Nonparametric Density Estimation for Human Pose Tracking

Thomas Brox[1], Bodo Rosenhahn[2], Uwe Kersting[3], and Daniel Cremers[1]

[1] CVPR Group, University of Bonn
Römerstr. 164, 53117 Bonn, Germany
{brox,dcremers}@cs.uni-bonn.de

[2] MPI for Computer Science,
Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
rosenhahn@mpi-inf.mpg.de

[3] Department of Sport and Exercise Science
The University of Auckland, New Zealand

Abstract

The present paper considers the supplement of prior knowledge about joint angle configurations in the scope of 3-D human pose tracking. Training samples obtained from an industrial marker based tracking system are used for a nonparametric Parzen density estimation in the 12-dimensional joint configuration space. These learned probability densities constrain the image-driven joint angle estimates by drawing solutions towards familiar configurations. This prevents the method from producing unrealistic pose estimates due to unreliable image cues. Experiments on sequences with a human leg model reveal a considerably increased stability, particularly in the presence of disturbed images and occlusions.

1 Introduction

This paper is concerned with the task of human pose tracking, also known as motion capturing (MoCap). It is a subtopic of pose tracking where the object/body consists of multiple parts, i.e. limbs, constrained by a kinematic chain [2]. The goal of pose estimation then is to determine the 3-D rigid body motion as well as the joint angles in the kinematic chain.

There are basically two ways to approach the problem. In the *discriminative* approach, one extracts some basic features from the image(s), the raw pixels in the simplest case, and directly learns a mapping from these observed features to the set of pose parameters from a large set of training data. Hence, the method does not care about the meaning of intermediate states, but solely acts as kind of a black box that yields a certain output given a certain input. A recent representative of the discriminative approach is the work in [1].

The *generative* approach, on the other hand, is model based, i.e., there is a more or less detailed object model that, for a given pose, can approximately generate the images that are seen by the camera. The pose parameters are optimized in such a way that the model optimally explains the images.

This paper builds upon a generative approach presented in [3] for rigid bodies with a free-form surface model given. The technique has been extended in [10] to kinematic chains. It determines the pose parameters by matching the projected surface to the object contours in the images. These contours are extracted by assuming a local Gaussian distribution of the object and background region and taking the projected surface model as shape prior into account.

Generative approaches like the one in [3, 10] can be described by Bayesian inference:

$$p(\chi, C|I) = \frac{p(I|C,\chi)p(C|\chi)p(\chi)}{p(I)} \tag{1}$$

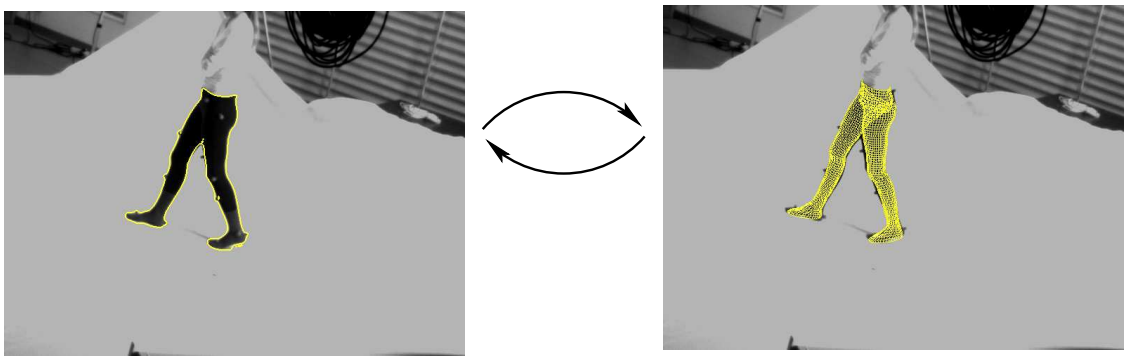


Figure 1: The MoCap system in [10]: **Left:** The object contours are extracted in the input images (just one frame is shown). **Right:** These are used for pose estimation. The pose result is applied as shape prior for the segmentation process and the process is iterated. The goal of the present paper is to extend this system to situations with heavily disturbed image data by supplementing a prior on the pose configuration.

