) Cross Subject**	**Accuracy (CV) Cross view**	**Time Complexity**
DGNN	89.9%	96.1%	O(Vertics+Edges)
STNN (proposed algorithm)	90.2%	96.3%	O(Vertices)

The features used to facilitate action recognition are the Magnitude and Orientation of the skeleton's bones and their corresponding joints. Fig. 8, shows a confusion matrix of the degree of discrepancy between the actual and perceived data.

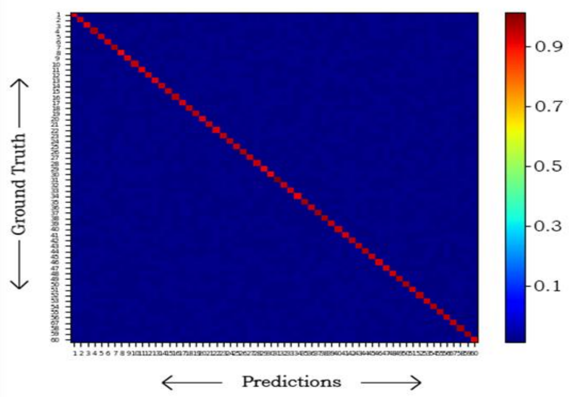

Fig. 8. Confusion Matrix of TSNN on NTU RGB+D 60 dataset axis is showing the number of actions.

VI. Conclusion

The paper presents a novel methodology for HAR that uses a tree structure to represent the human skeleton. The approach is intuitive as the skeletal structure resembles that of a simple inverted tree. The human's spine joint, the central joint of the body, is represented by the root node of the tree, which helps facilitate optimised performance in terms of storage requirement and response time. The results obtained show an improvement over the DGNN approach, which requires significant memory to store the features extracted and which is also slower to execute.

Acknowledgement

Portions of the research in this paper used the NTU RGB+D Action Recognition Dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore.

References

[1] "50 years of object recognition: Directions forward." *Computer vision and image understanding* 117, no. 8 (2013): 827-891. https://doi.org/10.1016/j.cviu.2013.04.005

[2] Shafaei, Alireza, and James J. Little. "Real-time human motion capture with multiple depth cameras." In *2016 13th Conference on Computer and Robot Vision (CRV)*, pp. 24-31. IEEE, 2016. https://doi.org/10.1109/CRV.2016.25

[3] Shahroudy, Amir, Jun Liu, Tian-Tsong Ng, and Gang Wang. "Nturgb+ d: A large scale dataset for 3D human activity analysis." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010-1019. 2016. https://ieeexplore.ieee.org/document/7780484/

[4] Liu, Jun, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex KotChichung. "Nturgb+ d 120: A large-scale benchmark for 3D human activity understanding." *IEEE transactions on pattern analysis and machine intelligence*, 2019. https://ieeexplore.ieee.org/abstract/document/8713892/

[5] Rusu, Radu Bogdan, Jan Bandouch, Zoltan Csaba Marton, Nico Blodow, and Michael Beetz. "Action recognition in intelligent environments using point cloud features extracted from silhouette sequences." In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 267-272. IEEE, 2008. https://ieeexplore.ieee.org/document/8713892

[6] Li, Meng, and Howard Leung. "Graph-based approach for 3D human skeletal action recognition." *Pattern Recognition Letters* 87 195-202, 2017. https://doi.org/10.1016/j.patrec.2016.07.021

[7] Yang, Xiaodong, and Ying Li Tian. "Effective 3D action recognition using EigenJoints." *Journal of Visual Communication and Image Representation* 25, no. 1 (2014): 2-11. https://doi.org/10.1016/j.jvcir.2013.03.001

[8] Munaro, Matteo, GioiaBallin, Stefano Michieletto, and Emanuele Menegatti. "3D flow estimation for human action recognition from coloured point clouds." *Biologically Inspired Cognitive Architectures* 5: 42-51, 2013. https://doi.org/10.1016/j.bica.2013.05.008

[9] Wu, Qingqiang, Guanghua Xu, Longting Chen, Ailing Luo, and Sicong Zhang. "Human action recognition based on kinematic similarity in real-time." *PloS one* 12, no. 10, 2017. https://doi.org/10.1371/journal.pone.0185719

[10] Shi, Lei, Yifan Zhang, Jian Cheng, and Hanqing Lu. "Skeleton-based action recognition with directed graph neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7912-7921. 2019. http://openaccess.thecvf.com/content_CVPR_2019/html/Shi_Skeleton Based_Action_Recognition_With_Directed_Graph_Neural_Networks _CVPR_2019_paper.html

[11] Rusu, Radu Bogdan, and Steve Cousins. "3D is here: Point cloud library (PCL)." In *2011 IEEE international conference on robotics and automation*, pp. 1-4. IEEE, 2011. https://pointclouds.org/assets/pdf/pcl_icra2011.pdf

[12] Yang, Zhengyuan, Yuncheng Li, Jianchao Yang, and Jiebo Luo. "Action recognition with spatio–temporal visual attention on skeleton image sequences." *IEEE Transactions on Circuits and Systems for Video Technology* 29, no. 8: 2405-2415, 2018. https://ieeexplore.ieee.org/document/8428616/.

DOI: http://dx.doi.org/10.24018/ejers.2020.5.8.2004

Muhammad Sajid Khan is Assistant Professor of Software Engineering at the Army Public College of Management & Sciences, Rawalpindi, Punjab Pakistan. His research interests include 3D face reconstruction, recognition, detection and identification.

Andrew Ware is Professor of Computer Science at the, University of South Wales, United Kingdom. He is the Editor in Chief of the Annals of Emerging Technologies in Computing (AETiC). His research centres on the application of AI to help solve real-world problems.

Misha Karim is a student at the Army Public College of Management and Sciences. She has a passion for developing and applying new technologies, particularly those related to AI and Data Science.

Nisar Bahoo is a student at the Army Public College of Management and Sciences. He has a keen interest in computer vision and machine learning. He also has significant experience as an Android developer

Muhammad Junaid Khalid is a student at the Army Public College of Management and Sciences. He has a passion for AI and Digital Image Processing.

DOI: http://dx.doi.org/10.24018/ejers.2020.5.8.2004