Backing Off: Hierarchical Decomposition of Activity for 3D Novel Pose Recovery

John Darby[1]
j.darby@mmu.ac.uk

Baihua Li[1]
b.li@mmu.ac.uk

Nicholas Costen[1]
n.costen@mmu.ac.uk

David Fleet[2]
fleet@cs.toronto.edu

Neil Lawrence[3]
neill@cs.man.ac.uk

[1] Department of Computing and Mathematics
Manchester Metropolitan University

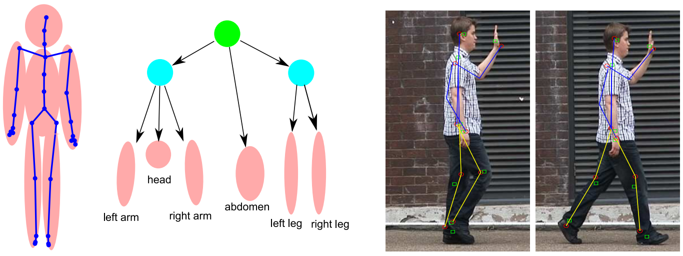
[2] Department of Computer Science
University of Toronto

[3] School of Computer Science
University of Manchester

In model-based analysis-by-synthesis approaches to pose estimation, pose candidates are synthesised using a geometric body model which is used for comparison against observation data [4]. As even simple models of the human body contain around 30 degrees of freedom, the synthesis step generally involves the exploration of a high-dimensional state space *e.g.* [1]. In order to constrain this search task a low-dimensional activity model is often learned from training data *e.g.* [5, 6]. Although such approaches show some capacity to generalise to intra-activity variations in style [6], when the activities to be tracked deviate significantly from those in the training data the global models are unable to cope and pose estimation fails *e.g.* [5].

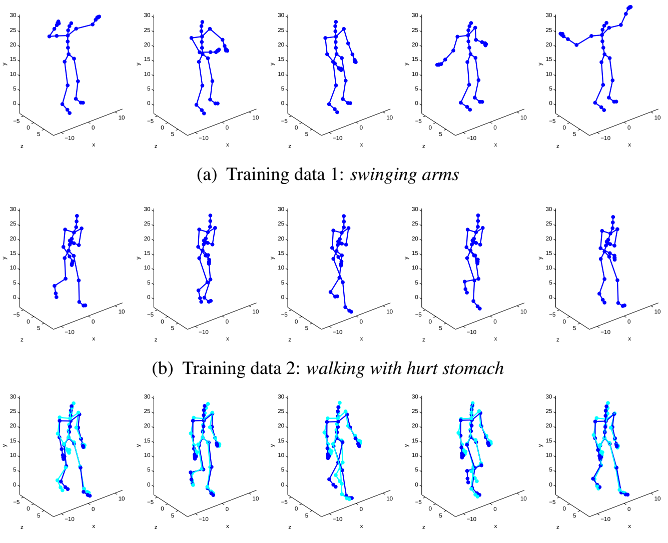
This paper proposes that for effective 3D pose estimation, some capacity to relax the constraints of these full-body models and exploit conditional independencies in the kinematic tree is desirable. We show that with a learned hierarchical model of body coordination for multiple activities, one can recover novel poses that comprise aspects of different activities. The motivation is well captured by the following pose estimation problem. Given training data for *(i)* a person walking and *(ii)* a person standing and waving, how can we construct and explore a model that can describe a person walking *whilst* waving? To address this problem we adopt the hierarchical Gaussian process latent variable model (H-GPLVM) [3] for activity modelling.

The Gaussian process latent variable model (GP-LVM) [2] represents high-dimensional data through a low-dimensional latent model, and a non-linear Gaussian process (GP) mapping from the latent space to the data space. This makes it ideal for the representation of human motion. The H-GPLVM [3] is a form of GP-LVM with a hierarchical latent representation (see Fig. 2(a)). The leaves of the latent model comprise a latent model for each limb or distinct body part; *i.e.*, each node is a GP-LVM for a single body part. To capture the natural coordination of body parts one can then model the joint distribution over latent positions in leaf nodes with a GP from a parent latent variable. For example, in Fig. 2(a) the *left leg* and *right leg* are coordinated by the *lower body latent variable*. Given a lower-body latent position, there is a GP mapping to latent positions for the left and right legs, from which there are GP mappings to the joint angles of the two legs.



(a) (b)

Figure 2: Hierarchical decomposition of skeleton (a) and *walking whilst waving* poses reconstructed from *walking* and *waving* training data (b).



(a) Training data 1: *swinging arms*

(b) Training data 2: *walking with hurt stomach*

(c) Backing off through the hierarchical activity model allows the recovery of novel poses.

Figure 1: MoCap training data (a,b) and resulting pose estimation results for a *walking* sequence found by searching in an H-GPLVM (c).

Lawrence and Moore [3] suggest that a “back off” method inspired by language modelling might be used for the recovery of poses not featured in the H-GPLVM’s training set. The idea is to descend the hierarchy and search nodes at the next level *independently*; this concept forms the basis for inference in this work. By shifting search down one level in the hierarchy we can gradually relax the level of coordination between body parts. While we may be unable to recover a novel test pose by inspection of full-body training poses at the root node, we may be able to fit the observations better by backing off to optimise the abdomen, upper body and lower body independently.

Given the non-linear form of the model, and the potential for ambiguity in pose estimation (*i.e.*, for multi-modality), we formulate pose estimation using a form of Monte Carlo inference. For efficiency, given the dimensions of the latent space and the pose space, we advocate the use of the annealed particle filter [1] with the H-GPLVM. In particular, we use a form of coarse-to-fine search, descending through the model from rough full-body pose estimates at the top level nodes of the model to the eventual refinement of partial pose parameters for each limb in the leaf nodes. The flexibility and accuracy of the approach is demonstrated with the recovery of novel 3D poses from 3D MoCap data (see Fig. 1), and by estimating 3D human pose from 2D monocular data (see Fig. 2(b)).

[1] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005.

[2] N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *JMLR*, 6:1783–1816, 2005.

[3] N. D. Lawrence and A. J. Moore. Hierarchical Gaussian process latent variable models. In *ICML*, 2007.

[4] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2): 90–126, 2006.

[5] C. Sminchisescu and A. Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *ICML*, pages 759–766, 2004.

[6] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *CVPR*, pages 238–245, 2006.