where χ denotes the sought pose parameters, I the input image, and C the object contour that is obtained together with the pose parameters. The technique in [10] uses a prior on the contour by means of $p(C|\chi)$, yet $p(\chi)$ has been ignored by assuming a uniform prior.
The goal of the present paper is to integrate such prior knowledge about the probability of pose configurations into generative approaches like the one in [10]. This is achieved by learning a probability density from training samples. To cope with the non-Gaussian nature of the configuration space, we suggest the approximation of the density by a nonparametric kernel density estimate. Such density estimates have been used in computer vision in the context of image segmentation [6, 12, 5] and shape priors [4].
While learning from training samples is a prerequisite in many discriminative approaches [13, 1] and well-known in the context of shape priors [17, 18, 10, 4], there is very few work with regard to prior knowledge in the context of 3-D generative models apart from the introduction of hard constraints such as explicit joint angle limits or prevention of self-intersections [16]. In [15] it has been suggested to learn a Gaussian mixture in a previously reduced space. Like the nonparametric density estimates suggested here, this work aims at capturing the complex, non-Gaussian configuration space of human pose.
Our experiments with a leg model having 6+12 degrees of freedom show a considerably increasing robustness when the prior is involved. Particularly in cases where the images yield few and unreliable information due to occlusion or noise, the prior helps to keep the result close to familiar pose configurations.

Paper organization. In the next section, we briefly review the technique described in [3, 10] which yields the image-driven part of the pose estimates. After that, Section 3 introduces the modeling of the pose prior, motivates the choice of a kernel density estimate, and demonstrates the integration of the prior into the numerical optimization scheme. In Section 4, we show the effect of the prior and compare the quality of the results to pose estimates obtained with an industrial marker based tracking system. The paper is concluded by a brief summary.

2 Image-driven Pose Tracking

2.1 Pose and Joints

To represent rigid body motions, we use the exponential form,

$$\boldsymbol{M} \quad = \quad \exp(\theta\hat{\xi}) = \exp\begin{pmatrix} \hat{\omega} & \boldsymbol{v} \\ 0_{3\times 1} & 0 \end{pmatrix}. \tag{2}$$

The matrix $\theta\hat{\xi}$ in the exponent is called a *twist*, which consists of two components, a 3×3 matrix $\hat{\omega}$ and a 3-D vector $\boldsymbol{v}$. The matrix $\hat{\omega}$ is restricted to be skew-symmetric, which means $\hat{\omega} \in so(3)$,

with $so(3) = \{\boldsymbol{A} \in \mathbb{R}^{3\times3} | \boldsymbol{A} = -\boldsymbol{A}^T\}$. The exponent of such a twist results in a rigid body motion [7], which is given as a screw motion with respect to a velocity θ. It is common to represent the components of a twist as a 6-D vector $\xi = (\omega_1, \omega_2, \omega_3, \boldsymbol{v})^T$. Twists have two advantages: firstly, they can easily be linearized and used in a fixed point iteration scheme for pose estimation [11]. Secondly, restricted screws (with no pitch component) can be employed to model joints. A kinematic chain is modeled as the consecutive evaluation of such exponential functions, i.e., a point at an endeffector, transformed by a rigid body motion is given as

$$X_i' = \exp(\theta\hat{\xi})(\exp(\theta_1\hat{\xi_1}) \dots \exp(\theta_n\hat{\xi_n}))X_i \tag{3}$$

For abbreviation, we will in the remainder of this paper note a pose configuration by the $(6+n)$-D vector $\chi = (\xi, \theta_1, \dots, \theta_n) = (\xi, \Theta)$ consisting of the 6 degrees of freedom for the rigid body motion ξ and the joint angles Θ. During optimization there is need to generate a transformation matrix from a twist and, vice-versa, to extract a twist from a given matrix. Both can be done efficiently by applying the Rodriguez formula, see [7] for details.

