Combining Sample-Based and Analytical Density Propagation for Monocular Tracking

Luc Van Gool[1,2] Tobias Jaeggli[1] Esther Koller-Meier[1]

[1]Swiss Federal Institute of Technology (ETH), D-ITET / BIWI, Zürich, Switzerland

[2]Katholieke Universiteit Leuven ESAT / VISICS, Leuven, Belgium

{*vangool, jaeggli*}*@vision.ee.ethz.ch*

1. Introduction

For image based body pose estimation, the relationship of body pose and image appearance has often been captured by the means of a geometric body model. In contrast to such handcrafted models, a learned model has many advantages; e.g. it can be learned from training data, and the computations can often be performed parametrically. Furthermore, learned probabilistic models are able to generalise over irrelevant variation such as the difference in the appearance of distinct subjects. The mapping from an image descriptor computed from the bounding box of the tracked person to its pose can be learned with regressors and therefore be computed analytically. However, when we consider the 2d bounding box tracking as an integral part of the body tracking and pose estimation problem, the learned regressors do not provide us with the necessary information, nor will it be possible to do all the computations parametrically. To overcome these problems, we propose to learn a model of the joint pdf of pose and image descriptors rather than the conditional, and, for inference we combine analytical and sample-based computation in a Rao-Blackwellised particle filter.

2. Approach

2.1. Learning the joint pdf

The joint pdf $p(\mathbf{x}, \mathbf{y})$ of appearance descriptors $\mathbf{y}$ and pose descriptors $\mathbf{x}$ is approximated by a mixture of Gaussians. While a single multivariate Gaussian captures linear dependencies between the variables, a mixture is able to approximate nonlinearities as well as multimodalities that occur due to ambiguities in monocular tracking. Since the rotation of the tracked person around her own axis is an obviously nonlinear transformation, several Gaussian components will be needed to capture the dependencies. We therefore chose to facilitate the learning process by assigning the Gaussian components of the model to certain view-angles. The overall model is thus learned as a mixture of view-specific models, that each is a mixture of Gaussians itself. An EM algorithm is used to learn the models from training data that are uniformly distributed over all view-angles.

2.2. Inference

Given an observed image descriptor and the learned model, we can directly infer a pdf over the pose by computing the conditional $p(\mathbf{x}|\mathbf{y}_{obs})$. On the other hand, our model also provides us with the information $p(\mathbf{y}_{obs})$ how likely the observation itself is; this will be needed to find the best bounding box, i.e. the one which is most likely to contain a human. This said, a straightforward strategy would be to use the well-proved formulation of Bayesian tracking in combination with a particle filter. This amounts to sampling in the full high-dimensional pose and bounding box search space, and evaluating the image likelihood $p(\mathbf{y}^i|\mathbf{x}^i)$ for each sample using the learned model. ($\mathbf{y}^i$ is the image descriptor computed at the bounding box location encoded by the ith sample). This approach is however inefficient for the high dimensional pose parametrisations that are typically used in tracking. Furthermore it doesn't exploit the fact that part of the problem can be solved analytically.

Rao-Blackwellised particle filter. In the Rao-Blackwellised particle filter (RBPF), the state space is partitioned into a part that is solved using a particle filter, and a part that is solved parametrically using the learned model. Each particle will consist of a sample for bounding box location $\mathbf{l}_t$ and view-orientation α_t, and a weight w_t^i. Additionally a parametric pdf over the possible body poses $p(\mathbf{x}_t|\mathbf{y}_{1:t}, \mathbf{l}_{1:t}^i, \alpha_{1:t}^i)$ is computed for each hypothesised bounding box and orientation. When computing this pdf, only those components of the learned model that

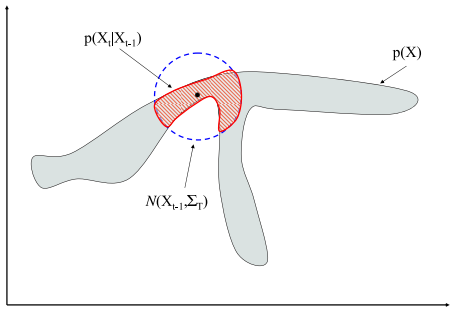

Figure 1. a) The overall prior (red, hatched) is defined as the product of the motion model (blue, dashed, for this illustration a Gaussian pdf around the state of $t-1$) and the static learned prior $p(\mathbf{X})$.

