Combining Skeletal Pose with Local Motion for Human Activity Recognition

Ran Xu, Priyanshu Agarwal, Suren Kumar, Venkat N. Krovi, and Jason J. Corso

rxu2@buffalo.edu

Computer Science & Engineering

State University of New York at Buffalo, NY, USA

University at Buffalo
The State University of New York

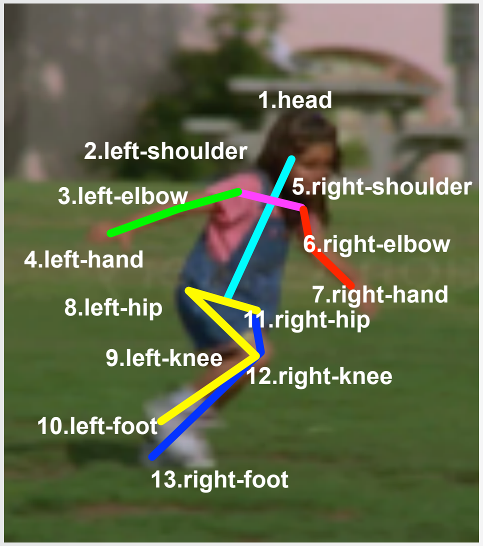
VII Conference on Articulated Motion and Deformable Objects, 12 July 2012

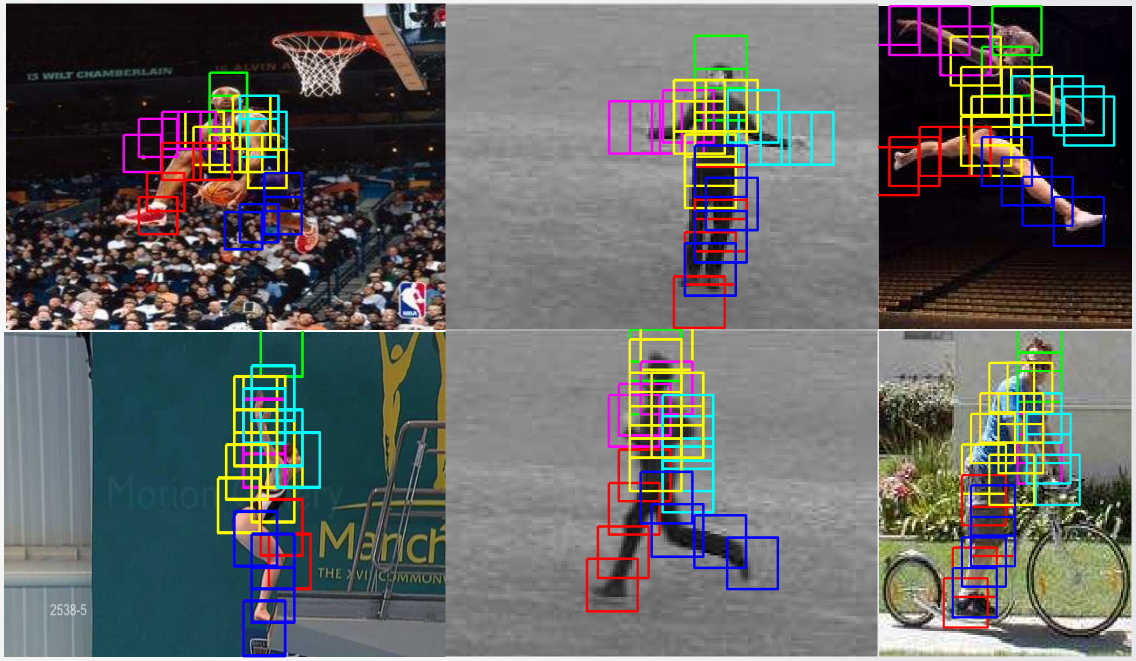
Motivation

1. Explicitly model articulated motion, instead of traditional space-time motion for activity recognition.
2. Distinguish distinct human action with similar pose distribution.