2.2 Model

Coupled extraction of the object contours and registration of the model to these contours can be described by minimization of an energy functional that contains both the pose parameters χ and the object contour as unknowns [3]:

$$\begin{aligned} E(\chi, \Phi) \quad = \quad & -\int_\Omega H(\Phi)\log p_1 + (1 - H(\Phi))\log p_2 \, dx \\ & + \nu \int_\Omega |\nabla H(\Phi)| \, dx + \lambda \int_\Omega (\Phi - \Phi_0(\chi))^2 \, dx. \end{aligned} \tag{4}$$

The contour is represented as the zero-level line of a level set function $\Phi : \Omega \to \mathbb{R}$, such that one can access the interior and exterior of the object region via the step function $H(s)$. Object and background are described by the probability densities p_1 and p_2, respectively. They are modeled by local Gaussian densities, as described in [3]. Hence, minimizing the first two terms yields a contour that maximizes the total a-posteriori probability of all pixel assignments.
The two remaining terms constitute a prior for the contour. The first term seeks to minimize the length of the contour. The second one depends on the pose parameters and seeks to draw the contour close to the projected surface model $\Phi_0(\chi)$. Vice-versa, this term relates the pose parameters to the image data by matching the surface model to the extracted contour, and thereby to the raw pixels. It is a generative model, since given the pose parameters one can use the projected surface Φ_0 and the region densities p_1 and p_2 to generate a simplified version of the image. The tuning parameters $\nu = 1.5$ and $\lambda = 0.05$ have been kept fixed in our experiments.
For $M > 1$ camera views, which are calibrated with respect to the same world coordinate system, the energy functional can easily be extended to M views by minimizing the joint energy

$$E(\chi, \Phi_1, ..., \Phi_M) = \sum_{i=1}^{M} E(\chi, \Phi_i). \tag{5}$$

Whereas the densities $(p_1)_i$ and $(p_2)_i$ are independent for each image, the contour extraction is coupled via the pose parameters χ that influence the contours due to the shape prior.

2.3 Optimization Scheme

We minimize energy (5) by alternating an optimization of the contours for fixed pose parameters and an update of the pose parameters for fixed contours. Keeping the pose parameters fixed yields the gradient descent

$$\partial_t \Phi_i = H'(\Phi_i)\left(\log\frac{(p_1)_i}{(p_2)_i} + \nu \operatorname{div}\left(\frac{\nabla\Phi_i}{|\nabla\Phi_i|}\right)\right) + 2\lambda\,(\Phi_0(\chi) - \Phi_i). \tag{6}$$

Obversely, by keeping the contours fixed, one can derive point correspondences between contour and surface points via shape matching. From the point correspondences, a nonlinear system

of equations can be formulated using the twist representation and Clifford algebra. Each point correspondence contributes three equations of rank 2. For details we refer to [10].
The nonlinear system can be solved with a fixed point iteration scheme. Linearizing the equations yields an over-determined linear system of equations, which can be solved with the Householder method in the sense of least squares. Updating the nonlinear system with the new estimates and linearizing again leads to a new linear system. The process is iterated until convergence.

