Nonparametric Density Estimation with Adaptive, Anisotropic Kernels for Human Motion Tracking [*]

Thomas Brox[1], Bodo Rosenhahn[2], Daniel Cremers[1], Hans-Peter Seidel[2]

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

[2] Max Planck Center for Visual Computing and Communication
Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany
{rosenhahn,seidel}@mpi-sb.mpg.de

Abstract. In this paper, we suggest to model priors on human motion by means of nonparametric kernel densities. Kernel densities avoid assumptions on the shape of the underlying distribution and let the data speak for themselves. In general, kernel density estimators suffer from the problem known as the curse of dimensionality, i.e., the amount of data required to cover the whole input space grows exponentially with the dimension of this space. In many applications, such as human motion tracking, though, this problem turns out to be less severe, since the relevant data concentrate in a much smaller subspace than the original high-dimensional space. As we demonstrate in this paper, the concentration of human motion data on lower-dimensional manifolds, approves kernel density estimation as a transparent tool that is able to model priors on arbitrary mixtures of human motions. Further, we propose to support the ability of kernel estimators to capture distributions on low-dimensional manifolds by replacing the standard isotropic kernel by an adaptive, anisotropic one.

1 Introduction

In recent years, human tracking has emerged as a vivid research area. In particular 3D human tracking, where one seeks to estimate the pose and joint angles of a 3D human model from 2D images, has attracted a lot of attention [9]. Having applications in surveillance and biomechanics, human tracking also serves as a playground for new machine learning techniques. Due to self-occlusions, inaccurate, corrupted, or missing data, it requires the use of prior knowledge on typical human poses and movements in order to avoid ambiguous solutions. Moreover, the solution space generally comprises multiple locally optimal solutions. This

[*] We thank U. Kersting for providing data from a marker-based system, as well as the German Research Foundation for their financial support.

is a great challenge for optimization algorithms if not being supported by predictions generated from strong priors.

Consequently, the literature provides numerous works on different learning techniques that can be used to exploit prior knowledge for human tracking. These works range from rather simple explicit joint angle limits [19, 6], over static pose priors [17, 2], to priors on motion dynamics [16]. Some recent dynamic models are based on sophisticated nonlinear regression methods including nonlinear dimensionality reduction [5, 20].

Most of these works stick to a maximum a-posteriori (MAP) formulation of the tracking problem. Given the input image I in the current frame and pose configurations in previous frames $\chi_{t-1}, ..., \chi_{t-k}$, one looks for the new configuration χ_t that maximizes

$$p(\chi_t | I, \chi_{t-1}, ..., \chi_{t-k}) \propto p(I|\chi_t) p(\chi_t|\chi_{t-1}, ..., \chi_{t-k}). \tag{1}$$

While the first factor considers how well a solution χ_t explains the image data, the second factor represents the conditional prior probability density of some pose given the poses of previous frames. One can directly model this prior, which leads to regression methods. Usually, such methods comprise a parametric component, which means that they cannot accurately model a prior consisting, for instance, of running and jumping motions, since the parametric model would mix up both motion patterns to yield an (unprecise) mean prediction. In order to handle such cases consisting of multiple motions, one has to employ a mixture of regressors [8, 11, 18], which includes many critical hyperparameters and is quite demanding with regard to optimization.

In this paper, we pursue an alternative strategy. Since $p(a|b) = \frac{p(a,b)}{p(b)}$, and we maximize with respect to a, $p(b)$ can be neglected as a constant factor and we may consider

$$p(\chi_t | I, \chi_{t-1}, ..., \chi_{t-k}) \propto p(I|\chi_t) p(\chi_t, ..., \chi_{t-k}). \tag{2}$$

Here the second factor is the joint prior density of poses in previous frames and the current one. Such an unconditional probability density can be estimated from training samples using a Parzen estimator. Since the density is fully nonparametric, it can easily model arbitrary mixtures of motion patterns. The Parzen estimator only implies the assumption of a locally smooth density. Consequently, it can capture *all* smooth densities provided there are enough training samples.

