Hierarchical Approach for Articulated 3D Pose-Estimation and Tracking (Extended Abstract)

Leonid Sigal
Brown University
Providence, RI 02912
ls@cs.brown.edu

Michael J. Black
Brown University
Providence, RI 02912
black@cs.brown.edu

1. Introduction

In the recent years we presented a number of methods for a fully automatic pose estimation [5, 7] and tracking [6] of human bodies in 2D [5] and 3D [6]. Initialization and failure recovery in these methods are facilitated by the use of loose-limbed body model [7] in which limbs are connected via learned probabilistic constraints. The pose estimation and tracking can then be formulated as an inference in a loopy graphical model and approximate belief propagation can be used to estimate the pose of the body at each time-step. Each node in the graphical model represents the position and orientation of the limb, and the directed edges between nodes represent statistical dependencies between limbs.

There are a number of significant advantages of this paradigm as compared to the more traditional methods for tracking human motion. Most traditional models of the body resort to the kinematic tree-based representations in 2D, 2.5D, or 3D leading to a high-dimensional search space. Searching for a body pose in this high dimensional space is impractical, and so most tracking methods rely on manual initialization or a canonical starting pose. Additionally, they often exploit strong priors characterizing the motions present, to speed up the search. The lack of automatic initialization from an arbitrary pose also makes it hard to recover from transient failures that often occur during tracking.

While the full body pose may be hard to recover directly, the location and pose of a sub-set of individual (visible) limbs is often much easier to compute. Many good head detectors exist and limb detectors based on the skin color, shading, and focus have been developed. This observation is what drives forth the loose-limbed body model paradigm, initially introduced in [7]. Here we would like to address the loose-limbed body model within the Bayesian hierarchical framework for 3D pose estimation and tracking from a monocular image sequence recently developed in [4] and [5].

2. Hierarchical inference framework

Recent work on 2D body pose estimation and tracking treats the body as a "cardboard person" in which the limbs are represented by 2D planar patches connected by joints. Such models are lower-dimensional than the full 3D model and recent work has shown that they can be estimated from 2D images [2, 3]. The results are typically noisy and imprecise but they provide exactly the kind of information necessary to generate *proposals* for the probabilistic inference of 3D human pose. Thus we simplify the 3D problem by introducing an intermediate 2D estimation stage.

To infer 2D body pose we adopt a generative bottom-up process. Simple body part detectors provide noisy probabilistic proposals for the location and 2D pose (orientation and foreshortening) of visible limbs (Fig. 1 **(b)**). To estimate the pose of the limbs we exploit the idea of a 2D loose-limbed body model [6]. We use a variant of non-parametric belief propagation (NBP) to infer probability distributions representing the belief in the 2D pose of each limb (Fig. 1 **(c)**). The inference algorithm also introduces hidden binary occlusion variables, and marginalizes over them to account for occlusion relationships between body parts. The conditional distributions linking 2D body parts are learned from ground truth data.

This process provides reasonable guesses for 2D body pose from which to estimate 3D pose. Sminchisescu et al [8] learned a probabilistic mapping from 2D silhouettes to 3D pose using a Mixture of Experts (MoE) discriminative model. We generalize their approach to learn a mapping from 2D poses (including joint angles and foreshortening information) to 3D poses. The approach uses a mixture of regularized linear regression models that are trained from a set of 2D-3D pose pairs obtained from motion capture data. Sampling from this model provides predicted 3D poses (Fig. 1 **(d)**), that are appropriate as proposals for a Bayesian temporal inference process (Fig. 1 **(e)**). Our multi-stage approach overcomes many of the problems inherent in inferring 3D pose directly from image features.

