Occlusion-tolerant and personalized 3D human pose estimation in RGB images

Ammar Qammaz, Antonis A. Argyros

Institute of Computer Science,
Foundation for Research and Technology – Hellas (FORTH)

AND

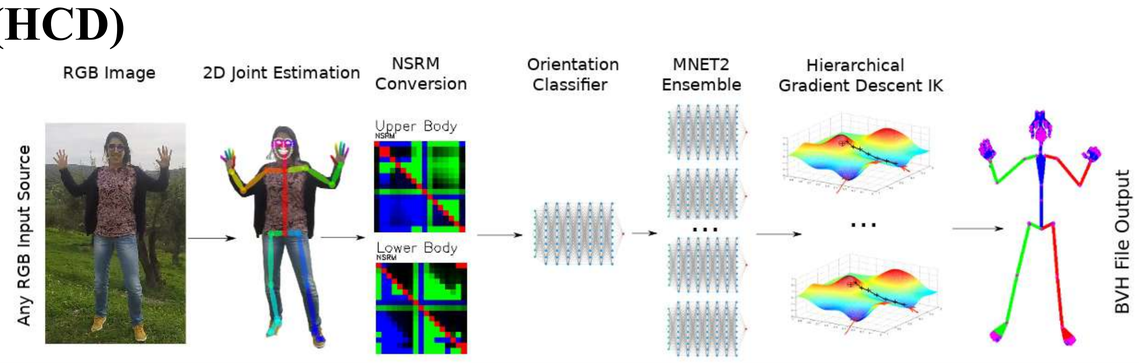
Department of Computer Science,
University of Crete – Hellas

OVERVIEW

We use neural network ensembles that we:

· train using a novel 2D skeleton descriptor we name **Normalized Signed Rotation Matrix (NSRM)**

· personalize and fine-tune their results using a novel optimization algorithm **Hierarchical Coordinate Descent (HCD)**



MAIN IDEAS

- **NSRM** skeleton descriptor encodes relative joint angles while being rotation invariant when aligned to a pivot point.
- We train on the CMU motion capture dataset after perturbing/filtering it.
- Our categorical cross-entropy classifier allows partitioning pose space in 4 view groups to simplify the estimation task of the neural networks.
- Bodies are split in upper and lower body hierarchies to make encoder estimations robust to heavy occlusions.
- 87 d.o.f. problem treated with conditionally independent encoders.
- **HCD** algorithm allows online body personalization without retraining the NN and performs very fast due to its parallelization potential.

ADVANTAGES

- **RGB** input makes our method applicable to most camera systems.
- **BVH** output makes method **plug and play** with popular 3D software.
- **Realtime** CPU only operation @ 70 fps for 2D joints → 3D pose.
- **33%** accuracy improvement on H36M-BP1 compared to baseline[1].

QUALITATIVE EXPERIMENTAL RESULTS

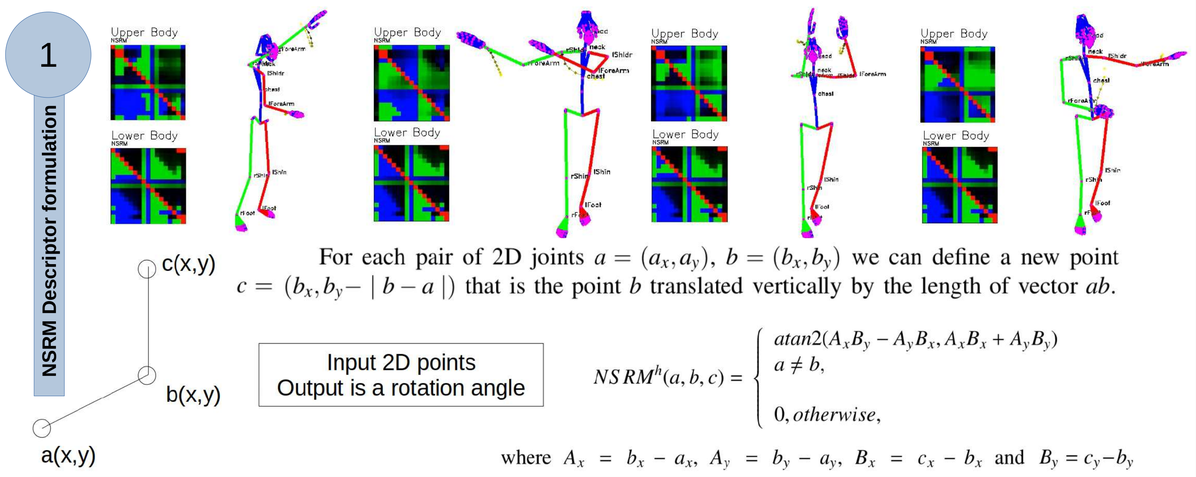
Our method can recover a great variety of poses when tested on the Leeds Sports Dataset.

