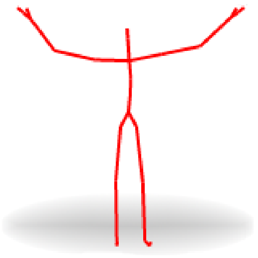
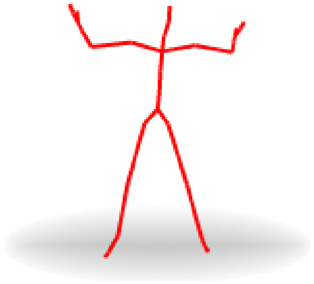




Motion Capture Data

Similarity | Classification

PETR ELIÁŠ

03/2015

DISA LABORATORY
FACULTY OF INFORMATICS
MASARYK UNIVERSITY

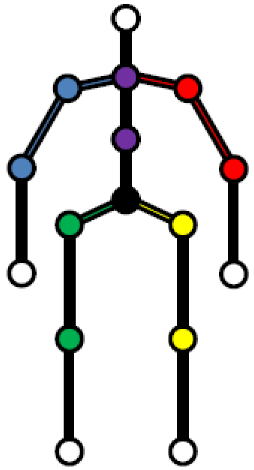
Contents – Motion Capture Data

I. Introduction and Challenges

II. Classification and Current approaches

III. Our approach – Principles and Challenges

IV. Our approach – Results and Future work

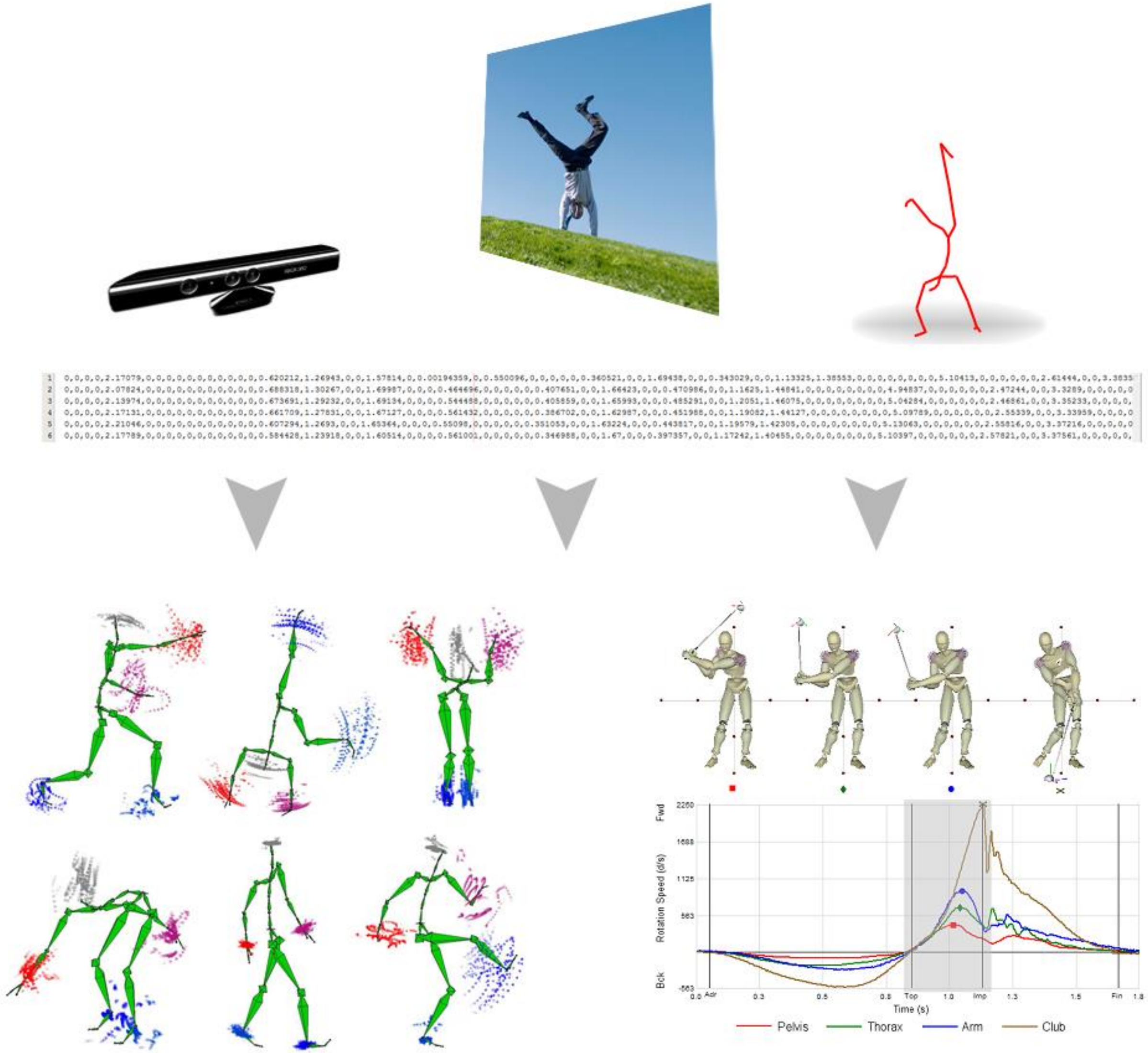


Introduction

Motion Capture (MOCAP) Data

Digital approximation *of* ***motions*** *carried out by* ***observed subjects*** *that are* ***captured*** *for further* ***inspection*** *and* ***applications****.*

- **Digital approximation** - (x, y, z) coordinate for each tracked joint and each frame (<120fps)
- **Motions** such as gait (walking), facial expression, interactions, whole-body actions
- **Observed subjects** are so far commonly individual humans
