Metric Learning Based Action Recognition From Three Dimensional Skeleton Data

Gebze Technical University Institute of Natural and Applied Sciences,
Graduate Research Symposium, 14-16 May 2018

Seyma YUCER, Prof.Dr.Yusuf Sinan AKGUL
VisLab, Department of Computer Engineering, Gebze Technical University, Kocaeli, Turkey
Email: {syucer, akgul}@gtu.edu.tr

Abstract

In this work, we present two different methods for analyzing human actions on 3D skeleton images. The first method is a representation-based solution for human action recognition problem. This method, called the geometric bag of joints based human action recognition, utilizes the spatial-temporal behavior of 3D skeleton frames and uses SoftMax regression method to classify human actions.

The other method is a deep neural network-based method. For this method, we designed a Siamese LSTM-DML network which learns the relationship between actions and recognizes human actions using this relationship information. For each action, sub-LSTM networks extract the temporal features of human actions. Using these features, the network is trained with two-way parameter sharing and learns the deep metrics of actions. Therefore, the end-to-end trainable network can classify human actions. Unlike the other method, this method is more generalizable and can learn the relationship between human actions.

As a part of this study, we created a GTU Action 3D dataset that contains daily indoor activities and indoor sports actions. Our methods are tested against Florence Action 3D, MSR Action3D, NTU RGB+D datasets as well as our own dataset. Also, we compared our methods to related works.

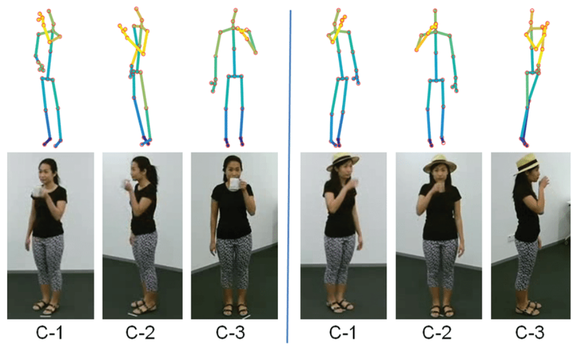

Figure I. 3D Skeleton Frames and RGB Images. [1]

Geometric Bag of Joints

The method is a representation-based solution for human action recognition problem. This method, called the geometric bag-of-joints based human action recognition, utilizes the spatial-temporal behavior of 3D skeleton frames and uses SoftMax regression method to classify human actions. A new indoor action dataset is collected by 8 subjects for this Project. This method outperforms the baseline BOW method.

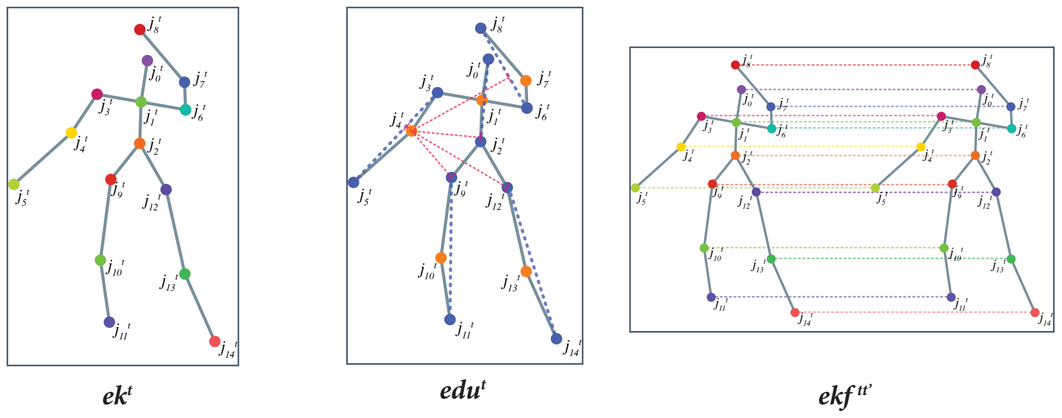

Figure II. Geometric Features: ek^t, edu^t, $ekf^{tt'}$.