Our method (green) greatly improves pose quality compared to the baseline method [1] (red).

[1] - A. Qammaz and A.A. Argyros, "MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images", In British Machine Vision Conference (BMVC 2019), BMVA, Cardiff, UK, September 2019.

METHOD OUTLINE

1 NSRM Descriptor formulation

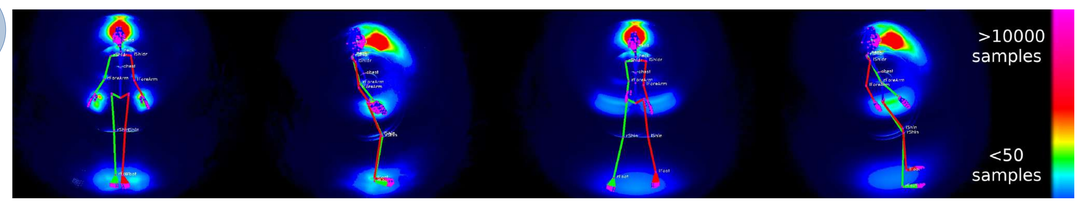


For each pair of 2D joints $a = (a_x, a_y)$, $b = (b_x, b_y)$ we can define a new point $c = (b_x, b_y - | b - a |)$ that is the point b translated vertically by the length of vector ab.

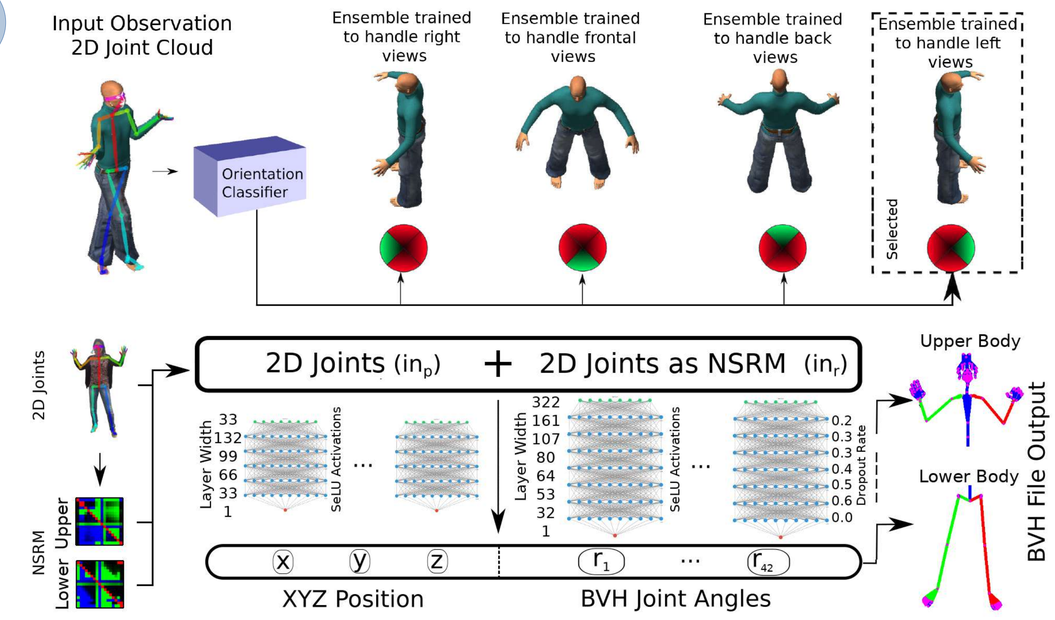
$$NSRM^h(a,b,c) = \begin{cases} atan2(A_x B_y - A_y B_x, A_x B_x + A_y B_y) \\ a \neq b, \\ 0, otherwise, \end{cases}$$