- **Captured by** devices based on various technologies (Kinect, OptiTrack, xSens, …)
- **Inspected** for analysis, action detection, action recognition, classification, reconstruction
- **Applications** in medicine, sports, security, entertainment (movies, games), robotics …


Rotation Speed (d/s)
Time (s)
Pelvis
Thorax
Arm
Club

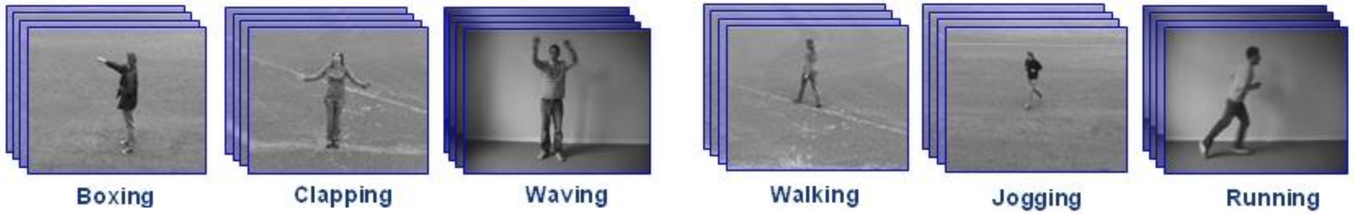
General Challenges

- Too much information on input (**complexity**)
- High cost of processing the original data (**efficiency**)
- Feature extraction and dimension reduction (**effectivity**)
- Various scenarios, various lengths of motions, various data sets (**adaptability**)

- **Applications are highly scenario-dependant**
  no general definition of MOCAP data similarity
  no accepted universal solution for action recognition or classification

Motion Data Classification

Identifying a category/categories of observed instance
on the basis of observations whose category membership is known.