Skeletal Pose

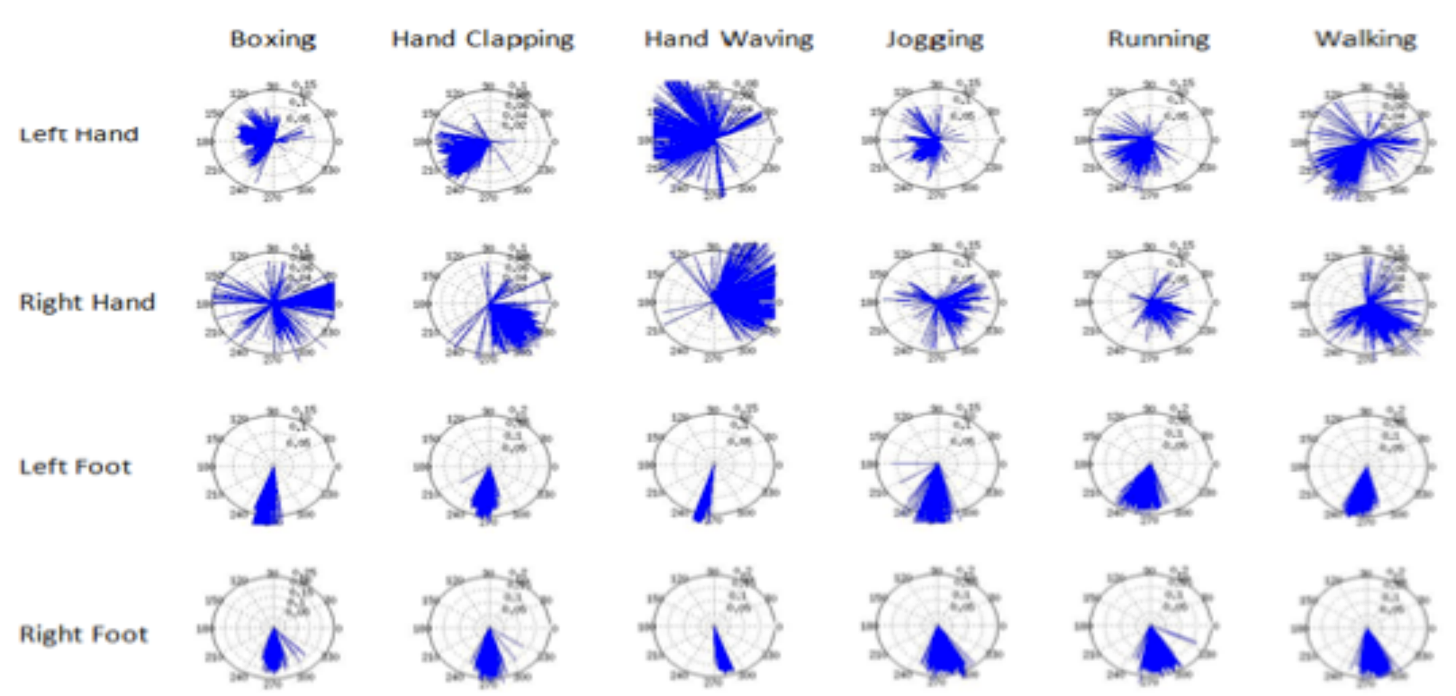
Data-driven human pose estimation[1] makes it plausible to model articulated motion explicitly.



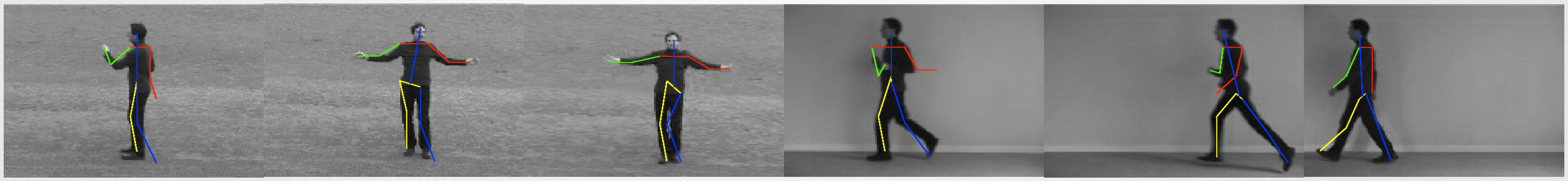


[1] Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)