3 Constraining the Pose by Kernel Density Estimates

The energy functional from the previous section can be motivated from a probabilistic point of view by considering the a-posteriori probability

$$p(\chi, \Phi|I) \propto p(I|\Phi)p(\Phi|\chi)p(\chi). \tag{7}$$

Maximizing this probability is equivalent to minimizing its negative logarithm, which leads to the energy in (4) plus an additional term that constrains the pose to familiar configurations:

$$E_{\text{Prior}} = -\log(p(\chi)). \tag{8}$$

As we want the prior to be independent from the translation and rotation of the body in the training sequences, we apply a uniform prior to the parameters ξ of the rigid body motion. The remaining probability density for the joint angle configuration $p(\Theta)$ is supposed to be learned from a set of training samples.
Fig. 2 visualizes the training data consisting of MoCap data from two walking sequences obtained by a marker based tracking system with a total of 480 samples. Only a projection to three dimensions (the three joint angles of the right hip) of the actually 12-dimensional space is shown. There are many possibilities to model probability densities from such training samples. The most common way is a parametric representation by means of a Gaussian density, which is fully described by the mean and covariance matrix of the training samples. Such representations, however, tend to oversimplify the sample data. Having, for instance, two training samples with the left leg in front and the right leg in back, and vice-versa, a Gaussian density would yield the highest probability for the configuration with both legs in the middle. Configurations close to the samples, on the other hand, would have a comparatively small probability. Although showing only a projection of the full configuration space, Fig. 2 clearly demonstrates that a walking motion cannot be described accurately by a Gaussian density. In the 12-D space, this becomes even more obvious.

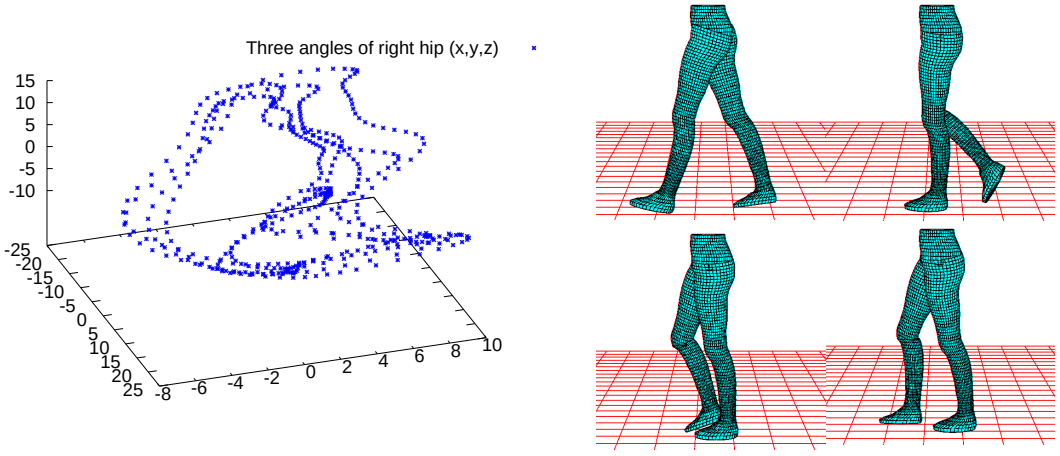


Figure 2: **Left:** Visualization of the training data obtained from two walking sequences. Only a 3-D projection (the three joint angles of the right hip) of the 12-D space is shown. **Right:** Some training samples applied to the body model.

For this reason, we suggest a nonparametric density estimate by means of the Parzen-Rosenblatt estimator [9, 8]. It approximates the probability density by a sum of kernel functions centered at the training samples. A common kernel is the Gaussian function, which leads to:

$$p(\Theta) = \frac{1}{\sqrt{2\pi}\sigma N} \sum_{i=1}^{N} \exp\left(-\frac{(\Theta_i - \Theta)^2}{2\sigma^2}\right) \tag{9}$$