The input space of human motion is rather high-dimensional. For a reasonable human body model, at least 20 degrees of freedom are needed. Looking only 4 frames into the past, already implies a 100-dimensional space. It is well known that estimating wide-spread densities in such spaces with a typical kernel estimator, would need huge amounts of training data [15]. However, in practice, this problem is often less severe. This is because high-dimensional spaces are often only sparsely populated, i.e., the density to be estimated concentrates on a small subspace, a low-dimensional manifold in the high-dimensional space. In this paper, we demonstrate that in case of human motion tracking, already the standard Parzen estimator can deal with a 121-dimensional space.

Nevertheless, this estimator is not optimal for such high-dimensional spaces. This is due to the fixed isotropic kernel, which does not adapt to the local structure of the subspace. Hence, the Parzen estimator looses predictive power in normal direction to the manifold. This drawback can be circumvented by introducing anisotropy in the estimation process. Therefore, we propose to replace the isotropic kernel of the standard Parzen estimator by adaptive anisotropic kernels. The same concept has been proposed in the context of general density estimation in [14, 21]. Also the work in [4] based on kernel PCA can be interpreted as sort of an anisotropic kernel density estimator. However, the latter has quadratic complexity in the test phase, which is problematic when the number of training samples becomes large.

2 Anisotropic Kernel Density Estimation

Consider some prior knowledge given by a set of training samples $\{x_i | i = 1, ..., N\}$. In order to integrate such knowledge into a Bayesian model, one must estimate a probability density from the samples. In contrast to typical parametric densities, such as a Gaussian density, which are very restricted in the priors they can model, this paper is concerned with nonparametric kernel densities. The classic Parzen-Rosenblatt density estimator employs an isotropic kernel $K(x, x')$ with a fixed width h. Given such a kernel, the estimated density reads [1, 12, 10]:

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} K_h(x, x_i). \tag{3}$$

A very common kernel is the Gaussian kernel

$$K_h(x, x') = \frac{1}{(2\pi h^2)^{\frac{D}{2}}} \exp\left(-\frac{\|x - x'\|^2}{2h^2}\right), \tag{4}$$

where D denotes the dimensionality of the data. This density estimator, though simple, reveals many advantages. Firstly, it can model arbitrary densities and one can show that in the limit, for $N \to \infty$ and $h \to 0$ adequately, the estimator converges to the true density [15]. Secondly, the estimator is very transparent. In contrast to many learning techniques that rely on modeling in an abstract feature space, the Parzen estimator is easily interpretable. Moreover, it contains only a single hyperparameter, the kernel width h, which can be estimated efficiently from the training data via cross-validation or, depending on the application, by even simpler criteria like average nearest neighbor distance. In contrast to Gaussian mixture models or related techniques, there is no need to determine the number of mixture components, which is a difficult non-convex optimization problem.

As mentioned in the introduction, the main weakness of the Parzen estimator appears when it is employed in high-dimensional spaces where the support of the density is located on a low-dimensional manifold. Then it looses predictive power in normal direction to this manifold due to the fixed isotropic kernel. This is the motivation for using adaptive, anisotropic kernels leading to an anisotropic version of the Parzen estimator. Again, the density is a sum of kernels centered at the training samples

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} K_i(x, x_i), \tag{5}$$

where now $K_i(x, x_i)$ is the locally adaptive anisotropic Gaussian kernel

$$K_i(x, x_i) = \frac{1}{|2\pi\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - x_i)^\top \Sigma_i^{-1}(x - x_i)\right). \tag{6}$$

Its window width and preferred direction is defined by the covariance matrix Σ_i. This covariance matrix is computed locally by means of

$$\Sigma_i = \alpha\mathbf{1} + \sum_{j=1}^{N} K_h(x_i, x_j)(x_i - x_j)(x_i - x_j)^\top, \tag{7}$$

where $\alpha\mathbf{1}$ denotes the identity matrix scaled by a regularization parameter α and $K_h(x, x')$ is the isotropic Gaussian kernel stated in (4).

This anisotropic kernel density estimator has several nice properties. Firstly, the absolute width of the kernel is locally adaptive. This allows for smaller windows in areas with many training samples, whereas sparsely populated areas can still be approximated by larger windows. Secondly, the windows have a preferred orientation in which the kernel size is increased. Since the kernel integrates to 1, this effect automatically decreases the kernel size in orthogonal directions. Such an anisotropy is particularly useful to model data on low-dimensional manifolds, as most of the kernel's power is focused on the tangential space of the manifold. In contrast to Gaussian mixture models, there is still no need to determine the number of mixture components. The estimator can be regarded as a degenerate version of a Gaussian mixture, where the number of components equals the number of training samples. Obviously, this also provides an increased accuracy in respect to the density's local structure compared to a Gaussian mixture with only a small number of components.

