**Paper Review Report:**

Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group, Raviteja Vemulapalli, Felipe Arrate and Rama Chellappa, 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, September 2014, Pages 588-595, Citation: **Doi:**[10.1109/CVPR.2014.82](https://doi.org/10.1109/CVPR.2014.82)

**1. Introduction**

This paper addresses problems such as variations in viewpoint, occlusions and background clutter that are found in skeletonization techniques that utilize monocular RGB data. Furthermore, it addresses recent advances in depth image methods that improve upon the monocular RGB methods. This paper notes that the depth image methods produce skeletons based on the absolute geometry between either joints or body-parts. Based on the 3D geometric relationships between various body parts, this paper proposes a body-part based approach for human action recognition which aims to improve upon current body-part and joint techniques by using the relative geometries between body-parts to form the skeleton.

**2. Method**

The proposed method can be split into two stages. First is the skeleton representation stage which aims to generate a skeleton by representing the relative geometry between pairs of body-parts using a Lie group called the Special Euclidian group. Second is the temporal modelling and classification stage which aims to use dynamic time warping (DTW) and the Fourier temporal pyramid (FTP) to generate the temporal sequence of the skeleton actions, and then use a linear support vector machine (SVM) to classify them.

**2.1 Skeleton representation**

In this stage a skeleton is generated by creating n body parts and (n + 1) joints. Next a local coordinate system is created to describe the relative geometry between a pair of body parts. This is done by using a transformation matrix consisting of a rotation and translation term that are functions of time, and multiplying it with the matrix that represents the coordinates of the first body-part in the global coordinate system. This causes one end of the first body-part to be set equal to the origin and its direction to be set equal to the direction of the x-axis in the global coordinate system. This establishes the local coordinate system of the first body-part. To obtain the relative geometry of the second body-part with respect to the local coordinate system of the first, the matrix representing the coordinates of the second body-part in the global coordinate system has to be multiplied by the same transformation matrix used to derive the local coordinate system. This will result in a matrix representing the coordinates of the second body-part in the local coordinate system of the first. In a similar manner the local coordinate system of the second body-part can be established and the relative geometry of the first body-part with respect to it can be obtained. It is important to note that the transformation matrix mentioned above is an element of the special Euclidian group. Once the relative geometries of all pairs of body-parts are established, an action curve consisting of all these relative geometries is created to represent a skeletal sequence describing an action. Finally, the action curves of all the skeletons are mapped from the special Euclidian group to their Lie algebra.

**2.2 Temporal modelling and classification**

In this stage certain issues with the skeletal sequences are fixed. The first of these issues are the rate variations of the skeletal sequence. They are reduced by generating a nominal curve and warping all the action curves used for training to the nominal curve using DTW. In order to use DTW the squared Euclidean distance in the Lie algebra of the action curves is utilised. The second set of issues are the temporal misalignments and noise found in the warped action curves. To handle them FTP is employed to remove high frequency coefficients in all dimensions of the warped action curves. After this the Fourier coefficients generated by FTP are concatenated to obtain the final feature vector. Once this is done the feature vector is fed into a linear SVM to conduct one vs all classification.

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

The novelty the paper contributes is seen in the use of relative geometric representations of skeletons to describe action, which is new since other methods like the one in “Y. Yacoob and M. J. Black. Parameterized Modelling and Recognition of Activities. In *ICCV*, 1998” utilise the absolute geometry between body-parts. For instance, the generation of the relative local coordinate systems between skeleton body-parts as part of the Lie group called the special Euclidian group is one novelty. Another is the representation of the skeletal action sequences as Lie algebra action curves, which subsequently utilises DTW and FTP to obtain feature vectors for SVM classification.

**3. Results from experimentation**

In the experiments conducted, the datasets used were the MSR-Action3D dataset (MSR), UTKinect-Action dataset (UTK) and Florence3D-Action dataset (F3D). Half of the examples were used for training and the other half for testing. All experiments used a three-level FTP with a quarter length of each segment as low frequency coefficients. Finally, all experiments used a linear SVM with parameter ‘C’ set equal to 1.

First the proposed method was compared against other skeleton representations which included: joint positions, pairwise relative positions of the joints, joint angles and individual body part locations. When compared to all these methods the proposed method outperformed them all on all datasets. For MSR, UTK and F3D the average accuracy for the proposed method was 92.46%, 97.08% and 90.88%, respectively. Moreover, the amount by which the proposed method outperformed some of the other methods was large. For instance, in the F3D dataset the proposed method outperformed the individual body part locations method by 10.1%.

Next the proposed method was compared against state-of-the-art methods. For the MSR dataset utilising the protocol in [1], the proposed method outperformed all the other state-of-the-art methods being compared by achieving an accuracy of 92.46%. Notably, it outperformed the Random forests [2] and Spatial and temporal part-sets [3] methods which themselves achieved accuracies over 90%. For the same dataset utilising the protocol in [4], the proposed method outperformed the Actionlets method [4] by achieving an accuracy of 89.49%. For the UTK dataset the proposed method again outperformed the others with an accuracy of 97.08%. Furthermore, it outperformed the Histograms of 3D joints method [5] which itself had an accuracy of 90.92%. Lastly, for the F3D dataset, the proposed method once again outperformed the other method (Multi-Part Bag-of-Poses [6]) it was being compared to with an accuracy of 90.88% which was 8.88% better.

**4. Conclusion / Future works**

In conclusion, the proposed method did very well, as it not only outperformed well known skeletonization methods but also outperformed all the state-of-the-art methods being compared. Based on this, it seems safe to say that this method is something that will need to be utilised in the future and improved upon. Further on, future work on this method would have to focus on developing methods that utilise the relative geometries between specific sets of body parts rather than all body parts. This is because, human actions tend to only be characterised by a few body parts. Additionally, to improve this method, future work will also need to focus on recognising actions performed by multiple people. This, if successful, would be highly beneficial as many human actions tend to be responses to group interactions.

**References:**

[1]  W. Li, Z. Zhang, and Z. Liu. Action Recognition Based on a Bag of 3D Points. In *CVPRW*, 2010.

[2] Y. Zhu, W. Chen, and G. Guo. Fusing Spatiotemporal Features and Joints for 3D Action Recognition. In *CVPRW*, 2013.

[3] C. Wang, Y. Wang, and A. L. Yuille. An Approach to Pose- based Action Recognition. In *CVPR*, 2013.

[4] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *CVPR*, 2012.

[5] L. Xia, C. C. Chen, and J. K. Aggarwal. View Invariant Human Action Recognition Using Histograms of 3D Joints. In *CVPRW*, 2012.

[6] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo, and P. Pala. Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses. In *CVPRW*, 2013.