Static Pose alone is not enough



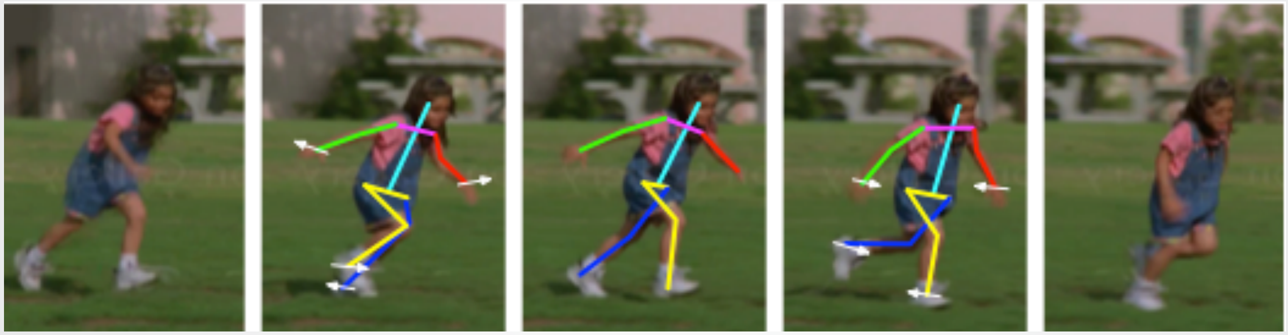
Polar histograms of limb-extreme points in human pose for the six actions in the KTH data set.

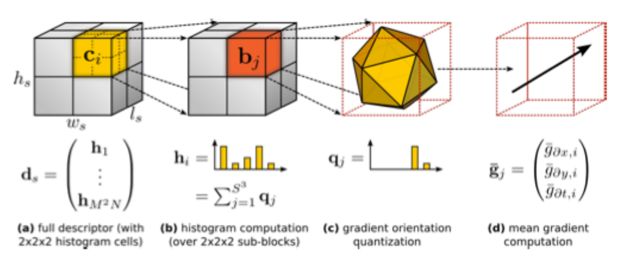


Dynamic Pose: A new mid-level representation

1. We extend the skeletal pose to incorporate local motion of the joint points, which we expect to add a richness to the pose-based representation for better descriptiveness.

2. To capture the local motion information of each skeletal joint point, we compute the histogram of oriented 3D gradients (HoG3D）in the neighborhood around the point.




$\mathbf{c}_i$
$\mathbf{b}_j$
h_s
w_s
l_s
$\mathbf{d}_s = \begin{pmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_{M^2N} \end{pmatrix}$
$\mathbf{h}_i =$
$= \sum_{j=1}^{S^3} \mathbf{q}_j$
$\mathbf{q}_j =$
$\bar{\mathbf{g}}_j = \begin{pmatrix} \bar{g}_{\partial x,i} \\ \bar{g}_{\partial y,i} \\ \bar{g}_{\partial t,i} \end{pmatrix}$

(a) full descriptor (with 2x2x2 histogram cells)

(b) histogram computation (over 2x2x2 sub-blocks)