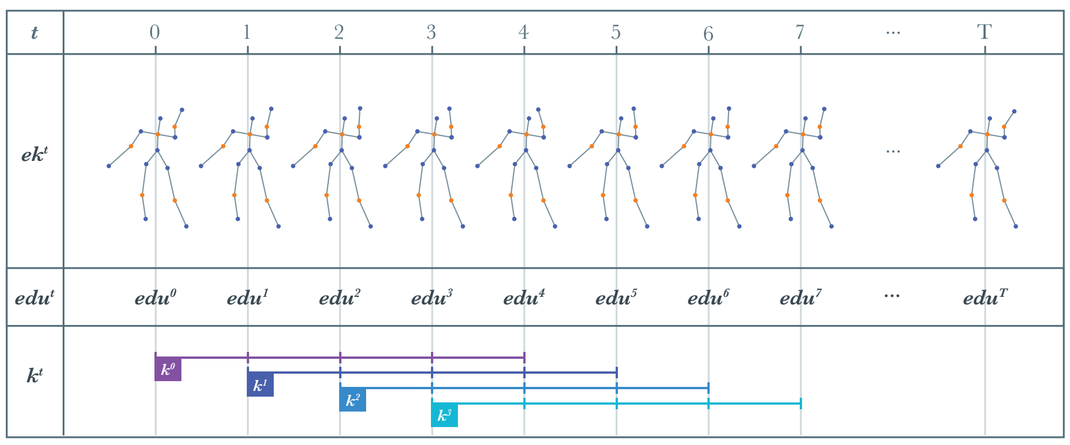

Figure III. Overview of Word Selection Phase

Siamese LSTM DML

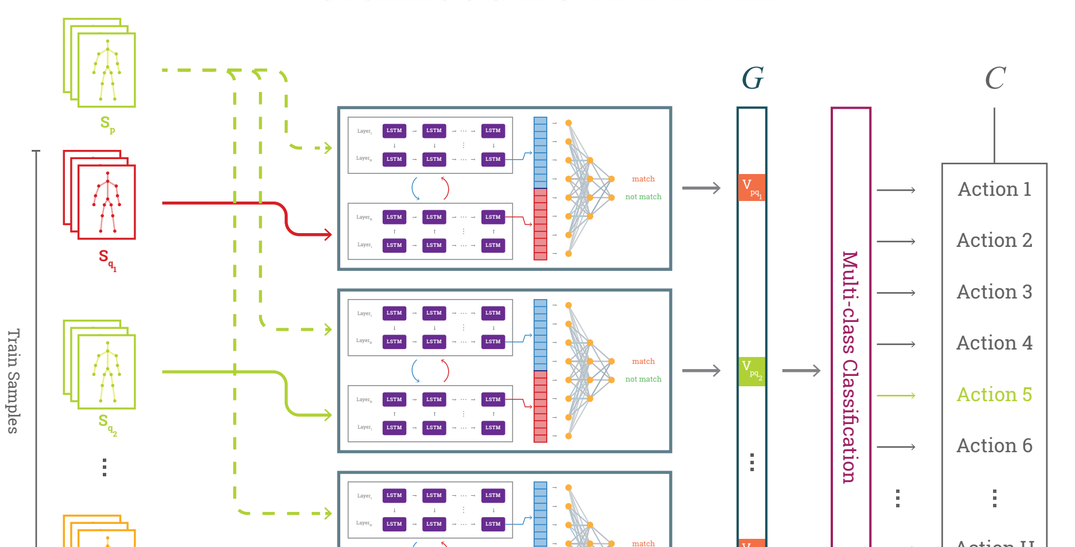

Figure IV. S-LSTM-based Deep Metric Learning Module and Multi-class Classification Module.

In this method, we pose the 3D human action recognition task as a Deep Metric Learning (DML) [7] problem which learns a similarity metric between two 3D joint sequence data using deep learning methods. One can compare two different 3D joint sequences using the automatically learned metric which can later be used for the classification of the compared sequences. The main advantage of this approach is that; we argue that it is easier to learn a similarity metric on smaller datasets than learning a classifier because it is possible to train the DML network with many different combinations of the available sequences.

Experimental Results

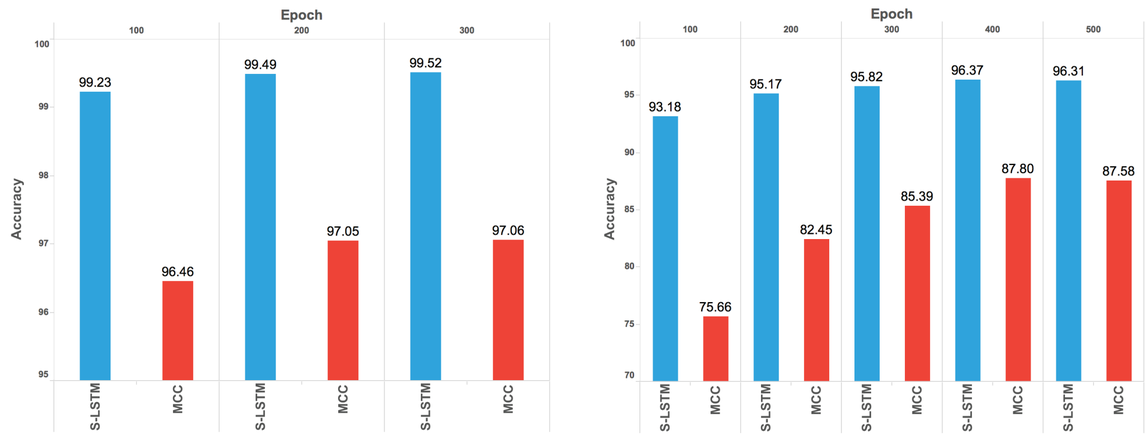

Figure V. Comparison between Siamese-LSTM Module Accuracy and Recognition Accuracy on Florence Action 3D and GTU Action 3D Datasets.

