**Paper Review Report:**

CNN-based Action Recognition and Supervised Domain Adaptation on 3D Body Skeleton via Kernel Feature Maps, Yusuf Tas, Piotr Koniusz, 2018, Citation: [arXiv:1806.09078](https://arxiv.org/abs/1806.09078)

**1. Introduction**

This paper addresses the various known methods for action recognition utilising Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs), and proposes a new method which uses CNNs fed and trained with texture like feature maps made from 3D body joint sequences via kernel linearization. This paper also addresses the utilisation of supervised domain adaptation (SDA) on time sequences of 3D body skeleton joints to transfer knowledge from multiple datasets in order to improve the performance of a trained CNN.

**2. Method**

The proposed method can be split into two parts. The first is the generation of the texture like feature maps from a sequence of 3D body skeleton joints to train the CNN. The second is the SDA pipeline that is applied to the feature maps made from the source and target training sets to enable transfer learning in order to improve the performance of the CNN.

**2.1. Generating Feature Maps**

Firstly, sequences of 3D body skeleton joints are created from the available datasets and are labelled with the action that they describe (i.e., walking, jumping, etc). These sequences are then compared with one another in pairs to measure the similarity between them in terms of the positions of the 3D body skeleton joints (spatially) and their progression in time (temporally).

This is conducted by a sequence kernel consisting of two sequences that are being compared. The sequence kernel consists of two sub kernels that aim to capture the similarity between the positions of the skeletal joints and the temporal alignment respectively. This sequence kernel is then linearized, by linearizing the sub kernels which involves rewriting them as the dot product of the concatenations of the spatial and temporal variables that they are a function of. These linearized sub kernels are then inputted back into the sequence kernel to provide a linearized sequence kernel. The linearized sequence kernel is further simplified by employing the Kronecker product to separate the spatial and temporal variables of the two sequences being compared, producing the feature maps of each sequence. This results in the sequence kernel being a function of the two feature maps that describe the action of the two original sequences.

**2.2. Conducting supervised domain adaption**

With the feature maps generated, the SDA algorithm, known as second- or higher-order scatter tensors (So-HoT) is used. The So-HoT algorithm only works for features of the same class and therefore this algorithm is only used between classes shared by the target and source datasets. The SDA pipeline is as follows: once the feature maps from the source and target datasets are generated, two separate CNNs are trained with the feature maps of each dataset, one for the source and the other for the target. While training, when the classes between the source and target batches match, an alignment loss is calculated via So-HoT to facilitate transfer learning between the two CNNs. When the classes do not match, only the classification losses for both the source and target are calculated. In the pipeline data from the source and target datasets with matching classes are processed first to calculate the alignment losses. The remaining datapoints with non-matching classes are then processed to compute the classification losses of the source and target CNNs.

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

The novelty that this paper contributes is seen in the generation and use of the feature maps that are able to be used to train off the shelf CNNs, as opposed to raw skeletal data. Additionally, the application of SDA to facilitate transfer learning is also new which is evidenced by the fact that this method is not present in any of the examples this paper references.

**3. Results from experimentation**

**3.1. Setup**

The experiment is setup as follows: the CNN used for both streams is ResNet-50 which has been pre-trained on the ImageNet dataset. Two streams for both the target and source datasets are set up with these CNNs and trained using stochastic gradient descent with a momentum of 0.9 in the SDA pipeline. The datasets used are the NTU RGB-D (NTU), SBUKinect Interaction (SBU) and UTKinect-Action3D (UTK) datasets.

**3.2. Experiments**

There are 3 types of experiments conducted on all the datasets used. Experiment one consists of training and testing the CNN on only the target dataset without SDA. Experiment two consists of training the CNN on a dataset that combines the source and target datasets, and then testing the CNN on the target set only without SDA. Lastly, experiment three is when the CNNs are trained on the SDA pipeline.

**3.3. Results**

Firstly, experiment one was conducted on the SBU and UTK datasets to test the effectiveness of the kernel feature maps method against other texture-based representations, specifically the representation mentioned in the paper ‘A new representation of skeleton sequences or 3d action recognition’ by Qiuhong Ke et al. The method with the best results was the kernel feature maps method outperforming the other method by about 1.8% for SBU and 1.4% for UTK, with a total accuracy of 91.13% and 96.5% for SBU and UTK, respectively. Secondly, all three types of experiments were conducted on all datasets to test if the SDA pipeline in experiment three would be superior to the methods of experiments one and two as well as other well-known methods using the same datasets that this paper referenced. The accuracy of experiment three which enabled SDA for SBU was 94.36%, outperforming experiments one and two by 3.23 and 1.84%, respectively. Similarly, for the UTK and NTU datasets experiment three yielded an accuracy of 98.9% and 75.35%, respectively, outperforming the results of experiment one by 2.4% and 0.8%, and experiment two by 1.4% and 0.7%, respectively. Moreover, in all the experiments the methods proposed by this paper outperformed almost all of the other well-known methods referenced such as ST-LSTM + Trust Gate as well as established methods like the raw skeleton method. But, one method, from the paper mentioned in the first sentence of this paragraph, outperformed the SDA pipeline for NTU. However, the difference in performance was small being only 0.38%.

**4. Conclusion / Future works**

The proposed methods are not just successful, but also open up new paths to explore. For instance, as these methods have been successful on off-the-shelf CNNs trained with time sequences, it shows that there is some potential to use off-the-shelf CNNs as opposed to building one yourself. Additionally, the fact that well known skeleton-based action recognition datasets, were leveraged in favour of generating new raw skeletal data seems to indicate that these datasets can be used to improve upon the performance of action recognition systems using techniques like SDA in future works.