where $A_x = b_x - a_x$, $A_y = b_y - a_y$, $B_x = c_x - b_x$ and $B_y = c_y - b_y$

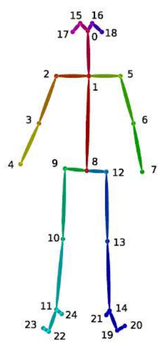
2 Training Data



3 Neural Network ensemble — Orientation Classifier and Neural



4 HCD optimization overview



Chain 1 hips shoulders neck
Chain 2 abdomen neck shoulders
Chain 3 right arm
Chain 4 left arm
Chain 5 right leg
Chain 6 left leg

Group 1
Group 2

We consider :

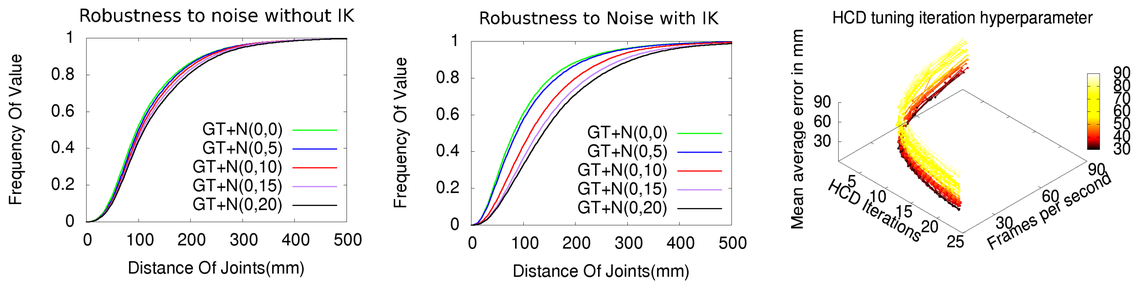
objective function E_{2DJ} that quantifies the mean squared error (MSE) of the 2D joints projected, compared to the 2D joints observed.

$$E_{J2D}(h,o) = \frac{1}{m_J} \sum_{i=1}^{m_J} |j_i^h - j_i^o|^2.$$

QUANTITATIVE EXPERIMENTAL RESULTS

Input	Dir	Dis	Eat	Gre	Pho	Pos	Pur	Sit	Smo	Pho	Wai	Wal	Dog	WaT	Sit.	**Avg**
Ours (NN+HCD)	69	78	92	78	100	79	134	141	97	89	84	85	102	81	165	**108**
Ours (NN only)	88	105	116	99	120	102	152	165	127	116	114	112	146	98	180	**122**
MocapNET [1]	135	140	145	143	153	137	174	215	156	150	151	156	166	134	246	**160**

Comparison of our method to baseline approach [1] with respect to MPJPE metric. Methods are trained on CMU and tested using H36M Blind protocol 1. Numbers in mm.



Proposed method accuracy with (left) and without (middle) the HCD module for various levels of Gaussian noise on H36M. Right: varying HCD iteration hyperparameter reveals a performance/accuracy sweet-spot at 5 HCD iterations.

FORTH INSTITUTE OF COMPUTER SCIENCE

For more information, **http://ammar.gr/mocapnet/**
or contact {ammarkov, argyros}@ics.forth.gr

GitHub YouTube

We acknowledge the **NVIDIA** Corporation for the donation of a Quadro P6000 GPU.
This work was partially supported by **EU H2020** project **Co4Robots** (Grant No 731869)