**Paper Review Report:**

View-Invariant Human Action Recognition Based on a 3D Bio-Constrained Skeleton Model, Qiang Nie , Jiangliu Wang, Xin Wang, and Yunhui Liu , IEEE Transactions on Image Processing, Volume 28, NO. 8, August 2019, Pages 3959 – 3972, Citation: **Doi:**[10.1109/TIP.2019.2907048](https://doi.org/10.1109/TIP.2019.2907048)

**1. Introduction:**

This paper addresses two main issues seen in current human action recognition algorithms. The first issue is noisy skeleton data obtained using popular skeleton extraction algorithms that utilise Microsoft Kinect. The second problem is the lack of view invariance mechanism in current skeletonization algorithms. Lastly this paper proposes a new motion recognition and visualisation technique that utilises convolutional neural networks (CNNs).

**2. Method:**

The proposed method is split into three parts. First is the skeleton pre-processing and recovery step that generate a 3-dimensional (3D) bio constrained skeletal model to remove noisy skeletal data that violate joint motion limits, occlusions and varying bone lengths from the Kinect based skeletons. Second, is the calculation of Joint Euler Angles (JEAs) and Euclidean distance matrices (JEDM) between the joints of the recovered skeleton to create images and establish view invariance. Last is the establishment of a CNN framework to classify these images.

**2.1. Skeleton recovery and pre-processing:**

The pre-processing stage is for setting a 3D bio-constrained skeletal model which will be used to recover the skeletal poses. The constraints applied to this skeleton model include joint motion limits and fixed bone lengths. In the first step of this stage, the structure of the 3D bio-constrained skeleton is built with fixed bone lengths. The second step establishes the joint motion limits. This is done by creating 5 joint parameters. The first parameter is a 3D camera space describing the joints position as a vector. Then, the joint orientation is created, being defined by 3 JEAs. Next, the initial orientation and joint limits are defined by the initial orientation of frame 0 in the skeleton and the neutral zero method respectively. Last is the joint type defined by a joint ID. If an angle in a joint is out of the range of the joint limits, then it becomes an illegal joint.

The recovery stage aims to utilise the bio-constrained skeleton to recover a correct 3D pose from the raw skeleton. The first step of this stage is to detect valid and invalid joints by applying the constraints of the bio-constraint skeleton onto the raw skeleton. If the joint is valid then its position information is preserved by imbedding it onto the 3D bio-constrained skeleton. If the joint is invalid, then the invalid joints of the current frame are replaced with joints from the bio-constrained skeleton of the previous frame. This is achieved by minimising a function which describes the whole skeletonization process. The invalid joints are replaced by utilizing the rotation matrix of the joint being evaluated, where the rotation matrix consists of the JEAs of the joint between the frame being evaluated and the previous frame

**2.2. Motion visualization and recognition:**

The first part of the visualization stage is to encode the motion features. To do so a concatenate state vector (CSV) is produced to represent the human body state. This CSV contains the global orientation variable, a local pose consisting of the JEAs of the skeleton and a variable describing the pose state of the skeletal hands. To make it view-invariant, the global orientation variable is created by multiplying the transpose of the global rotation matrix with the gravity vector of the camera space. However, this state vector mainly contains temporal data. To incorporate more spatial data, JEDMs of the skeleton are created. Finally, a mapping function which takes the CSVs (derived from the JEAs) and JEDMs as body frames is used to convert them into two motion images.

The second part is the recognition stage. It utilises the ResNet CNN as the main framework for classifying the motion images. Batch normalisation alongside the activation function are applied to generate the output of every layer in the CNN. The predicted label is the probability of the input being the action and is the output of the last layer. It is set equal to the sum of the outputs of all the previous layers multiplied by the transpose of the weight vector, which is then inputted into a mapping function. The loss function used for training is a cross entropy loss which includes regularization. Two CNN streams are trained separately with the motion images of the JEAs and JEDMS as separate inputs. The mean of the outputs of these two streams is used to predict the action of the skeleton.

**2.3.** **Novelty of the proposed methods:**

The first novelty this paper contributes is seen in the skeleton recovery method which utilizes a bio-constrained skeleton to recover a non-erroneous skeletal pose from Kinect based skeleton. The second is the use of JEAs and JEDMs to generate motion images containing temporal and spatial data, and establishing view invariance, which according to this paper has not been conducted by anyone else.

**3. Results from experimentation:**

The experiment utilises three datasets: North-Western UCLA (NUCLA), MSRC-12 (MSRC) and NTU RGB+D (NTU). The cross-subject training protocol (CS) is applied to MSRC, the cross-view training protocol (CV) is applied to NUCLA, and both are applied to NTU. The CNN used is ResNet with 34 layers, and the activation function used is ReLU. The optimization algorithm used is the stochastic gradient descent with a momentum of 0.9, weight decay of 0.0001 and learn rate of 0.015.

For the MSRC dataset, the JEA images, the JEDM images and the use of both images combined yielded an accuracy of 88.91%, 92.93% and 94.20%, respectively. The best results were from the use of both images, outperforming almost all the other methods being compared. One exception was the method in “M. Liu, H. Liu, and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition”, *Pattern Recognition.*, vol. 68, pp. 346–362, Aug. 2017” which outperformed this method by 0.2%. For the NUCLA dataset, the JEA images, the JEDM images and the use of both images combined yielded an accuracy of 86.40%, 91.47% and 94.40%, respectively. The results from both the use of the 2 types of motion images and the JEDM images only yielded results that outperformed all the other handcrafted skeleton, RNN and CNN methods like LARP, TS-LSTM and ESV. For the NTU dataset, the JEA images, the JEDM images and the use of both motion images combined yielded an accuracy of 75.89 and 81.75%, 78.55 and 84.54% plus 86.68 and 91.79% for CS & CV, respectively. Again, the best results came from the use of both images, having outperformed all the other methods being compared. Lastly, for all datasets and all experiment types, the use of the recovered skeletons yielded better results when compared to the use of the original Kinect based skeleton.

**4. Conclusion/Future works:**

In conclusion, the methods are not only successful, but also open up new paths to explore. For example, the use of JEAs and JEDMs to generate images creates a new way in which action recognition could be conducted. As a future work to explore new methods that will produce better spatio-temporal representations of skeletal data, the performance of CNN based methods can be improved much more. This would also help to eliminate the lack of methods to address view invariance. Lastly, future research into methods like the 3D bio-constrained skeleton to remove errors in poses would also be of great benefit.