**Paper Review Report:**

Learning to recognise 3D human action from a new skeleton-based representation using deep convolutional neural networks, Huy-Hieu Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, Sergio A. Velastin, IET Computer Vision, [Volume 13, Issue 3](https://digital-library.theiet.org/content/journals/iet-cvi/13/3;jsessionid=mhpcbc6l865a.x-iet-live-01), April 2019, Pages 319-328, Citation: **Doi:**  [10.1049/iet-cvi.2018.5014](https://doi.org/10.1049/iet-cvi.2018.5014)

**1. Introduction**

This paper begins by addressing various approaches to human action recognition (HAR). It discusses traditional studies on HAR focusing on 2D acquisition methods, while pointing out the difficulties of capturing spatio-temporal data using such methods. It then presents 3D methods like RGB-D cameras as the solution to the problems of 2D methods. After this, the paper examines recognition methods using convolutional neural networks (CNNs), noting advantages such as the higher performance of deep CNNs as well as the ability of CNNs to learn complex features far better than hand crafted techniques. Therefore, this paper proposes a new skeleton-based representation utilising RGB images as well as an end-to-end learning framework consisting of deep CNNs.

**2. Method**

The method can be split into two stages. The first is the skeleton representation stage which aims to utilise a skeleton encoding method to transform 3D skeleton joint coordinates into RGB images that effectively capture the spatio-temporal data of the action that the skeleton represents. Second, is the end-to-end framework made from deep CNNs called residual networks (ResNets), which are trained with RGB images to classify human actions.

**2.1. Image based skeletal representation**

In order to utilise deep CNNs to classify human actions, the input data needs to be transformed from raw skeletal frames into images. The transformation converts the 3D joint coordinates of each skeleton in a sequence of frames into RGB images. The procedure is as follows: for joint *i* in frame *n*, a transformation function calculates the quotient of the difference between the coordinate of joint *i* and the minimum value of all the coordinates of frame *n,* and the difference between the maximum and minimum value of the coordinates of frame *n*. This quotient is then multiplied by 255 to provide an RGB colour value for the position of joint *i* in frame *n*. This procedure is repeated for all joints in all frames of a skeleton sequence. Once this is done, the frame of each sequence becomes an RGB image where the spatial information of the joints is maintained as the RGB colour values of the image. In order to maintain the temporal information, the frames are sorted according to the chronological order of the original raw skeletal frames.

**2.2. Deep CNN framework**

The framework that is used to classify the sequence of RGB images obtained from Section 2.1 utilises deep CNNs, specifically CNNs called ResNets. ResNets are used because they solve two key problems encountered by other deep CNNs, namely the vanishing gradients problem, and the degradation phenomenon which causes higher training and testing error. Five ResNets containing 20, 32, 44, 56, and 110 layers will be tested in the experiment phase.

Normally, ResNets are constructed from building blocks with 2 layers. The first layer of the building block receives an input and passes it through a convolution (Conv) function with 3 x 3 filters. This is followed by a batch normalization function (BN) and a rectified linear unit (ReLU) activation function. The second layer takes the output of the first layer as its input and passes it through a Conv and BN function. The result is then summed with the original input of the first layer before passing it through a ReLU activation function to provide the final output of the ResNet block. The modification that this paper makes to the ResNet block is to add a dropout function with a dropout rate of 0.5 between the 2 layers of the ResNet block to prevent overfitting. The ResNets are designed to take 32 x 32 images and are trained from scratch using stochastic gradient descent. The cost function that is minimised by the ResNet utilises the cross-entropy loss between the labelled actions of the RGB images and the predicted labels.

**2.3. Novelty of the proposed methods**

The novelty this paper presents is the use of a deep CNN framework constructed from ResNets, together with a simple method to encode a skeleton to an image. This paper points out that many deep learning methods for HAR utilise recurrent neural networks (RNNs), e.g., [1]. Moreover, while CNN methods have been used, e.g., [2], these methods only use simple CNN architectures and rely on complex skeleton encoding methods instead.

**3. Results from experimentation**

In the experiments conducted, the benchmark datasets used to test the ResNets were the NTU-RGB+D dataset (NTU) and the MSR Action3D dataset (MSR). As mentioned in Section 2.2, there are 5 ResNets that were trained on skeletal images, hence these ResNets will be tested on the datasets. When tested on the MSR dataset, the ResNet configuration with the best result was the one with 32 layers which achieved an average accuracy of 99.53%. When tested on the NTU dataset, the same 32-layer ResNet came out on top. For NTU using both cross-subject and cross-view settings the 32-layer ResNet achieved accuracies of 73.1% and 80.4%, respectively. When compared to state-of-the-art-results on the MSR dataset, the 32-layer ResNet outperformed all the other methods. Notably, it outperformed other deep learning methods, like the method in [3] which uses recurrent neural networks as well as the depth motion maps method in [4], both of which achieved accuracies above 95%. When compared to state-of-the-art-results on the NTU dataset, the 32-layer ResNet prevailed by outperforming the rest. Again, it outperformed other deep learning methods like the LSTM method of [5] for both the cross-subject and cross-view settings. However, the 32-layer ResNet was outperformed by the 20-layer ResNet on the NTU dataset utilising the cross-subject setting.

In terms of computational efficiency, the proposed 32-layer ResNet was tested using a subset of the MSR dataset. The tests were implemented in MATLAB and the training was conducted on the NVIDIA GeForce GTX 1080 Ti GPU. When trained using the RGB skeleton images on this GPU without parallel processing, the average time to process a sequence was 4.15 × 10−3s. During the testing stage on the MSR subset, the average time the trained ResNet took to classify a sequence was 21.84 × 10−3s. Since the time it took to train and test per sequence was very small the method has been deemed to be computationally efficient.

**4. Conclusion/Future works**

In conclusion, the proposed method did very well, as it outperformed all the other state-of-the-art methods on both test datasets. Based on this, it seems wise to suggest that future work on this method would be very beneficial. Notably, this paper mentions several things that will need to be researched or worked upon in future. The first thing this paper proposes is to use new skeleton encoding methods where the Euclidean distances between joints are exploited. It also proposes better neural network architectures such as inception-ResNets to improve feature learning and classification. Lastly, to counter the limitations in Kinect sensors this paper suggests the use of deep learning approaches which estimate human body points, such as kurtosis wavelet energy methods.

**References:**

[1] Zhu, W., Lan, C., Xing, J.*, et al.*: ‘Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks’, arXiv preprint arXiv:1603.07772, 2016

[2] Wang, P., Li, Z., Hou, Y.*, et al.*: ‘Action recognition based on joint trajectory maps using convolutional neural networks’. Proc. 2016 ACM on Multimedia Conf., Amsterdam, The Netherlands, 2016, pp. 102–106

[3] Du, Y., Wang, W., Wang, L.: ‘Hierarchical recurrent neural network for skeleton based action recognition’. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, Massachusetts, USA, 2015, pp. 1110– 1118

[4] Jin, K., Min, J., Kong, J.*, et al.*: ‘Action recognition using vague division depth motion maps’, *J. Eng.*, 2017, 1, (1), pp. 77–84.

[5] Liu, J., Shahroudy, A., Xu, D.*, et al.*: ‘Spatio-temporal LSTM with trust gates for 3D human action recognition’. European Conf. on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016, pp. 816–833