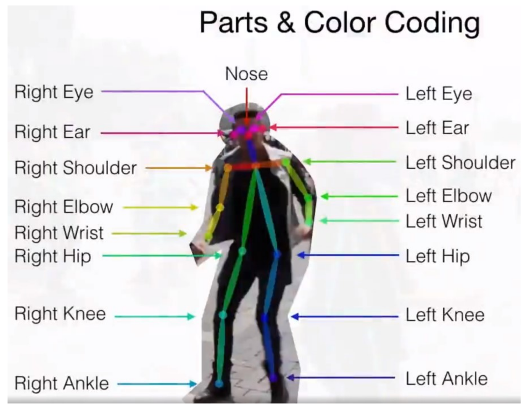
A Hierarchical Model for Action Recognition Based on Body Parts

[1]Zhejiang University of Technology, Hangzhou, China,
[2]City Univ. of Hong Kong, HK, China, [3]Soochow Univ., Suzhou, China
Zhanpeng Shao[1], You-Fu Li[2], Yao Guo[2], Jianyu Yang[3], Zhenhua Wang[1]

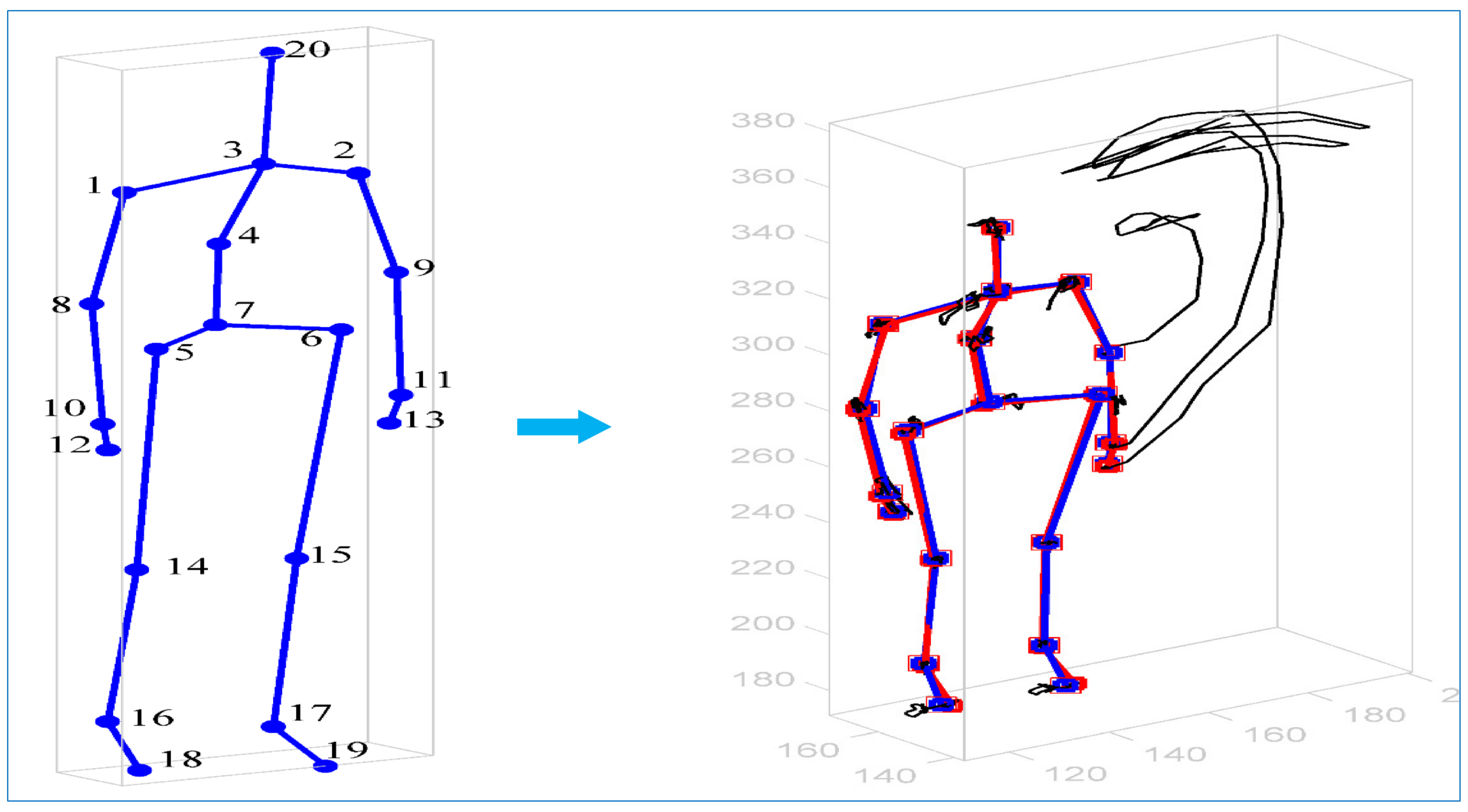
Motivation

Pose Estimation from Video [1]