The density estimator still imposes only two hyperparameters h and α. These hyperparameters can be estimated from the training data via leave-one-out (LOO) cross validation, i.e., one minimizes the following loss function based on Kullback-Leibler divergence

$$E(h, \alpha) = -\log\left(\sum_{i=1}^{N} \hat{p}_{i,h,\alpha}(x_i)\right), \tag{8}$$

where $\hat{p}_{i,h}(x)$ denotes the estimated probability density with parameters h and α when sample i has been removed from the training set. In the application case of human tracking, we found that one can simplify the parameter optimization by setting h to the average nearest neighbor distance of all training samples and

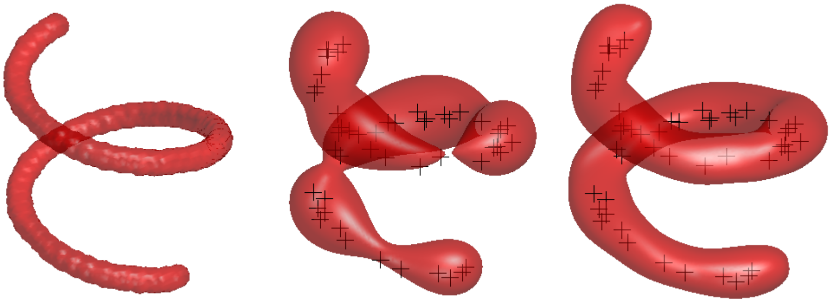

Fig. 1. From left to right: (a) Intrinsically one-dimensional **density** in three-dimensional space. **(b)** Density estimate with the conventional **Parzen** method (ISE $= 12.2 \cdot 10^{-4}$). **(c)** Density estimate with the **anisotropic Parzen** method (ISE $= 9.3 \cdot 10^{-4}$).

$\alpha = \frac{h}{5}$. This is reasonable since training data is obtained via motion capture systems with a fixed frame rate, i.e., samples always come in larger groups.

Figure 1 demonstrates the qualitative difference between the isotropic kernel density estimator and the anisotropic one. Having some data points sampled from the true density (left), the isotropic estimator yields a density estimate that approximates the true density quite well but lacks the ability to interpolate in some of the gaps (middle). In contrast, the anisotropic estimator focuses better on the structure of the density. This is also reflected by the lower integrated mean square error (ISE) between the true and the estimated density.

3 Kernel Densities in Human Tracking

Density estimators can be a valuable component in an application like human motion tracking. In this task, we expect a given surface model consisting of several limbs that are interconnected by predefined joints. The sought pose configuration χ at each frame consists of a global rigid body motion, represented by the six parameters of a twist ξ, as well as a number of joint angles $\Theta = (\theta_1, ..., \theta_M)^\top$. Estimation of these parameters at frame t from image data and poses from previous frames can be regarded in a MAP setting

$$p(\chi_t | I, \chi_{t-1}, ..., \chi_{t-k}) \propto p(I|\chi_t) p(\chi_t, ..., \chi_{t-k}), \qquad (9)$$

where the conditional prior density $p(\chi_t | \chi_{t-1}, ..., \chi_{t-k})$ has already been replaced by the joint density $p(\chi_t, ..., \chi_{t-k})$, as explained in the first section of this paper. The right hand side consists of a data fidelity factor and the prior density of certain sequences of pose configurations.