TABLE I. Results on the Florence Action 3D Dataset

Methods	Accuracy
Geometric bag-of-joints	88.00%
Multi-part Bag-of-Poses [2]	82.00%
Riemannian Manifold [3]	87.04%
Latent Variables [4]	89.67%
Lie Group [5]	90.88%
Feature Combinations [6]	94.39%
Softmax	61.61%
1-Layer LSTM	76.99%
2-Layer LSTM	72.32%
Siamese-LSTM DML	89.51%

TABLE II. Results on the GTU Action 3D Dataset

Standard Methods	Accuracy
Geometric bag-of-joints	96.50%
SOFTMAX	75.99%
1-Layer LSTM	90.21%
2-Layer LSTM	95.47%
Siamese-LSTM DML	**97.06%**

Future Work

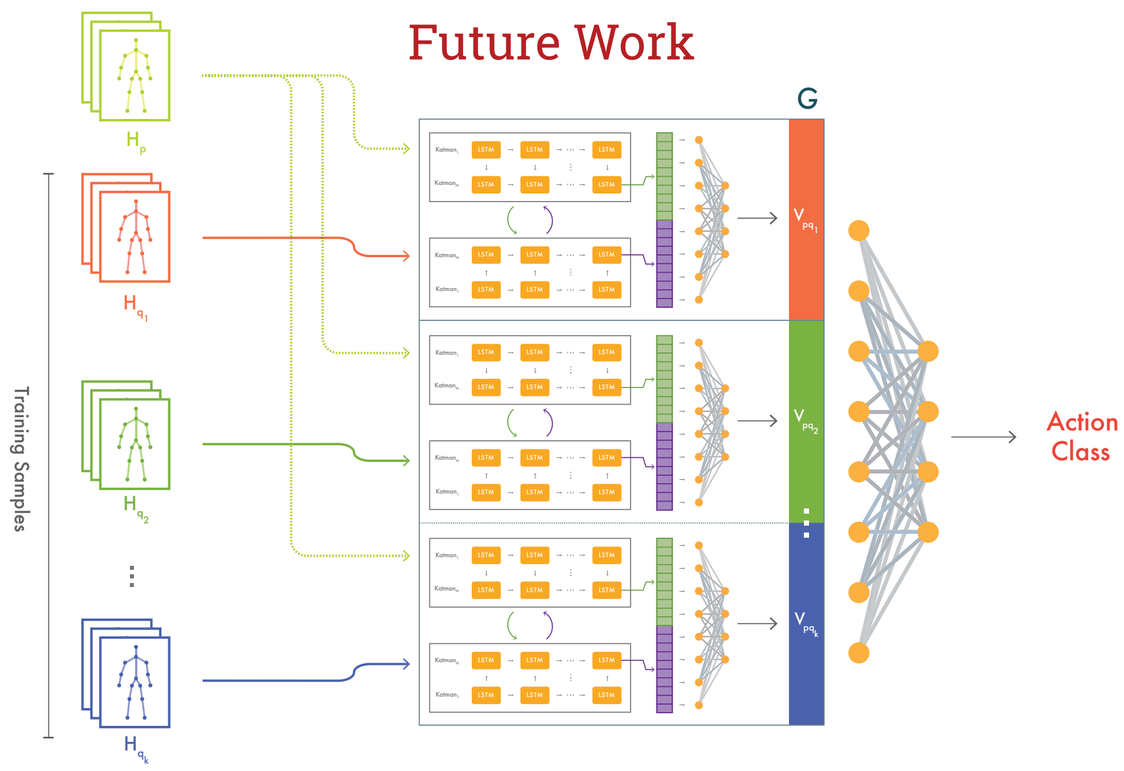

Figure VI. End-to-end Siamese LSTM Network for Action Recognition.

References

[1] Skepxels: Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Skeleton-and-RGB-sample-frames-from-the-NTU-RGB-D-Human-Activity-Dataset-30-Three_fig3_321125021 [accessed 14 May, 2018]

[2] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, ve P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses", içinde IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2013, ss. 479–485.

[3] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, ve A. Del Bimbo, "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold", IEEE Trans. Cybern., c. 45, sayı 7, ss. 1340–1352, 2015.

[4] R. Anirudh, P. Turaga, J. Su, ve A. Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables", içinde Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, c. 07–12–June, ss. 3147–3155.

[5] R. Vemulapalli, F. Arrate, ve R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group", içinde Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, ss. 588–595.

[6] D. Carbonera Luvizon, H. Tabia, ve D. Picard, "Learning features combination for human action recognition from skeleton sequences", Pattern Recognition Letters, 2016.

[7] D. Yi, Z. Lei, ve S. Z. Li, "Deep Metric Learning for Practical Person Re-Identification", Icpr, c. 11, sayı 4, ss. 1–11, 2014.