where N is the number of training samples $\Theta_i \in \mathbb{R}^{12}$. This probability density estimator involves the kernel width σ as a tuning parameter. Whereas small kernel sizes lead to an accurate representation of the training data, the estimated density may not generalize well, i.e., unseen test samples may be assigned a too small probability. Large kernel sizes are more conservative, leading to a smoother approximation of the density, which in the extreme case comes down to a uniform distribution. Numerous works on how to optimally choose the kernel size are available in the statistics literature. A detailed discussion can be found in [14]. In our work, we fix σ as the maximum nearest neighbor distance between all training samples, i.e., the next sample is always within one standard deviation. This ensures a smooth approximation between samples while it still keeps the density model flexible.
Note that (9) does not involve a projection but acts on the full 12-dimensional configuration space of the 12-D joint model. This means, also the interdependency between joint angles is taken into account.
The gradient descent of (8) in Θ reads

$$\partial_t \Theta = -\frac{\partial E_{\text{Prior}}}{\partial \Theta} = \frac{\sum_{i=1}^{N} w_i(\Theta_i - \Theta)}{\sigma^2 \sum_{i=1}^{N} w_i} \tag{10}$$

$$w_i := \exp\left(-\frac{|\Theta_i - \Theta|^2}{2\sigma^2}\right). \tag{11}$$

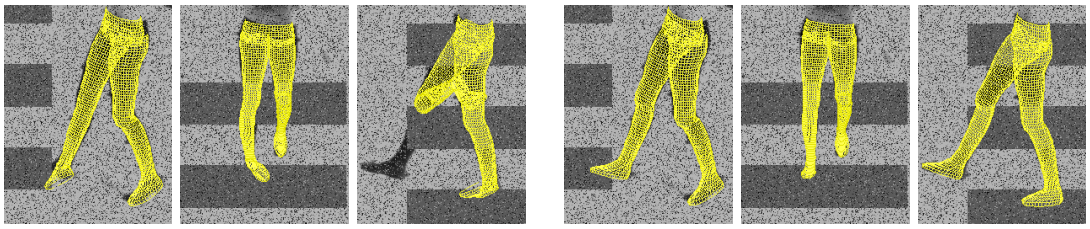
This can be interpreted as the pose configuration being drawn to the next local maximum of the probability density, i.e., the local mode. We integrate this equation into the linear system of the fixed point iteration scheme from Section 2 by appending for each joint j an additional equation $\theta_j^{k+1} = \theta_j^k + \tau \partial_t \theta_j^k$ to the linear system. These equations are weighted by the number of point correspondences in order to achieve an equal weighting between the image- and the prior-driven part. In our experiments, the step size parameter $\tau = 0.125\sigma^2$ yielded stable results.

In contrast to a simple alternation between the image-driven and the prior-driven part, the integration of (10) into the linear system efficiently allows to compensate discrepancies not only locally in the respective joint angle, but globally in all pose parameters including the overall rigid body motion. Therefore, a large discrepancy in the angle of the leg, for instance, can also be compensated by a rotation of the hip.
A second advantage is the implicit regularization of the equation system. Assume a foot is not visible in any camera view. Without prior knowledge, this would automatically lead to a singular system of equations, since there are no correspondences that generate any constraint equation with respect to the joint angles at the foot. Due to the interdependency of the joint angles, the prior equation draws the joint angles of the invisible foot to the most probable solution given the angles of the visible body parts.

4 Experiments

For the experiments we used a four-camera set-up and grabbed image sequences of a female lower torso. The cameras were calibrated using a calibration cube, synchronized via a genlock interface, and we grabbed with 60 frames per second. The person wore a black leg suit (see Fig. 1).
To allow for a quantitative error analysis, we installed parallel to this set-up a second camera ring for a marker based system. Markers were attached to the leg suit and tracked by a commercially available MoCap system[1]. We grabbed a series of sequences and are able to compare our marker-free approach with the marker based system.

[1] We used the Motion Analysis system with 8 Falcon cameras.



Without prior With prior

Figure 3: Relevance of the learned configurations for the tracking stability. Distracting edges from occlusions locally disturb the image-driven pose estimation. This can finally cause a global tracking failure. The prior couples the body parts and seeks the most familiar configuration given *all* the image data.