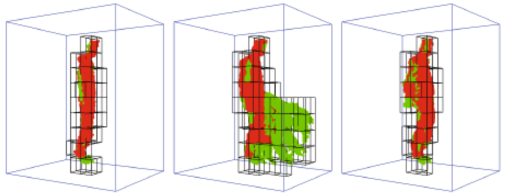
Challenges

- Different actions are performed differently by different actors
- Scope ranging from microgestures (mimics) to complex exercises (dancing)
- Relative vs absolute moves (jog vs jog on place)
- Rotation of actor (run vs run in circle)
- Various frame rates, body sizes, data quality, number of tracked joints, …

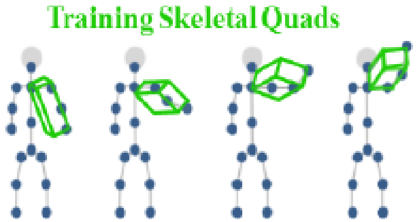
Classification Approaches

Features (generally simple)

relative distances or angles between joints, most informative joints, velocity changes, absolute coordinates, space-time occupancy, skeletal quads, covariance of 3D Joints, flexible dictionary of action primitives, …



combined with



Classifier (generally complex)

Distance Based: Dynamic Time Warping, k-NN, … **and Machine learning based:** Support Vector Machines, Neural Networks, Hidden Markov Models, Boltzman machines, …

Our Approach – Main Idea

1) Find effective transformation from (dynamic) motion capture data into (static) images.

2) Classify image based on their visual similarity to others based on known approaches (k-NN classifier on Caffe descriptors)

Our Approach – Motivation

- **Visualization** of motion data provides humans with **better understanding** compared to set of high-dimensional vectors
- **Comparing** visual similarity of **images is a known concept** nowadays - it achieves high precision and many techniques might be employed
- Instead of finding complex solution to a problem sometime it is easier to **reduce the problem** into another problem that **already has known solution**
- **Universality** (scenario independance) of this approach - by selecting a proper transformation function that **visually** differentiates target classification categories

Our Approach – Process

MOTIONS AS IMAGES

MOCAP data

hdm05, 1464 motions
120fps, 31 joints
15 categories
(rotate arms, punch, …)

CONVERT

Images

1 move = 1 image
Width = #no of frames
Height = #no of joints

EXTRACT

Caffe Descriptors

Convolutional Neural Network

Trained on 1.2M set of images (mostly photographs)