(c) gradient orientation quantization

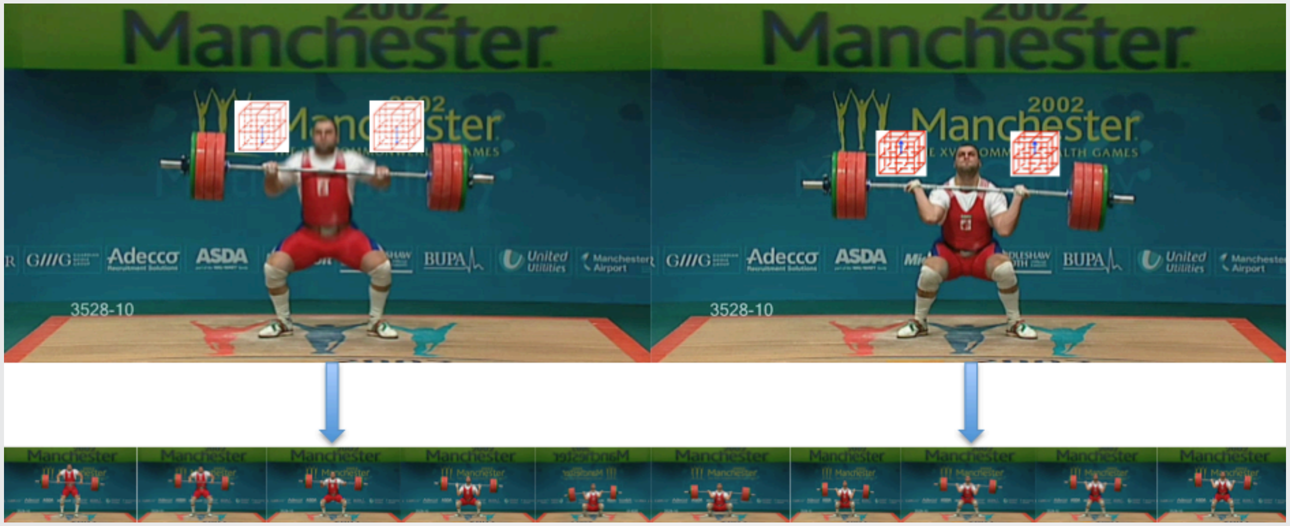
(d) mean gradient computation

Distance Function

We define threshold γ and large distance β, Euclidean distance of skeletal poses p and q is $d_i(p,q)$. At each joint i for pose p, denote the local space-time HoG3D histograms as $h_p(i)$.The distance between two dynamic poses is:

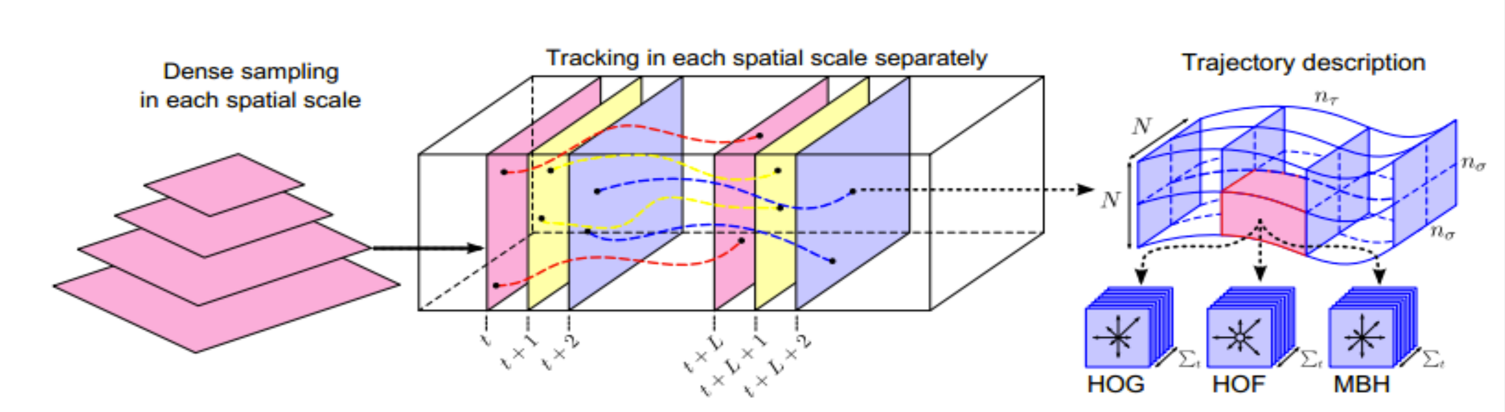
$$\delta(i) = \left\{ \begin{array}{ll} 1 - \min(h_p(i), h_q(i)) & if \;\; d_i(p,q) < \gamma \\ \beta & if \;\; d_i(p,q) >= \gamma \end{array} \right\}$$