Fig. 4 and 5 visualize results of a walking sequence in which we replaced 25% of all pixels by a uniform random value. Additionally, we added heavy occlusions of two different types to all camera views. In the first case, box-shaped occlusions of random size and gray value where randomly distributed across the images. In the second case, we added enduring horizontal stripes to the images. For the last quarter of the sequence, the person is not visible in the first camera anymore. All these difficulties frustrate the acquisition of contour data needed for pose estimation.

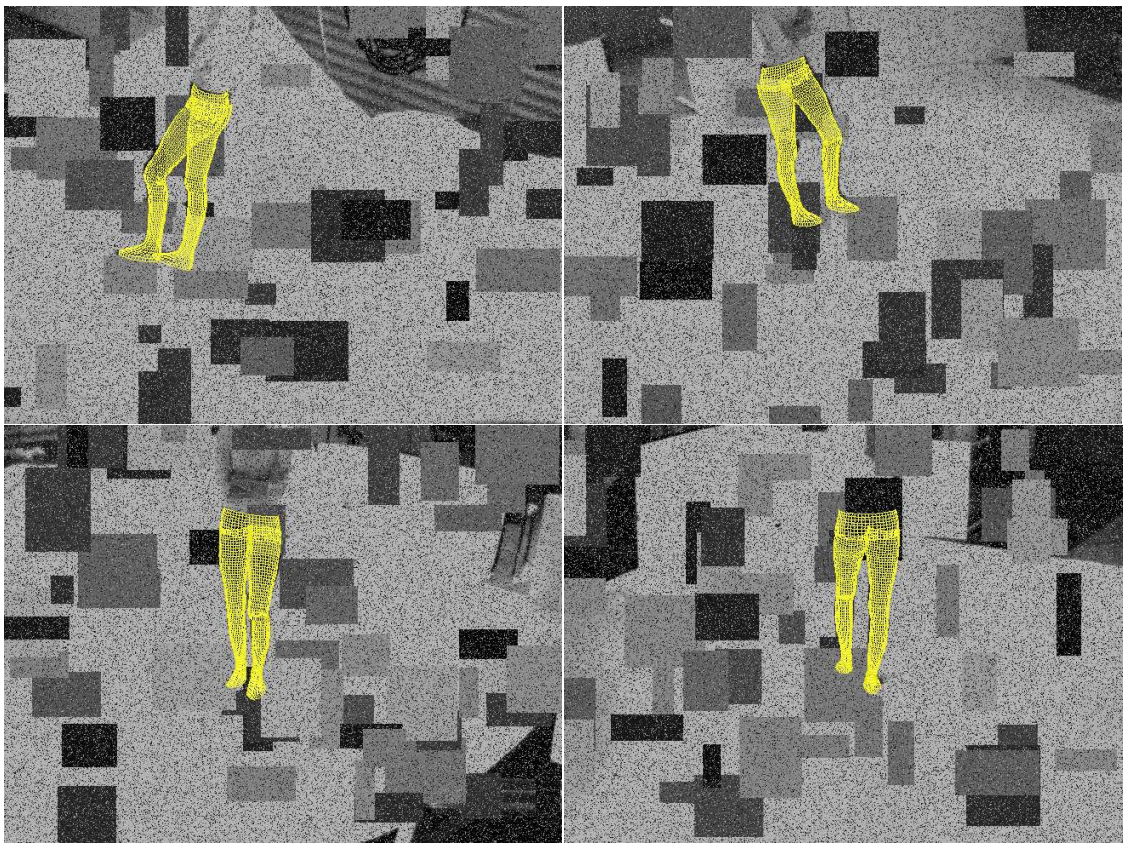


Figure 4: Pose estimates in a sample frame disturbed by 50 varying rectangles with random position, size, and gray value and 25% uncorrelated pixel noise.