3.1 Modeling the data fidelity

There are several ways to model the data fidelity, such as keypoint tracking or silhouette constraints. Since this issue is not the focus of this paper, we stick to an existing silhouette based method [13], where (9) is expanded to

$$p(\chi_t, \Phi | I, \chi_{t-1}, ..., \chi_{t-k}) \propto p(I|\Phi) p(\Phi|\chi_t) p(\chi_t, ..., \chi_{t-k}) \tag{10}$$

by introducing the silhouette represented as the zero level of a function $\Phi : \Omega \to \mathbb{R}$. Maximizing the probability in (10) is equivalent to minimizing its negative logarithm. With certain model assumptions on the appearance of the object and background region [13], this yields the energy

$$\begin{aligned} E(\chi_t) = &- \int_\Omega H(\Phi) \log p_1 + (1 - H(\Phi)) \log p_2 \, dx \\ &+ \lambda \int_\Omega (\Phi - \Phi_0(\chi_t))^2 \, dx - \log p(\chi_t, ..., \chi_{t-k}), \end{aligned} \tag{11}$$

where $H(s)$ is the step function that distinguishes the object and background region, p_1 and p_2 are densities of the intensity in these regions, $\Phi_0(\chi_t)$ is the level set function representing the silhouette of the projected model given the pose χ_t, and $\lambda = 0.05$ is a weighting parameter that steers how much the contour Φ may deviate from the model silhouette Φ_0.

In contrast to many tracking works that use a sampling strategy to minimize similar energies as the one in (11), we use a gradient descent in Φ and χ, which yields the next local minimum starting from some initialization χ^0. In the tracking context, finding the next local minimum can be sufficient, especially if the model is supported by a prior density that allows for reasonable predictions of poses in successive frames. For this reason, we now concentrate on the last term in (11), which comprises this prior density.

3.2 Modeling the prior density

For building a prior density, a set of training samples with certain motion patterns is required. A database with a rather large variety of motions is available at Carnegie Mellon University [3]. We used this database to assemble training samples for estimating a prior density.

In order to ensure certain invariance properties, and to keep the required number of samples as well as the dimensionality as small as possible, we arrange the sample vectors in the following way. Firstly, we restrict the degrees of freedom of our model to the 29 most important ones. , i.e., 3 dof at each shoulder, 1 dof at each elbow, 1 dof at each hand, 3 dof at each upper leg, 1 dof at each knee, 2 dof at each foot, and 1 dof at the neck. Together with the global rigid body motion, this yields a total of 29 dof. Further, since we are interested in invariance with regard to the location and orientation of the person, we only consider the joint angles at previous frames, not the global twist. Finally, for keeping the dimensionality small and nonetheless considering configurations that are several frames in the past, the time axis is non-uniformly sampled. In detail, we assemble vectors x_i, where the first six components are the twist parameters representing the rigid body motion between $t-1$ and t. The next $M = 23$ components are the absolute joint angles in t. There follow successively the M joint angles in

$t-1$, in $t-2$, $t-5$, and $t-10$. Similar to the so-called snippets in [7], this yields training vectors x_i of dimension $D = 121$, from which a density can be estimated according to Section 2. In case the frame rate of the input sequence does not match the training sequences, the prior is scaled accordingly.

3.3 Density gradient and pose prediction

For the minimization of (11) we are not interested in the absolute density, but in the local gradient of its logarithm. This gradient corresponding to the anisotropic density estimator in (5) reads:

$$\begin{aligned} K_i(x, x_i) &:= \exp\left(-\frac{1}{2}(x - x_i)^\top \sigma_i^{-1}(x - x_i)\right) \\ \frac{\partial \log p(x)}{\partial x} &= -\frac{1}{2}\frac{\sum_{i=1}^N K(x, x_i)\mathit{\Sigma}_i^{-1}(x - x_i)}{\sum_{i=1}^N K(x, x_i)}. \end{aligned} \tag{12}$$

Note that only the first $6 + M$ components of $\frac{\partial \log p(x)}{\partial x}$ are needed, since the pose at previous frames is fixed. Starting from some point x^0 and ignoring the data fidelity term, gradient ascent will converge to the local mode of the density in the $(6 + M)$D subspace, i.e., the most likely pose configuration in a local neighborhood given the poses at previous frames. In combination with the data fidelity term, the result is the local maximum a-posteriori solution given the image data *and* the prior density.

In some cases, one is indeed interested in the local mode of the density alone, starting from some motion vector x^0. In human tracking, this situation arises when predicting the pose in a successive frame, irrespective the image data, which may be unreliable without a good hypothesis of the pose in the new frame. For prediction, it is beneficial to estimate a density where the absolute joint angles at t in x_i are replaced by relative angles between $t-1$ and t. This prevents predictions far from the tracked motion in case the training data are sparse and rather dissimilar from the tracked motion.