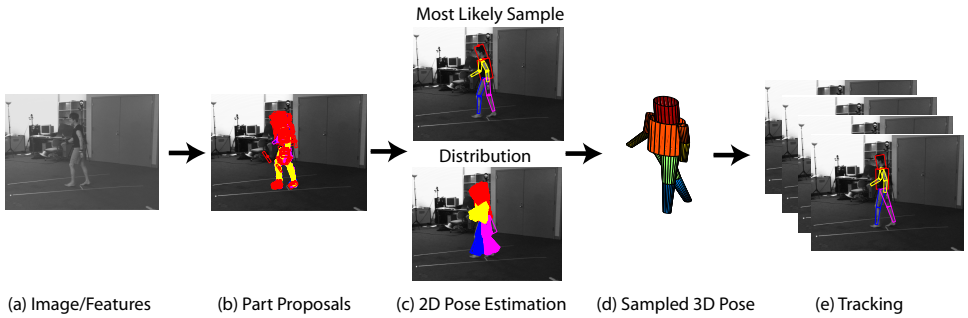

Figure 1. **Example of the hierarchical inference process. (a)** monocular input image with bottom up limb proposals overlaid **(b)**; **(c)** distribution over 2D limb poses computed using non-parametric belief propagation; **(d)** sample of a 3D body pose generated from the 2D pose; **(e)** illustration of tracking.

3. Related Work

Our approach for 2D pose-estimation, that is based on loose-limbed body model, is similar in spirit to the dynamic programming (DP) methods for estimating pictorial structures but has a number of advantages. Unlike [2] we do not need to run limb detectors at a dense, discretized set of joint positions and orientations. Rather we work with a much sparser set of detections and allow NBP to solve for continuous valued joint parameters. The pictorial structures approaches also tend to have problems with multiple parts explaining the same image regions, leading to sub-optimal MAP estimates. This is due to the inherent assumption that a global likelihood can be decomposed into a set of independent local terms. Alternatively, our approach [5] allows for long-range interactions that are required for proper occlusion reasoning; consequently these long-range interactions create loops in the graph and disallow the use of DP methods that only work for tree-based graphical models.

The inference of the 3D body pose from intermediate features has received a good deal of attention with a variety of machine learning methods being employed [1, 8]. These previous approaches have focused on directly inferring 3D pose from 2D silhouettes which may be difficult to obtain. In general, silhouettes contain less information than our 2D models which represent all the limbs, the joint angles, and foreshortening. This helps reduce the ambiguities found in matching silhouette to 3D models but does not remove ambiguities altogether.

4. Conclusions

Bottom-up generative approaches such as those in [5] and [6] provide an attractive paradigm for articulated pose-estimation and tracking. Their current limitations however lie in the inference algorithms (such as NBP) that are while tractable, are still computationally expensive and are inadmissible for real-time or close to real-time performance on current hardware. We believe that there is still work to be done in improving efficiency of these inference algorithms, as well as in developing novel algorithms for inference in real-valued graphical models.

Discriminative approaches [1, 8] tend to be faster, but also tend to generalize worse when it comes to test data that does not conform to the statistics of the data on which these discriminative models have been initially trained. This is a real problem since most discriminative approaches thus far have resulted to learning from perfect silhouette data (due to the lack of real data), but are faced with inferring the body pose from an incomplete and often eroded silhouettes.

We believe that there is something to be gained by combining the discriminative and generative approaches. The hierarchical framework for 3D pose estimation discussed here is a step in that direction. We believe that similar to the object recognition community, where there is currently much focus on combining generative and discriminative models for better overall performance, there is need in more forthcoming work that tries to combine these two schools of approaches for pose estimation and tracking.

References

[1] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. *CVPR*, vol. 2, pp. 882–888, 2004.

[2] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, (61)1:55-79, Jan. 2005.

[3] D. Ramanan and D. Forsyth. Finding and tracking people from the bottom up. *CVPR*, vol. 2, pp. 467–474, 2003.

[4] L. Sigal and M. J. Black. Predicting 3D People from 2D Pictures. *in submission*.

[5] L. Sigal and M. J. Black. Measure Locally, Reason Globally: Occlusion-sensitive Articulated Pose Estimation. *CVPR*, 2006.

[6] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. *CVPR*, vol. 1, pp. 421–428, 2004.

[7] L. Sigal, M. I. Isard, B. H. Sigelman, and M. J. Black. Attractive people: Assembling loose-limbed models using non-parametric belief propagation. *NIPS*, 2003.

[8] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. *CVPR*, 2005, San Diego.