Output is 4096 dimensional vector

1s extraction
for each image

IMPORT

Metric Space Instances

Using MESSIF framework

CLASSIFY

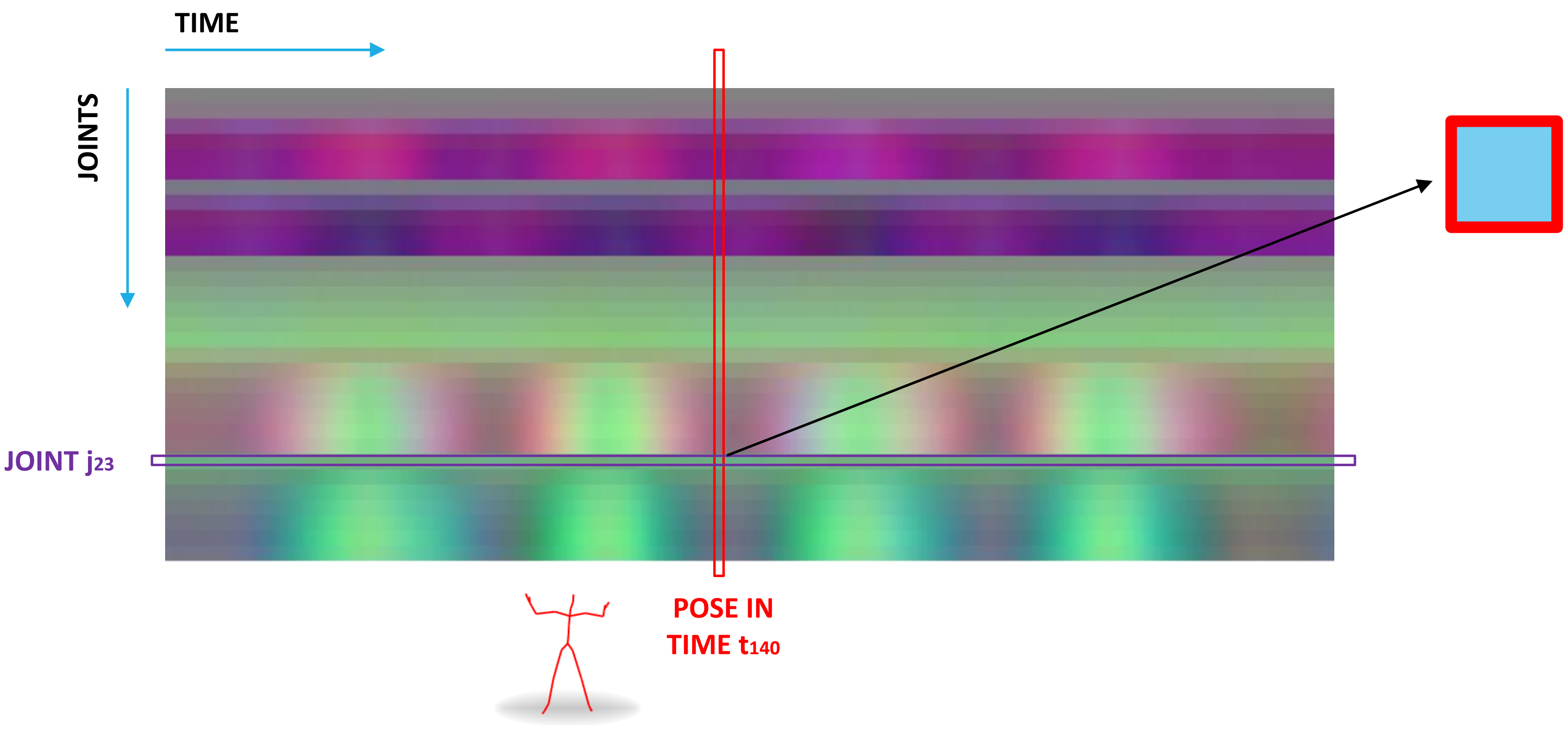
Classifier

1-NN

Weighted k-NN

Our Approach – Motions as Images

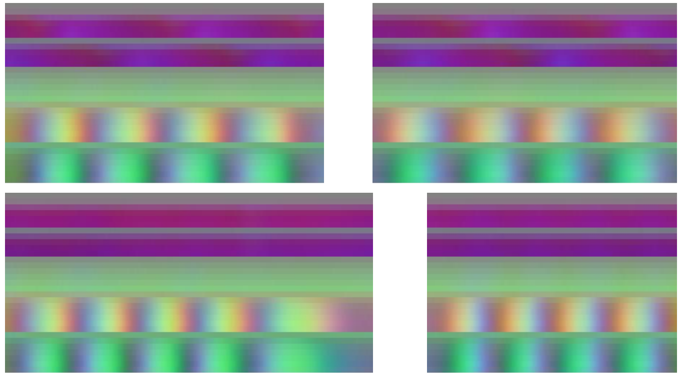
Every motion is a time series of (x, y, z) coordinates of all tracked joints.



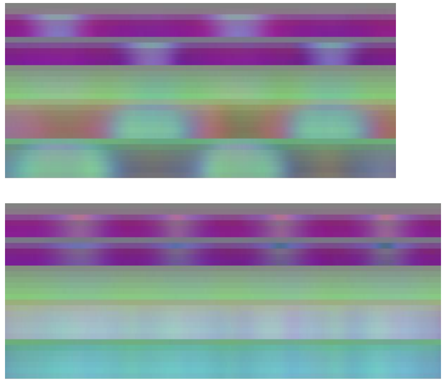
Color of pixel (140, 23) is given by RGB(X, Y, Z) where X, Y, Z are the coordinates of joint 23 in time 140 normalized* over the whole dataset to range (0, 255).