$$D(p,q) = \sum_{i=1}^{12} \delta(i)$$

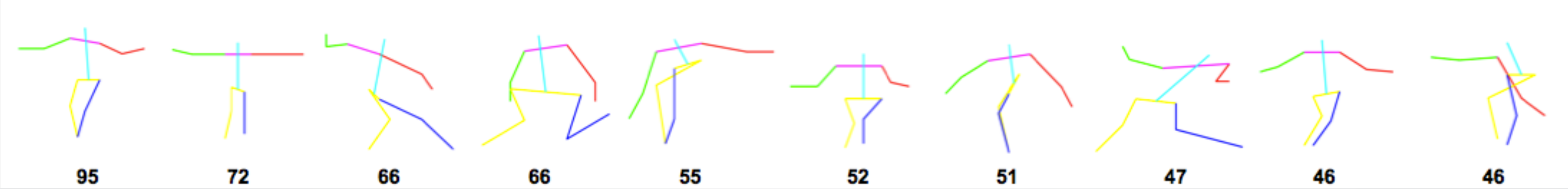


Dynamic Pose based Activity Recognition

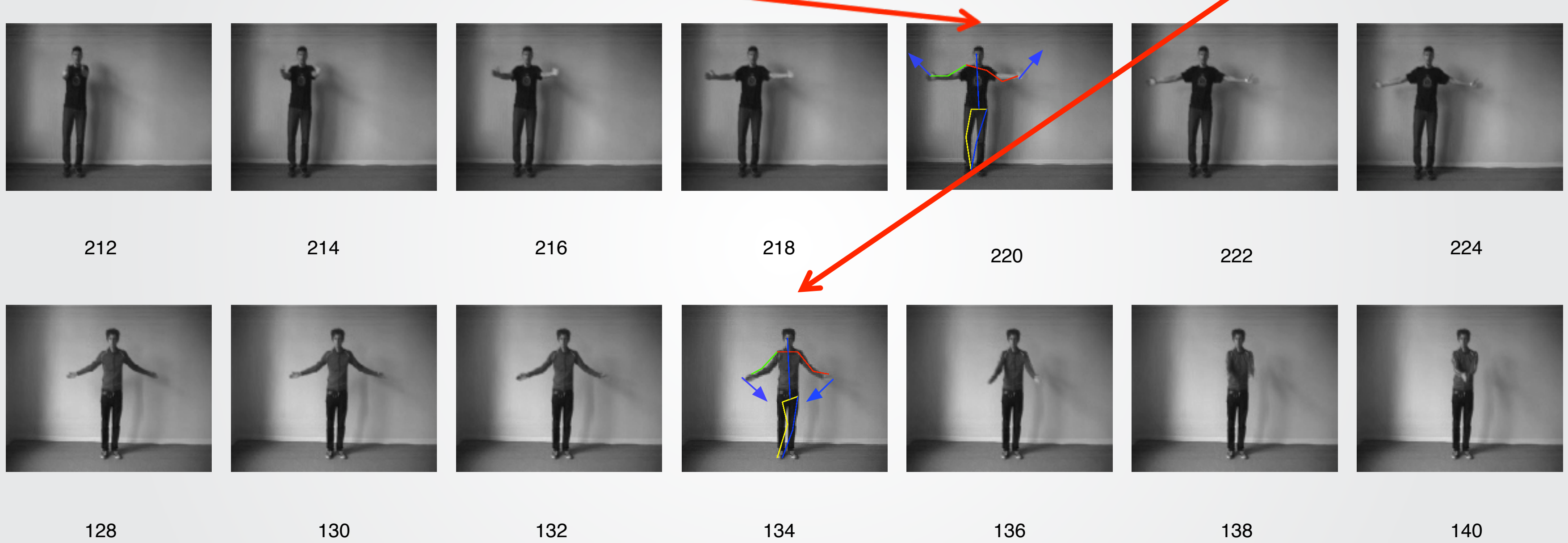
1. For skeletal pose, we construct a k-means codebook of 1000 visual words from 24-dimensional skeletal pose data using Euclidean distance. For dynamic pose, we construct codebook using our specific distance function.
2. For classification we use many one-versus-one histogram intersection kernel SVMs.
3. When fuse dynamic pose with global motion context, such as HoG3D and Dense Trajectory, we get better recognition results.



Visualization of Dynamic Pose Codebook



The top ten canonical dynamic poses in the learned codebook.