Thanks to the joint angle prior, however, the sequence is tracked reliably in both cases despite these disturbances. The training set did *not* contain the test sequence. The diagram in Fig. 5 compares the obtained tracking curves to the marker based result, which can be regarded as ground truth $\pm$ 3 degree (0.05 radians), and the result obtained when the joint angle prior is ignored. Despite the occlusions, the errors are almost within the accuracy of the marker based system. Without the prior, however, tracking fails nearly right from the beginning.

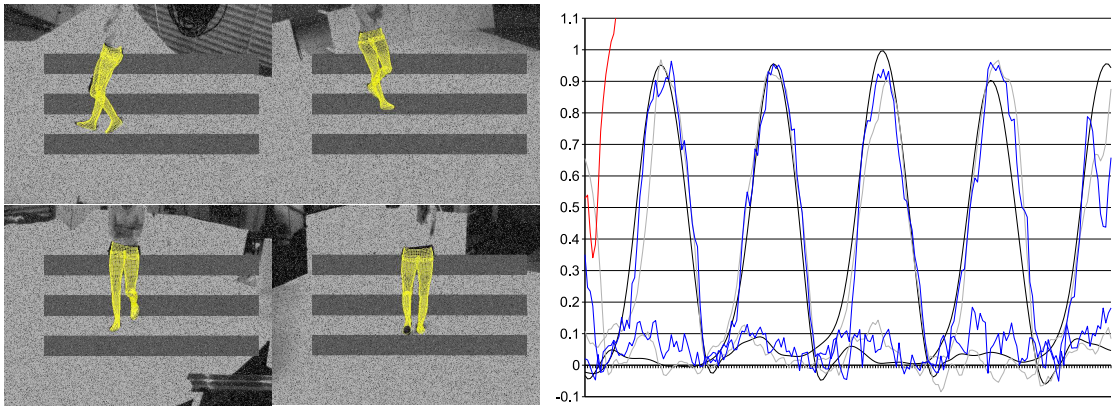


Figure 5: **Left:** Pose estimates in a sample disturbed by three enduring gray bars and 25% uncorrelated pixel noise. **Bottom:** Joint angles in radians of the left and right knee, respectively. **Black:** marker based system. **Gray:** occlusion by permanent bars. **Blue:** occlusion by random rectangles (see figure 4). **Red:** tracking without prior fails after a couple of frames.

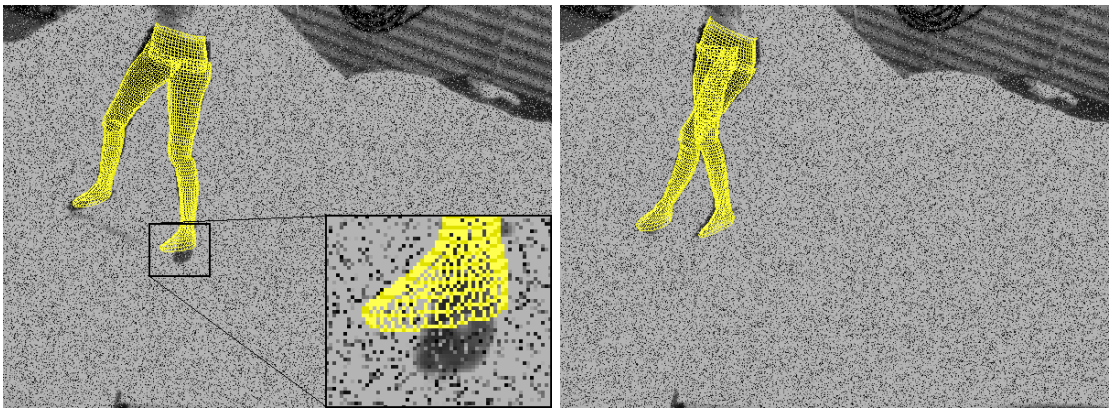


Figure 6: Generalization capabilities of the prior: two frames from a jumping sequence tracked with solely training data from walking sequences available. **Left:** The enlarged part reveals inaccuracies, as the prior prevents the foot angle from further bending. **Right:** However, the prior is able to accurately handle many other configurations not consistent with those of a walking person.