Human Skeleton

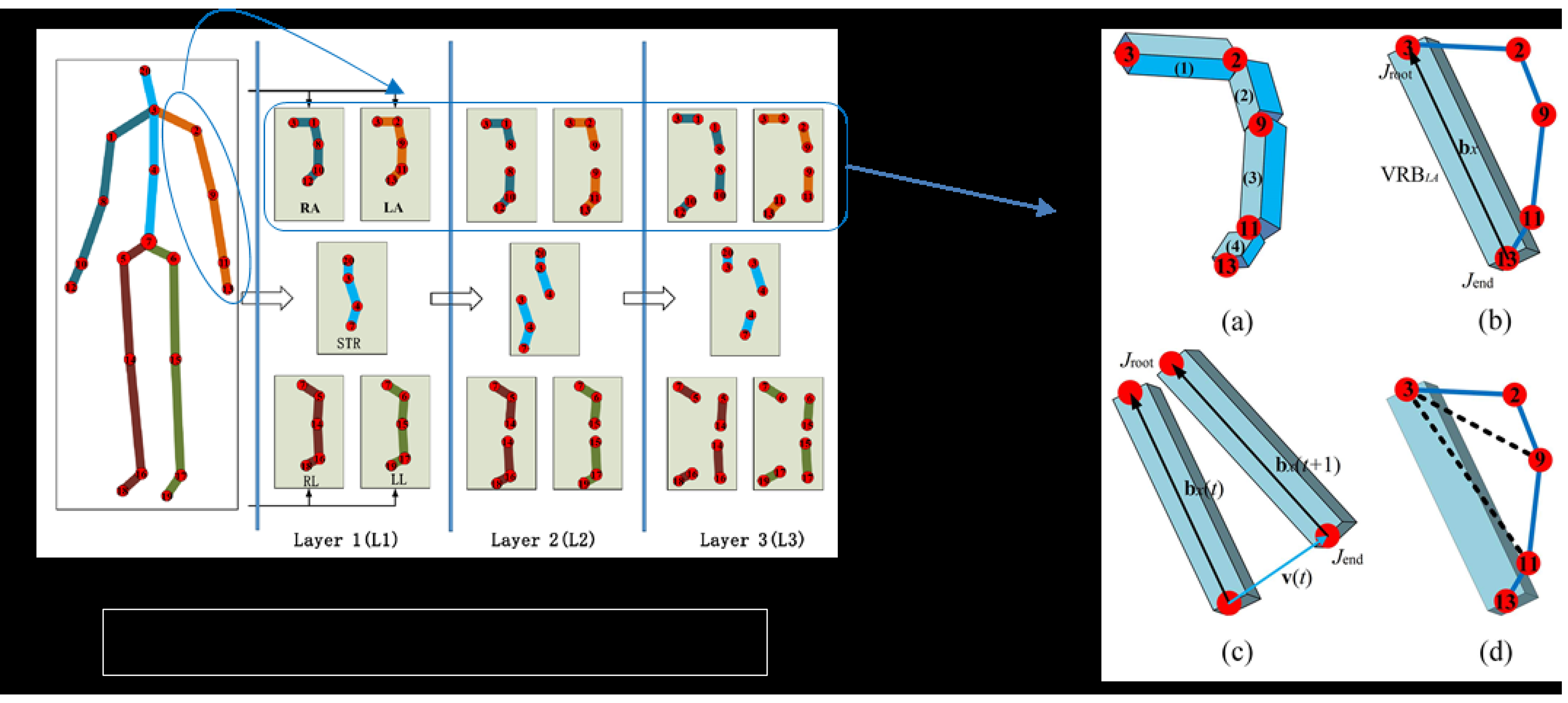
RGB+D+S channels using Depth Cameras [2]



A human action can been seen as a set of concurrent motions on multiple body-parts of the human skeleton, which could provide a very compact representation for understanding human actions