4 Experiments

Our experiments demonstrate that kernel density estimators in general, but in particular the one based on anisotropic kernels, are well suited to model dynamic motion priors in human tracking. Firstly, Figure 2 and a video in the supplementary material show that one can generate an enduring cyclic motion from the anisotropic density. For this motion generation, we simply provide a short sequence of poses. Starting the gradient descent (12) from the last such pose, and ignoring the image-driven part, yields an enduring running motion. This means, from previous poses alone, the density can predict a reasonable succession of poses like a regressor would do.

An important challenge in human tracking are monocular sequences. Since only few limbs are visible in a single view, the problem is generally underconstrained and prior assumptions, such as the suggested prior density, are needed

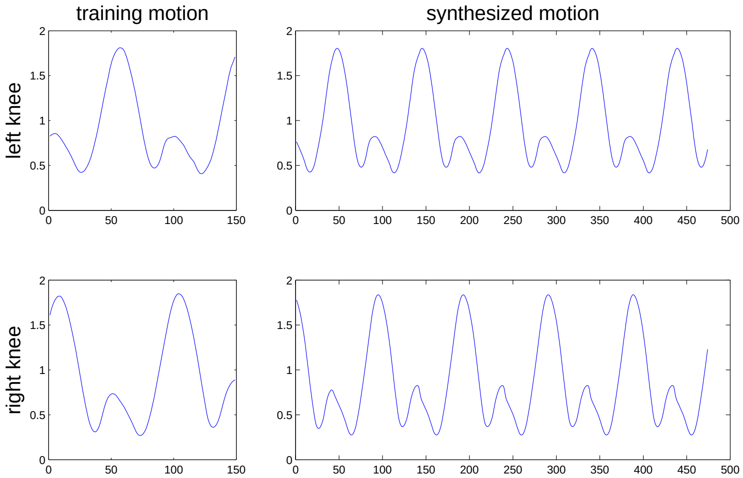

Fig. 2. Synthesis of a cyclic motion from the density estimated from one (or multiple) training motions. Only the left and right knee angles are depicted.

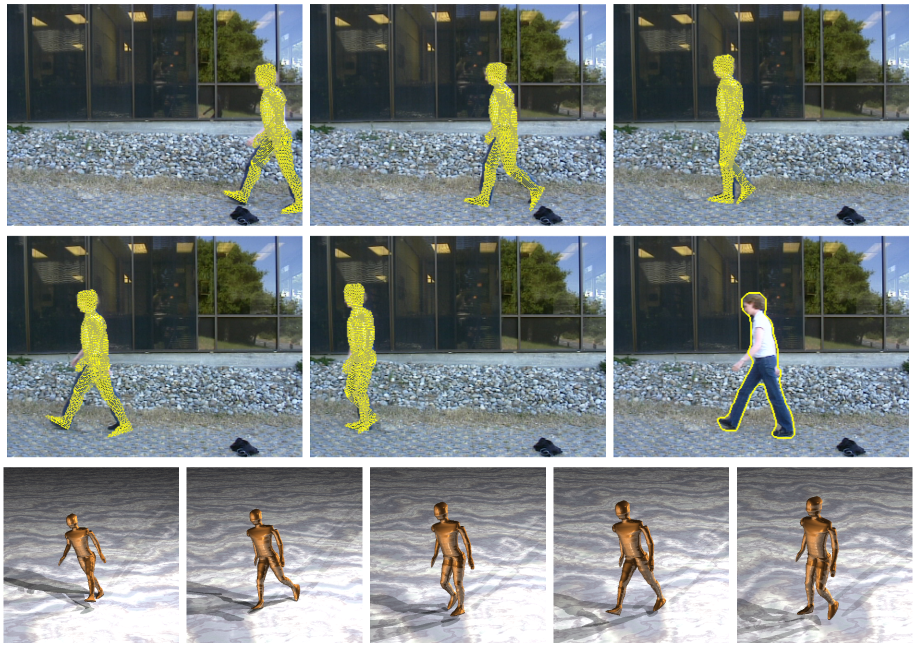

Fig. 3. Tracking result for a monocular sequence, where the density has been estimated with anisotropic kernels. **Center right:** Input image with extracted contour. **Bottom row:** Synthesized view generated from the tracked pose.