The 1st and 9th dynamic pose codebook centroids visualized in video sequence. The first row corresponds to 220th frame of video person23_handclapping_d4_uncomp.avi, and the second row corresponds to 134th frame of video person25_handclapping_d4_uncomp.avi.

Experimental Result

Method	KTH	UCF-Sports	Method	KTH	UCF-Sports
BoP	76.39%	71.33%	Kovashka [2]	94.53%	87.27%
BoDP	91.20%	81.33%	Brendel [3]	94.20%	77.80%
HoG3D	82.41%	76.67%	Gaidon [4]	94.90%	90.30%
Dense Trajectory	95.33%	83.33%	Yao [5]	92.00%	86.60%
BoDP+HoG3D	89.35%	86.67%	Tran [6]	95.67%	88.83%
BoDP+DT	97.22%	87.33%	Sadanand [7]	98.20%	95.00%

[2] res
for

[3] Brendel, W., Todorovic, S.: Activities as time series of human postures. In: ECCV 2010.

[4] Gaidon, A., Harchaoui, Z., Schmid, C.: A time series kernel for action recognition. In: BMVC 2011.

[5] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In CVPR, 2010.

[6] Tran, K.N., Kakadiaris, I.A., Shah, S.K.: Modeling motion of body parts for action recognition. In: BMVC, 2011.

[7] Sreemananth Sadanand, Jason Corso. Action Bank: A High-Level Representation of Activity in Video. In CVPR, 2012.

BoP vs BoDP in KTH

BoP

	hw	bx	wk	jg	cl	rn
handwaving	0.89	0.03	0.06	0.03	0	0
boxing	0	0.64	0	0.03	0.28	0.06
walking	0.03	0.06	0.86	0.03	0	0.03
jogging	0	0	0.14	0.83	0.03	0
clapping	0.03	0.25	0	0	0.61	0.11
running	0	0.14	0	0	0.11	0.75

BoDP

	hw	bx	wk	jg	cl	rn
handwaving	1	0	0	0	0	0
boxing	0	0.81	0	0	0.11	0.08
walking	0.06	0	0.92	0.03	0	0
jogging	0	0	0	1	0	0
clapping	0	0.17	0.03	0	0.78	0.03
running	0	0.03	0	0	0	0.97

BoP vs BoDP in UCF Sports

BoP

	dv	gf	kk	lf	rd	rn	sk	sb	hs	wk
diving	0.71	0	0	0	0.14	0.07	0	0.07	0	0
golfing	0.06	0.78	0.17	0	0	0	0	0	0	0
kicking	0	0.05	0.70	0	0	0.10	0.10	0.05	0	0
lifting	0	0	0	0.67	0.17	0	0.17	0	0	0
riding	0	0.08	0.08	0	0.58	0.08	0	0.17	0	0
running	0	0	0.08	0	0.08	0.77	0	0	0	0.08
skating	0.17	0.08	0.08	0	0	0	0.08	0	0.08	0.50
swing-bench	0.05	0	0	0	0.05	0	0	0.90	0	0
h-swinging	0.08	0	0	0	0	0	0.08	0	0.85	0
walking	0	0	0.05	0	0	0	0.14	0	0	0.82

BoDP

	dv	gf	kk	lf	rd	rn	sk	sb	hs	wk
diving	1	0	0	0	0	0	0	0	0	0
golfing	0	0.83	0.06	0	0	0	0.11	0	0	0
kicking	0	0.05	0.90	0	0	0.05	0	0	0	0
lifting	0	0.17	0	0.67	0	0	0.17	0	0	0
riding	0	0.08	0.08	0	0.67	0.08	0.08	0	0	0
running	0	0	0.23	0	0	0.62	0.08	0	0	0.08
skating	0	0.17	0.17	0	0	0.08	0.33	0	0.08	0.17
swing-bench	0	0	0	0	0	0	0	1	0	0
h-swinging	0	0	0	0	0	0	0.08	0	0.92	0
walking	0	0	0	0	0	0	0.14	0	0	0.86

Fusion with Dense Trajectory in KTH

DT

	hw	bx	wk	jg	cl	rn
handwaving	1	0	0	0	0	0
boxing	0	0.97	0	0	0.03	0
walking	0.06	0	0.94	0	0	0
jogging	0	0	0	1	0	0
clapping	0	0.19	0	0	0.81	0
running	0	0	0	0	0	1