Proposed Method

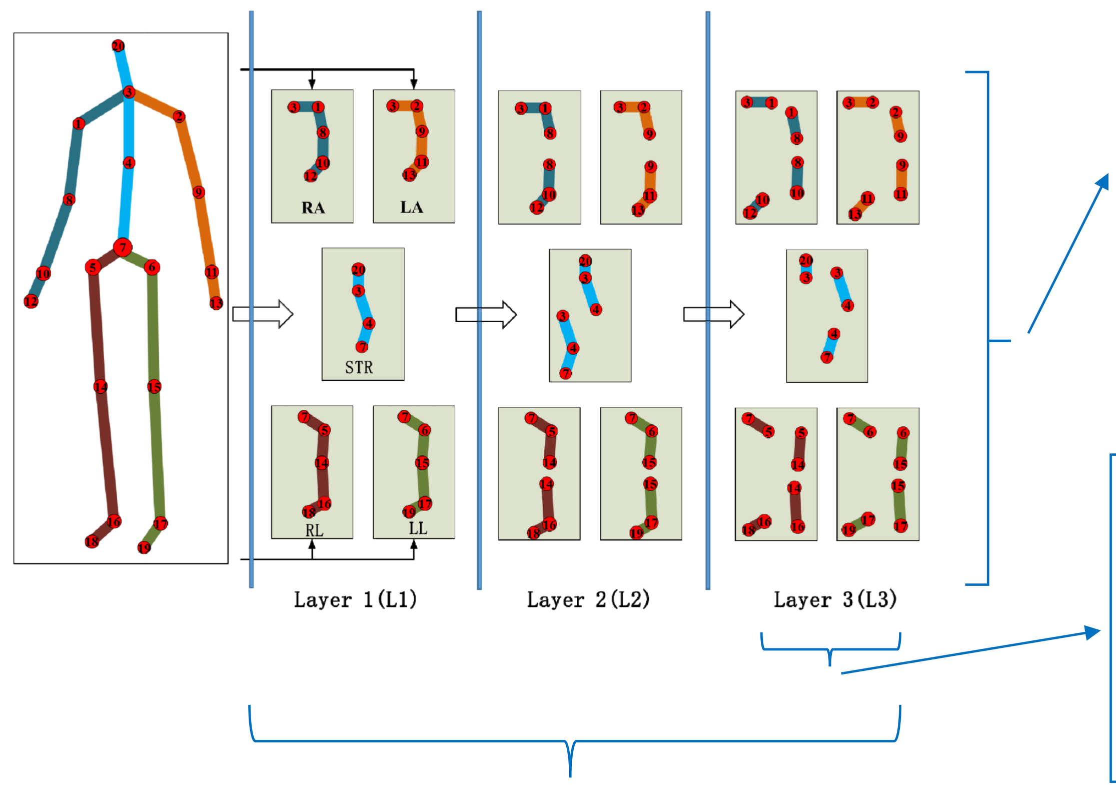
1. Hierarchical action representation



All possible virtual rigid bodies (body-parts) in the left arm

RRV descriptor of each body-part in the hierarchical model → Concatenate RRV descriptors of all body-parts in part-wise and layer-wise orders to build a hierarchical RRV descriptor for the model → FV encoding on the HRRV descriptor → Hierarchical body-parts learning

2. Hierarchical body-parts learning



(2) Apply L1 norm among body-parts (group sparsity) at the same layer to achieve active body-parts selection.

(1) Apply L4 norm based regularization inside the features of each body-part to encourage certain degree of "diversity" induction.