Minimum over all x, y, z coordinates over all joints over all poses over all seqeuences will get 0. Maximum 255 respectively.

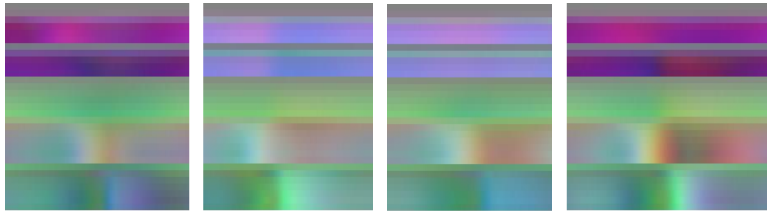
Our Approach – Motions as Images



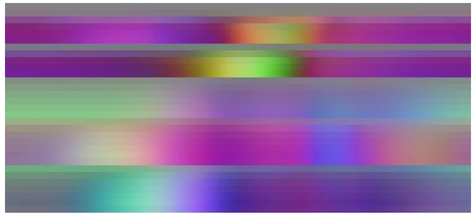
Rotate arms



Exercise



Throw right hand



Cartwheel

Kick

Our Approach – Challenges

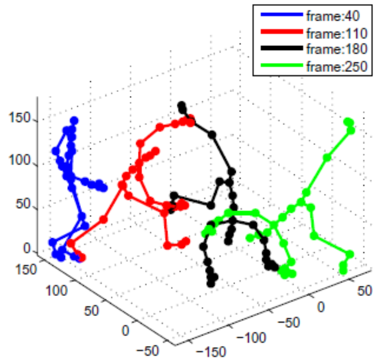
- Notion of time
  - Various speed of performances
  - Various lengths of actions
- Normalization
  - Initial rotation of subject (rotate by hips, first frame, all frames)
  - Centering in space (put root joint in (0, 0, 0), first frame, all frames)
  - Human skeleton size (infant vs adult, bones size normalization)
  - Range normalization (into RGB or other target space)
- Segmentation
- Action recognition in longer sequences

Normalization

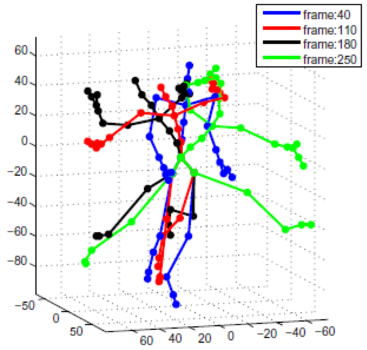
I. Pose centering
Root joint to (0, 0, 0)

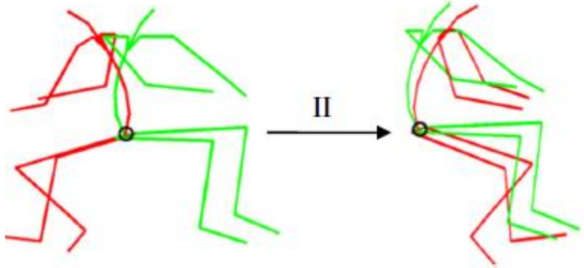
II. Pose rotation by angle φ
Rotation along y-axis by angle φ
is determined as an angle between
z-axis and straight line connecting left and right hip
in a y-projected 2D space (x, z)

III. Coordinates values normalization
Reduction to desired range such as RGB or (0, 1)



I





Results – Confusion Matrix

hdm05 | 1464 motions | 15 categories | 1-NN classification | 93.17% precision

ID	MOVE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	#Ns
1	cartwheel	100															6
2	grabDepR		96		4												105
3	kick			98	2												49
4	move		0,2		93		2				0,5					4	430
5	punch					100											48
6	rotateArms				11		89										46
7	sitLieDown		2		2			95									43
8	standUp				2				95							2	43
9	throwR				4					96							23
10	jump				12						84	4					25
11	hopOneLeg				6							94					18
12	neutral												83	1		16	75
13	tpose								1				2	98			198
14	exercise				11						5				84		19
15	turn		0,3		2								7			91	336