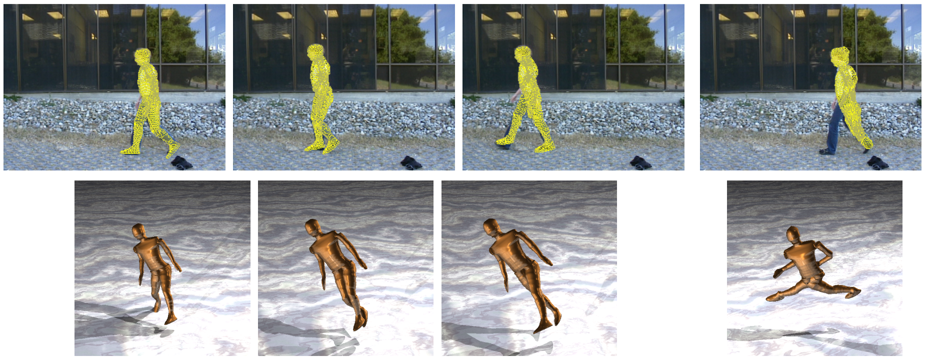

Fig. 4. Top row, from left to right: Tracking result for the sequence in Figure 3, but using the isotropic kernel density estimator. Results are not as good as in the anisotropic case. **Rightmost:** Without any motion prior, tracking fails already after a few frames. **Bottom row:** Synthesized view generated from the tracked pose.

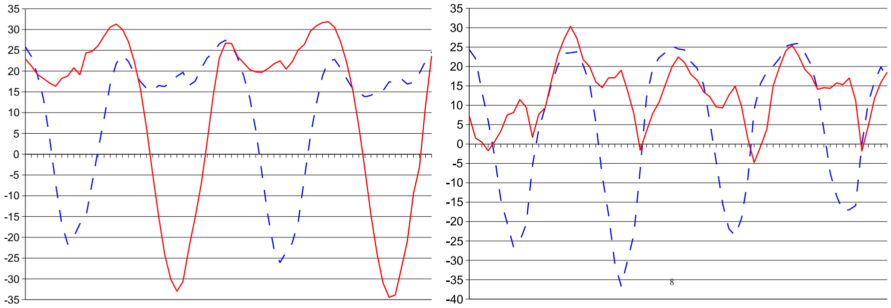

Fig. 5. Tracking curves of the left and right knee. **Left:** Anisotropic kernel density. **Right:** Classic kernel density. The isotropic kernel, due to its weaker predictive power, leads to a hopping motion, as the left and right leg are partially interchanged.

for a unique solution. Figure 3 shows the tracking result for a standard test sequence[1]. The synthesized views confirm that the estimated 3D pose is very accurate thanks to the prior density estimated from two standard and one exaggerated walking sequences. In contrast, the isotropic kernel estimator in Figure 4 yields results that explain the 2D image data very well, but since the density is less distinct, the estimated pose is not as good as with the anisotropic kernel. As Figure 5 and the video in the supplementary material show, the isotropic kernel density estimator partially mixes up the left and the right leg. The rightmost image in Figure 4 shows that replacing the prior density by some static pose prediction fails completely. Due to the the weak prediction, the gradient descent runs into a suboptimal local minimum and not even consistency with the 2D image data can be ensured.