In order to test the generalization capabilities of the Parzen estimator, we further applied the method to a sequence where the person was asked to perform a series of jumping jacks. Again we added 25% uniform noise to the images. As pose configurations of this type of motion pattern were not contained in the training data, one expects problems concerning the accuracy of tracking. Indeed Fig. 6 reveals errors for some frames where the true configuration was too far away from the training samples (enlarged in Fig. 6). Nevertheless, the tracking remains stable and yields sufficiently accurate non-walking configurations. For all experiments we used the same internal parameters.

5 Summary

We have suggested to learn joint angle configurations from training samples via a Parzen density estimator and to integrate this prior via Bayesian inference into a numerical scheme for contour based human pose tracking from multiple views. The learned density draws the solution towards familiar configurations given the available data from the images. In case the image does not provide enough information for a unique solution, the most probable solution according to the prior is preferred. The experimental evaluation demonstrates that this allows to handle situations with seriously disturbed images where tracking without knowledge about reasonable angle configurations is likely to fail.

Acknowledgements

We gratefully acknowledge funding by the DFG project CR250/1 and the Max-Planck Center for visual computing and communication.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, Jan. 2006.

[2] C. Bregler, J. Malik, and K. Pullen. Twist based acquisition and tracking of animal and human kinematics. *International Journal of Computer Vision*, 56(3):179–194, 2004.

[3] T. Brox, B. Rosenhahn, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose estimation. In W. Kropatsch, R. Sablatnig, and A. Hanbury, editors, *Pattern Recognition*, volume 3663 of *LNCS*, pages 109–116. Springer, Aug. 2005.

[4] D. Cremers, S. Osher, and S. Soatto. Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. *International Journal of Computer Vision*. To appear.

[5] T. Kadir and M. Brady. Unsupervised non-parametric region segmentation using level sets. In *Proc. Ninth IEEE International Conference on Computer Vision*, volume 2, pages 1267–1274, 2003.

[6] J. Kim, J. Fisher, A. Yezzi, M. Cetin, and A. Willsky. A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Transactions on Image Processing*, 14(10):1486–1502, 2005.

[7] R. Murray, Z. Li, and S. Sastry. *Mathematical Introduction to Robotic Manipulation.* CRC Press, Baton Rouge, 1994.

[8] E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[9] F. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

[10] B. Rosenhahn, T. Brox, U. Kersting, A. Smith, J. Gurney, and R. Klette. A system for marker-less motion capture. *Künstliche Intelligenz*, (1):45–51, 2006.

[11] B. Rosenhahn and G. Sommer. Pose estimation of free-form objects. In T. Pajdla and J. Matas, editors, *Proc. 8th European Conference on Computer Vision*, volume 3021 of *LNCS*, pages 414–427. Springer, May 2004.

[12] M. Rousson, T. Brox, and R. Deriche. Active unsupervised texture segmentation on a diffusion based feature space. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 699–704, Madison, WI, June 2003.

[13] G. Shakhnarovich, P. Viola, and T. Darell. Fast pose estimation with parameter sensitive hashing. In *Proc. International Conference on Computer Vision*, pages 750–757, Nice, France, Oct. 2003.

[14] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, New York, 1986.

[15] C. Sminchisescu and A. Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *Proc. International Conference on Machine Learning*, 2004.

[16] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22(6):371–391, 2003.

[17] J. Zhang, R. Collins, and Y. Liu. Representation and matching of articulated shapes. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 342–349, June 2004.

[18] J. Zhang, R. Collins, and Y. Liu. Bayesian body localization using mixture of nonlinear shape models. In *Proc. International Conference on Computer Vision*, pages 725–732, Beijing, China, Oct. 2005.