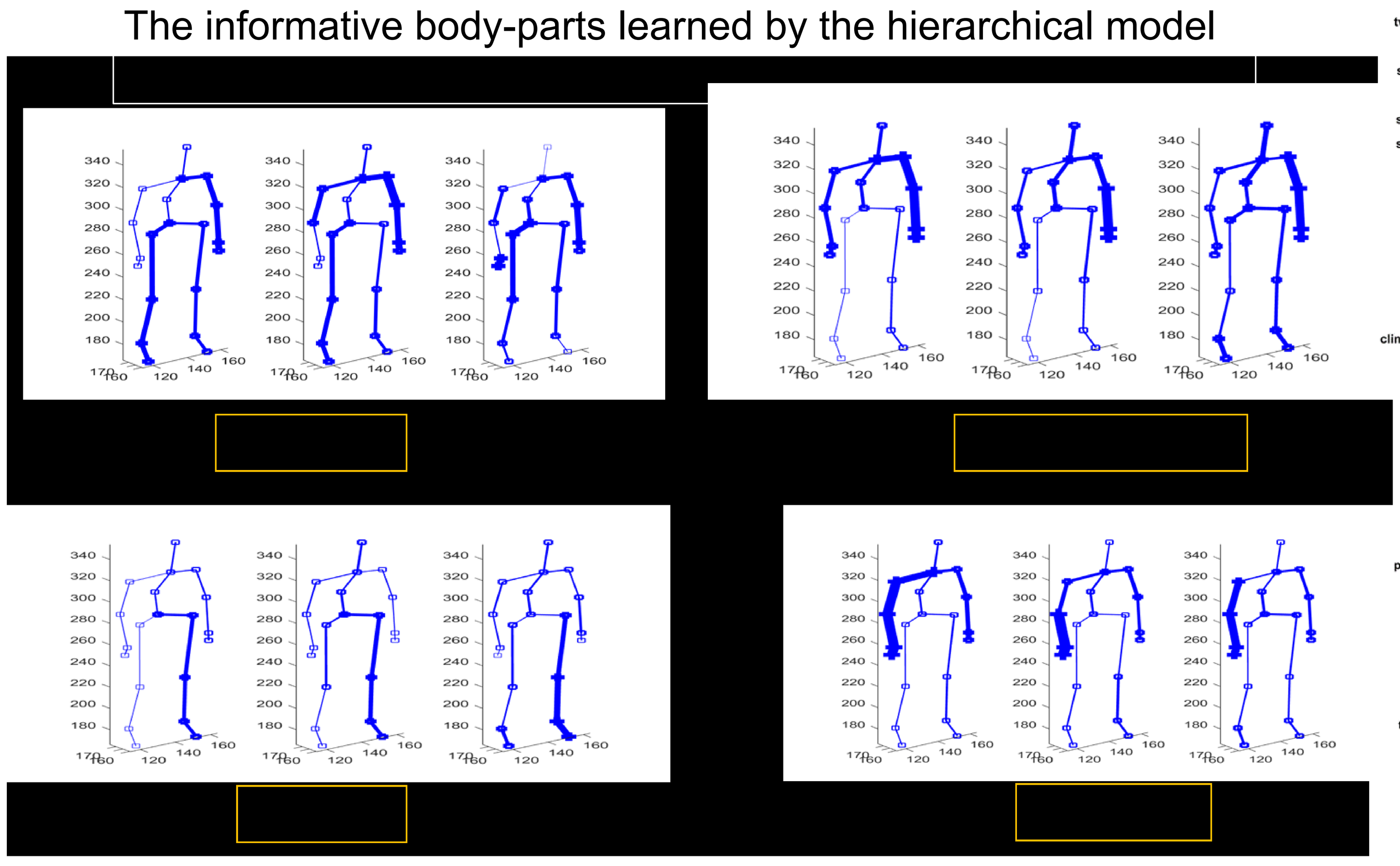
(3) Apply L2 norm over feature groups from all layers to fuse multiple bundles of body-parts.

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \left\|\mathbf{U}^{\mathrm{T}}\mathbf{W} - \mathbf{Y}\right\|_{\mathrm{F}}^2 + \lambda \sum_{c=1}^{C}\sum_{l=1}^{3}\left\|\mathbf{w}_c^l\right\|_{4,1}^2$$

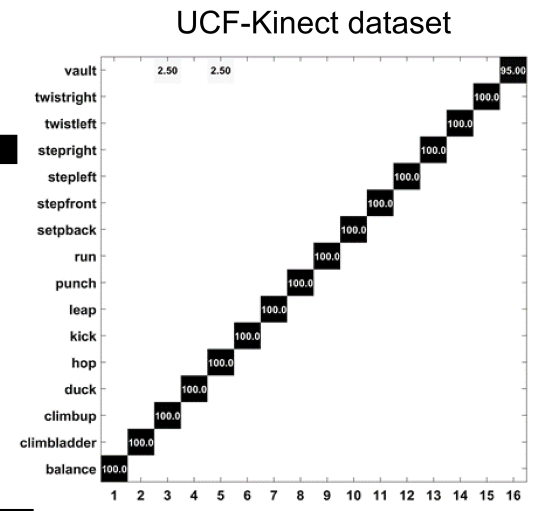
$$\mathbf{W}^* = \arg\min_{\mathbf{W}} \left\|\mathbf{U}^{\mathrm{T}}\mathbf{W} - \mathbf{Y}\right\|_{\mathrm{F}}^2 + \lambda_1 \sum_{c=1}^{C}\sum_{l=1}^{3}\left(\sum_{k_l=1}^{K_l}\left\|\mathbf{w}_c^{l,k_l}\right\|_4\right)^2 + \lambda_2 \sum_{k=1}^{d}\left\|\mathbf{w}^k\right\|_2$$