Other Approach Comparison

Action	N_s	N_f	pos	pw	cen	key
cartwheelLHandS	21	8627	100	100	100	100
clapAboveHead(1)	14	6102	100	100	100	100
depositLowR	28	7767	100	75.0	100	100
elbowToKnLeS (7)	13	5756	100	100	100	100
hitRHandHead	13	2943	84.6	92.3	7.69	92.3
hopBothLegs	36	3462	61.1	91.7	41.7	91.7
hopLLeg	41	3080	100	100	95.1	100
hopRLeg	42	3107	100	100	100	100
jogLeftCircleRS	17	4142	100	94.1	100	100
JumpingDown	14	3952	92.9	7.14	76.5	92.9
jumpingJack (6)	13	5589	100	100	0	100
kickLFront (5)	14	6422	78.6	78.6	0	78.6
kickLSide	26	6063	76.9	88.5	92.9	88.5
kickRFront	15	6728	100	86.7	53.9	86.7
kickRSide	15	7020	93.3	100	80.0	66.7
punchLFront	15	5924	80.0	73.3	67.7	86.7
punchLSide	15	5324	86.7	66.7	53.3	26.7
punchRFront (4)	15	6450	93.3	86.7	60.0	73.3
punchRSide	14	5140	85.7	85.7	28.6	78.6
rotateArmsBBack	16	5111	100	100	100	100
rotateArmsLBack	16	1725	93.8	93.8	43.8	100
rotateArmsRBack	16	1685	100	56.3	43.8	100
sitDownChair (2)	20	6377	90.0	70.0	100	90
sitDownFloor (3)	20	8154	95.0	100	80.0	100
sitDownKnTS(10)	17	10978	100	100	100	100
sitDownTable	20	5411	85.0	60.0	35.0	85.0
skierLstart	30	4240	100	100	90.0	100
squat (8)	13	7619	100	100	100	100
staircaseDownRS	15	3338	100	100	86.7	100
standUpKnTS (9)	17	3094	100	100	82.4	100
standUpSitChair	20	5919	90.0	85.0	100	100
standUpSitFloor	20	8060	90.0	100	95.0	95.0
standUpSitTable	20	5000	85.0	65.0	30.0	70.0
throwBasketball	14	5710	78.6	92.9	0	78.6
throwSitHighR	14	4192	100	100	78.6	100
throwStandingLR	14	4957	100	85.7	0	100
turnLeft	30	5882	76.7	43.3	40.0	80.0
turnRight	30	5908	93.3	86.7	70.0	86.7
walkLstart	31	4818	96.8	93.6	83.9	96.8
walkRightCrossF	16	5369	100	100	93.8	100
Average			92.7	86.5	66.9	91.1

Luo, J., Wang, W., & Qi, H. (2014). Spatio-Temporal Feature Extraction and Representation for RGB-D Human Action Recognition. *Pattern Recognition Letters*. doi:10.1016/j.patrec.2014.03.024

Summary

Advantages

- Difference between motions can be observed directly by visual comparison
- Interesting approach combining known technologies to solve challenging problem
- Potential for scenario independent solution
- Sub motion and repetitive action recognition using NN
- Quite robust and toletant to various lengths (even 50x resized images still obtain similar precision)

Disadvantages

- No solution for segmentation
- Not suitable for online action recognition
- Computationally and time demanding computing of image descriptors (order of minutes)

Future Work

- Action recognition based on segmentation
- Motion classfication using Convolutional Neural Network trained on subset of motion images or better Convolutional Neural Network trained on MOCAP data
- Comparison with DTW approach (centered, rotated, normalized poses)
- Optimize the speed of feature extraction – Caffe descriptor is a current bottleneck