are compatible with the hypothesised orientation α_t have to be considered, all other components will have a very low weight in the mixture and can thus be ignored. This advantage follows from assigning model components to view-orientations, as described in section 2.1. The computation of $p(\mathbf{x}_t|\mathbf{y}_{1:t}, \mathbf{l}^i_{1:t}, \alpha^i_{1:t})$ using the bayesian tracking formulation requires the definition of a temporal prior that includes both a motion model and learned information about likely body poses. We used a combination of Brownian motion around a linear prediction of the state, and the learned (static) prior $p(\mathbf{x})$, as illustrated in fig. 1. For details concerning RBPF, we refer to [1].

3. Experiments

The experiments were done using an image descriptor that is based on the silhouette of the tracked subject, obtained by background subtraction. To encode these segmented images using a descriptor of moderate size, we use signed distance functions, that assign to each point on a regular grid a signed value indicating the distance to the closest point on the silhouette. The dimensionality of the descriptor is further reduced by PCA. The pose descriptor uses 3d joint locations of ankle, hip, shoulder, elbow, wrist and the head location, and PCA dimensionality reduction. The training silhouettes were synthetically generated from motion capture of human locomotion. The algorithm was tested using real and synthetic data with groundtruth.

References

[1] K. Murphy and S. Russel. Rao-blackwellized particle filtering for dynamic bayesian networks. *In A. Doucet, N. de Freitas adn N. Gordon, editors,* Sequential Monte Carlo Methods in Practice, *pp 499-515, Springer*, 2001. 2

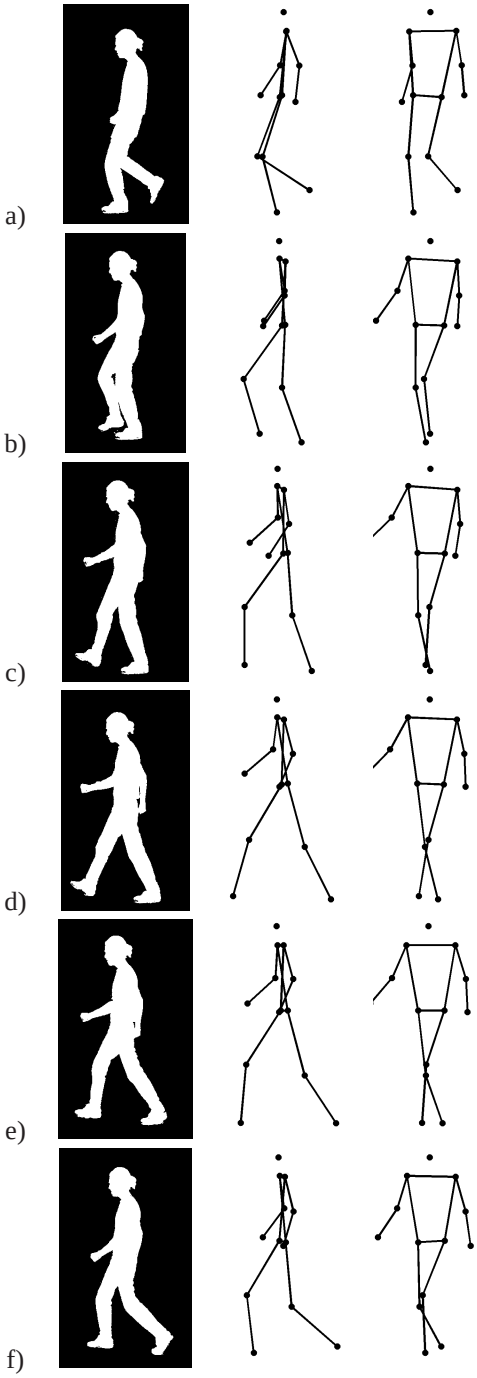

Figure 2. Tracking through a real sequence. For each of the selected frames, the left column shows the tracked bounding box. The other columns show the estimated pose from side view resp. 45 degrees. To visualise a single pose per frame, we chose the mean of the component with the highest weight from the GMM that corresponds to the sample with the highest weight of the sample set.