Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations

Mohamed E. Hussein[1], Marwan Torki[1], Mohammad A. Gowayyed[1], Motaz El-Saban[2]

1 Alexandria University

2 Microsoft Research Advanced Technology Labs Cairo

Problem Domain

Human action recognition from videos is a challenging machine vision task with multiple important application domains, such as human robot/ machine interaction, interactive entertainment, multimedia information retrieval, and surveillance. We present a novel approach to human action recognition from 3D skeleton sequences extracted from depth data.

Approach

- *We propose a novel fixed-size descriptor that is based on the covariance matrix of 3D skeleton joint locations, extracted from depth data.* [Fig 1]
- *We encode the temporal dependencies in the action sequence by using a hierarchical temporal pyramid of covariance descriptors on subintervals from the sequence.* [Fig 2]
- *We use an SVM classifier with linear kernel for classifying the action sequence based on the constructed descriptor.*

Experiments

- *The covariance descriptor with an off-the-shelf classification algorithm outperforms the state of the art in action recognition on multiple datasets (MSRC-12 Kinect Gesture, MSR-Action3D and HDM05-MoCap*) [Table 1]
- *We made our own annotations for MSRC-12 Kinect Gesture dataset and performed comprehensive evaluation of our descriptor on it.* [Table 2]

Descriptor Construction

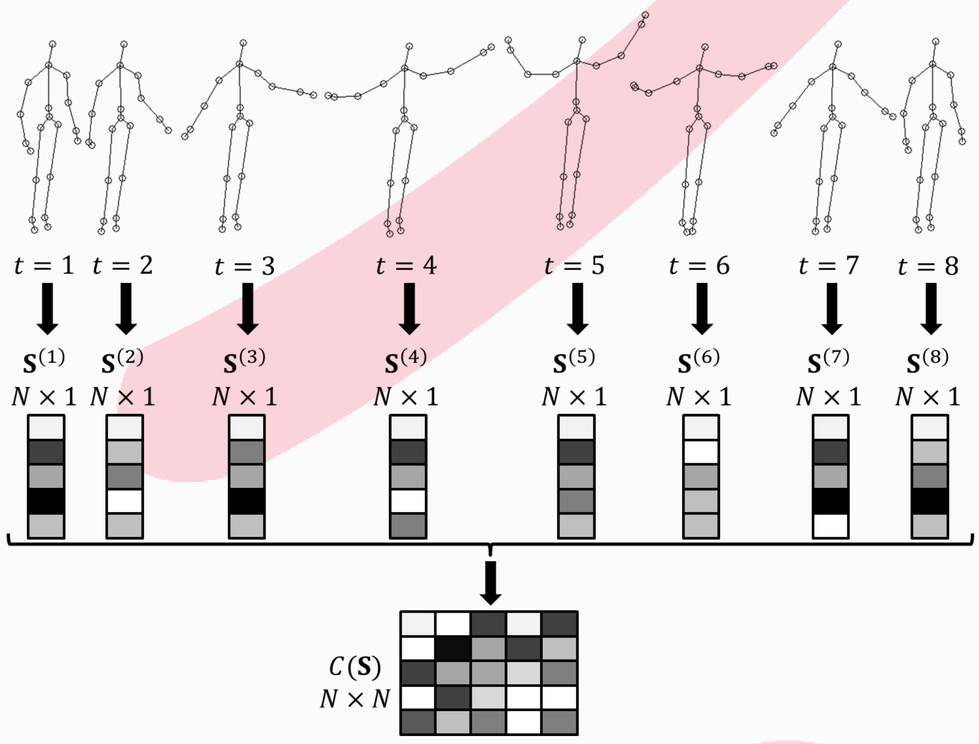

Fig 1: *A sequence of 3D joint locations of T = 8 frames is shown at the top for the "Start System" gesture from the MSRC-12 dataset. For the i^{th} frame, the vector of joint coordinates, S(i) is formed. The sample covariance matrix is then computed from these vectors.*

Temporal Construction of the Covariance Descriptor.

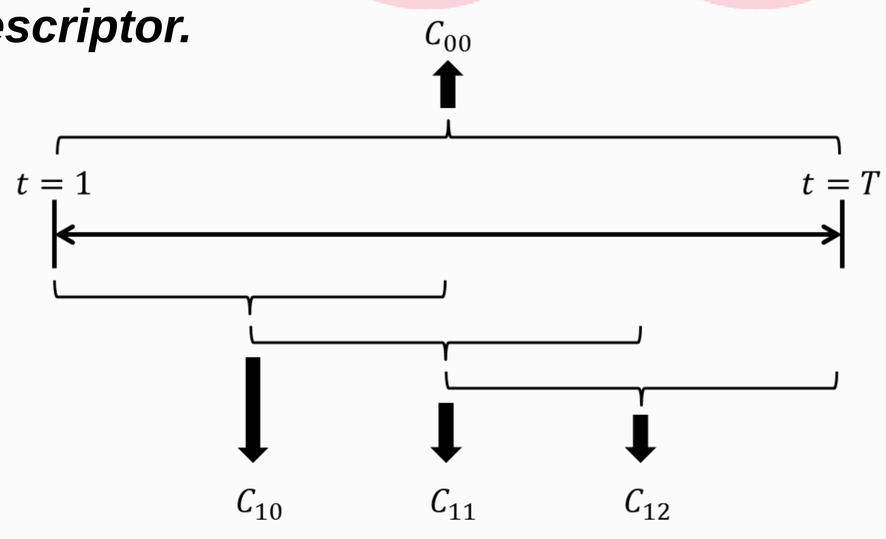

Fig 2: *C_{li} is the i^{th} covariance matrix in the l^{th} level of the hierarchy. A covariance matrix at the l^{th} level covers $T/2^l$ frames of the sequence, where T is the length of the entire sequence.*

Results

Method	Acc(%)
Rec. Neural Net. [Martens and Sutskever,2011]	42.50
Hidden Markov Model [Xia et al.,2012]	78.97
Action Graph [Li et al., 2010]	74.70
Random Occupancy Patterns [Wang et al., 2012a]	86.50
ActionLets Ensemble [Wang et al., 2012b]	88.2
Proposed Cov3DJ	90.53

Table 1: Comparative results on MSR-Action 3D dataset

Method	L=1	L=2	L=2 , OL
Leave One Out	92.7	93.6	93.6
50% subject split	90.3	91.2	91.7
1/3 Training	97.7	97.8	97.9
2/3 Training	98.6	98.7	98.7
[Ellis et al., 2013]	89.6	90.9	91.2

Table 2: Classification accuracy results for experiments on the MSRC-12 dataset with different experimental setups and different descriptor configurations.

References

[Wang et al., 2012a] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In European Conference on Computer Vision (ECCV), 2012.

[Wang et al., 2012b] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[Xia et al., 2012] L. Xia, C.C. Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3d joints. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pages 20–27, 2012.

[Li et al., 2010] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In IEEE International Workshop on CVPR for Human Communicative Behavior Analysis, 2010.

[Martens and Sutskever, 2011] J. Martens and I. Sutskever. Learning recurrent neural networks with hessian-free optimization. In Proc. 28th Int. Conf. on Machine Learning, 2011.

[Ellis et al., 2013] Chris Ellis, Syed Zain Masood, Marshall F. Tappen, Joseph J. Laviola, Jr., and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. Int. J. Comput. Vision, 101(3):420–436, February 2013.