Experimental Results

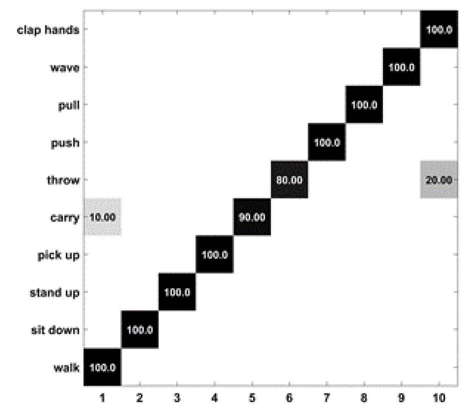
The informative body-parts learned by the hierarchical model



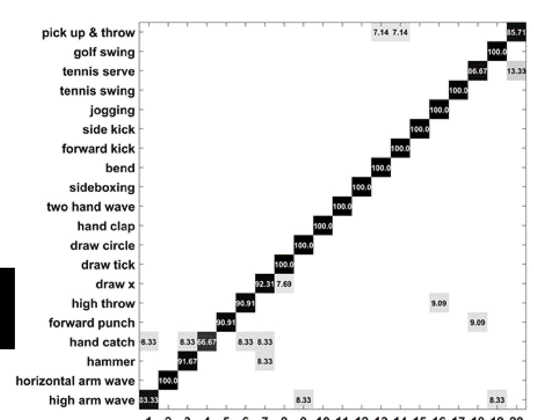
UCF-Kinect dataset



UT-Kinect dataset



MSR-Action3D dataset



Recognition Performance on the Three Datasets Using Different Variations of our HBPL Method

Methods	MSR-Action3D	UT-Kinect	UCF-Kinect
HRRV–SVM	84.98%	94.0%	98.28%
HBPL–ℓ_2 Norm	88.28%	96.0%	98.91%
HBPL–SJD	71.43%	87.0%	94.14%
HBPL(L1)	91.94%	94.0%	99.22%
HBPL(L2)	92.67%	95.0%	99.53%
HBPL(L3)	89.38%	94.0%	99.38%
HBPL(L1+L2)	93.41%	96.0%	99.53%
HBPL(L1+L3)	91.94%	95.0%	**99.69%**
HBPL(L2+L3)	92.67%	95.0%	99.77%
HBPL(L1+L2+L3)	**94.87%**	**97.0%**	**99.69%**

TABLE III

Recognition Performance on the Three Datasets Using Different Methods.

MSR-Action3D	Modality	Accuracy(%)
EigenJoints [26]	S	82.3
DMM & HOG [27]	D	85.5
Actionlet Ensemble [1]	S, D	88.2
HON4D [2]	D	88.9
DSTIP [28]	D	89.3
Motion Trajectories [29]	S	92.1
MMMP [6]	S, D	93.1
Lie Group [10]	S	89.5
Elastic Functional Coding [12]	S	85.2
LTBSVM [3]	S	91.2
Range Sample [25]	D	**95.6**
SNV [30]	S, D	93.1
Random forests [14]	S, D	94.3
Deep CNN [24]	D	**100.0**
HBPL(Ours)	S	**94.9**

UT-Kinect	Modality	Accuracy(%)
Histograms of 3D Joints [22]	S	90.9
DSTIP [28]	D	85.8
Random Forests [14]	S, D	91.9
Lie Group [10] (reported by [31])	S	93.6
Elastic Functional Coding [12]	S	94.9
Motion Trajectories [29]	S	91.5
LTBSVM [3]	S	88.5
SNV [30] (reported by [24])	S, D	88.9
Deep CNN [24]	D	90.9
ST-LSTM [31]	S	95.0
HBPL(Ours)	S	**97.0**

UCF-Kinect	Modality	Accuracy(%)
EigenJoints [26]	S	97.1
Motion Trajectories [29]	S	99.2
LAL [4]	S	95.9
LTBSVM [3]	S	97.9
Hankelets [11]	S	97.7
HBPL(Ours)	S	**100.0**

[1] Shih-En Wei et al., Convolutional pose machines, CVPR2016
[2] Shahroudy, J. Liu, T. T. Ng, and G. Wang, Ntu rgb+d: A large scale datase for 3d human activity analysis, CVPR2016