BoDP+DT

	hw	bx	wk	jg	cl	rn
handwaving	1	0	0	0	0	0
boxing	0	1	0	0	0	0
walking	0	0	1	0	0	0
jogging	0	0	0	1	0	0
clapping	0	0.17	0	0	0.83	0
running	0	0	0	0	0	1

Fusion with HoG3D in UCF-Sports

HoG3D

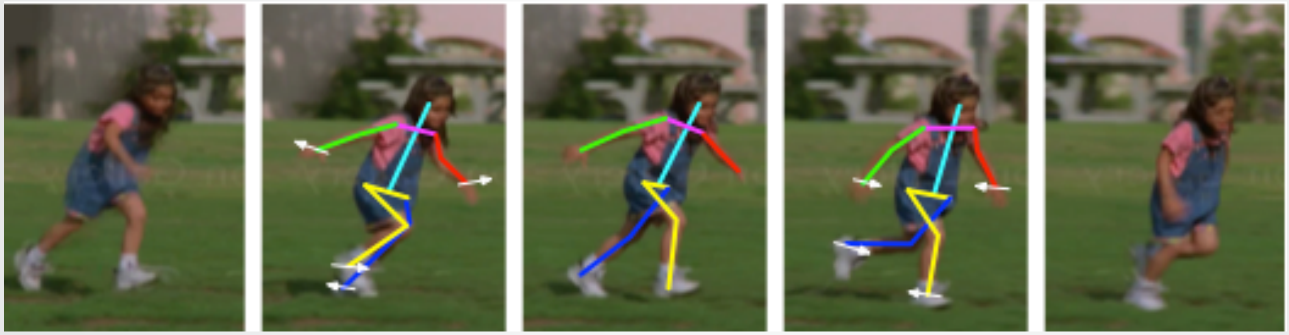
	dv	gf	kk	lf	rd	rn	sk	sb	hs	wk
diving	1	0	0	0	0	0	0	0	0	0
golfing	0	0.78	0	0	0.06	0	0.06	0	0	0.11
kicking	0	0	0.70	0	0.15	0.05	0.05	0	0	0.05
lifting	0	0	0	0.67	0	0	0	0	0	0.33
riding	0	0.08	0.42	0	0.50	0	0	0	0	0
running	0.08	0.08	0.31	0	0	0.46	0	0.08	0	0
skating	0	0.17	0	0	0	0	0.42	0.08	0.08	0.25
swing-bench	0	0	0	0	0	0	0	1	0	0
h-swinging	0	0	0	0	0	0	0.08	0	0.92	0
walking	0	0	0.05	0	0	0	0.05	0.05	0	0.86

HoG3D+BoDP

	dv	gf	kk	lf	rd	rn	sk	sb	hs	wk
diving	1	0	0	0	0	0	0	0	0	0
golfing	0	1	0	0	0	0	0	0	0	0
kicking	0	0	0.90	0	0	0.10	0	0	0	0
lifting	0	0	0	0.67	0	0	0	0	0	0.33
riding	0	0	0.17	0	0.83	0	0	0	0	0
running	0	0	0.08	0	0	0.77	0	0	0	0.15
skating	0	0.17	0	0	0	0	0.50	0	0	0.33
swing-bench	0	0	0	0	0	0	0	1	0	0
h-swinging	0	0	0.08	0	0	0	0.08	0	0.85	0
walking	0	0	0.05	0	0	0	0.09	0	0	0.86

Conclusion

- We propose a new middle level representation of articulated human action—dynamic pose that adds local motion information to skeletal joint points.
- We have implemented our representation in an activity recognition setting using bag of features with kernel intersection SVM as the base classifier.
- When fusion with global motion context, dynamic pose achieves state-of-the-art activity recognition results on two benchmark dataset.



Acknowledgement This work was partially supported by the National Science Foundation CAREER grant (IIS-0845282), the Army Research Office (W911NF- 11-1-0090), and the DARPA Mind's Eye program(W911NF-10-2-0062).

Thank You

Q&A