Sources

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., & Bajcsy, R. (2014). Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, *25*(1), 24–38. doi:10.1016/j.jvcir.2013.04.007

Poppe, R., Van Der Zee, S., Heylen, D. K. J., & Taylor, P. J. (2014). AMAB: Automated measurement and analysis of body motion. *Behavior Research Methods*, *46*, 625–33. doi:10.3758/s13428-013-0398-y

Chen, X., & Koskela, M. (2013). Classification of RGB-D and Motion Capture Sequences Using Extreme Learning Machine. *Image Analysis*, 640–651. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-38886-6_60

Luo, J., Wang, W., & Qi, H. (2014). Spatio-Temporal Feature Extraction and Representation for RGB-D Human Action Recognition. *Pattern Recognition Letters*. doi:10.1016/j.patrec.2014.03.024

Vieira, A. W., Nascimento, E. R., Oliveira, G. L., Liu, Z., & Campos, M. F. M. (2012). STOP: Space-Time Occupancy Patterns for 3D action recognition from depth map sequences. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7441 LNCS, pp. 252–259). doi:10.1007/978-3-642-33275-3_31

Evangelidis, G., Singh, G., & Horaud, R. (2014). Skeletal Quads : Human Action Recognition Using Joint Quadruples. doi:10.1109/ICPR.2014.772

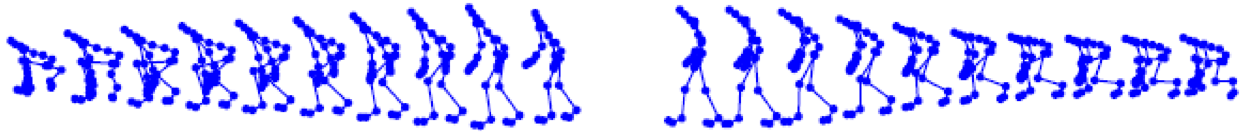
Hussein, M. E., Torki, M., Gowayyed, M. a., & El-Saban, M. (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. *IJCAI International Joint Conference on Artificial Intelligence*, 2466–2472.

Our approach formally

We denote the k -th body pose in a motion sequence of length T as a vector:

$p^t = (j_1^t, \dots, j_n^t)$ with t $\in \{1, \dots, T\}$

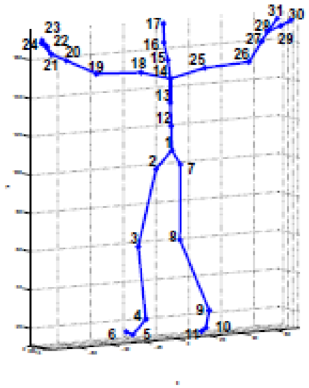
for a recording with T frames.



Each component j_i^t of the vector p^t corresponds to

a joint $i \in \{1, \dots, n\}$ position measurement,

and is denoted by a triplet (x, y, z).



Our approach formally (2)

Let

$$s = \{p^{t_1}, p^{t_2}, \dots p^{t_T}\}$$

be a sequence of poses constituting some motion and let

$$img_{\gamma(\alpha,\beta)} \epsilon \{\alpha \times \beta \times \gamma(\alpha, \beta)\} \ \alpha \in \{1, \dots, maxWidth\}, \beta \in \{1, \dots, maxHeight\}, \gamma \in RGB$$

be an image of size $maxWidth \times maxHeight$ and $\gamma(\alpha, \beta)$ is an information how to color pixel at position (α, β).

Finally we seek to find a transformation function

$$\varphi\colon\ |TIME| \times\ |JOINTS| \times |\mathbb{R}^3| \rightarrow |\mathbb{N}^2| \times |RGB|$$

Such that

$$\varphi(s) = img_{\gamma(\alpha,\beta)}$$