[1] The sequence is available at `www.nada.kth.se/∼hedvig/data.html`

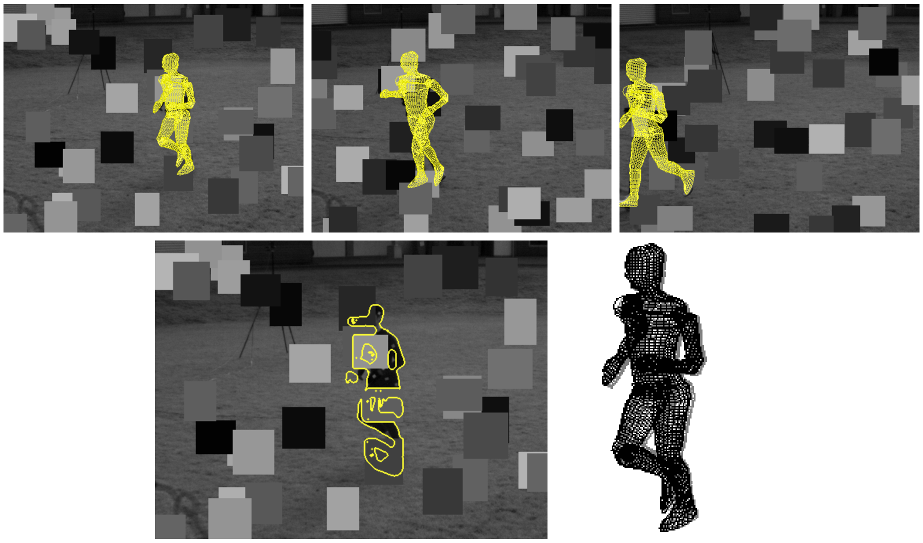

Fig. 6. Tracking of a jogging sequence with 40 occlusions randomly placed in each image. Only one out of four camera views is shown. For the whole sequence, see the supplementary online material. **Top row:** Tracking result. Despite the substantially disturbed image cues (see contour), reasonable poses are computed. **Bottom left:** Contour used for estimating the pose. **Bottom right:** Prediction of the pose in a new frame (black) relative to the previous frame (gray) by means of the prior density.

Partial occlusions are another challenge in human tracking. Figure 6 demonstrates the robustness of the proposed technique in the presence of severely corrupted image data. 40 occluding boxes have been randomly added to the sequence. In contrast to pixel noise, image data is not only missing, but even misleading, since the occluding boxes create false object boundaries like real occlusions. The contour shown in Figure 6 demonstrates this negative effect on the contour extraction. The prior density estimated from 9 different running and jogging motions, which were subsampled to yield a total of 606 points, keeps the solution close to a jogging motion and, hence, allows for successful tracking.

Setting	mean error	std. dev.
0 boxes	4.01°	±3.3°
20 boxes	5.47°	±4.7°
40 boxes	5.71°	±4.5°
additional samples, 40 boxes	6.13°	±4.9°
walking samples only, 20 boxes	39.58°	±35.3°
isotropic kernel density, 0 boxes	3.64°	±2.4°

Table 1. Mean error and standard deviation of knee and elbow joints between tracking results of the jogging sequence in Figure 6 and the outcome of a marker-based tracking system (ground truth).

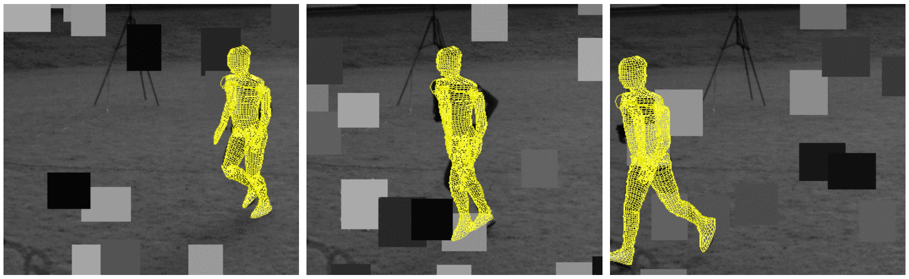

Fig. 7. Tracking of the jogging sequence with 20 occluding boxes and only samples from a walking motion being available for density estimation. The image contains few information on the arms. Hence, the arm pose is hallucinated from the walking prior. The legs, however, are tracked well, despite the unfitting prior.

We also investigated whether the image/prior tandem is able to generalize to sequences where the motion seen in the image does not perfectly fit to any of the prior motions. Figure 7 shows a result where the density has been estimated only from samples of a single walking sequence. Due to poorly constrained image data, the arms reflect the walking motion of the prior. The legs, though, fit well to the jogging motion seen in the image. This shows that the prior can be voted down by clear image data.

For getting a better insight in what happens in the high-dimensional space, Figure 8 depicts on the left the training data consisting of 9 running and jogging motions projected into 2D space via multidimensional scaling. In blue one can see the trajectory of the tracked jogging sequence in this space. Clearly, the pure running prior has a very simple structure. In contrast, the bottom figure shows the situation when additional motions are added to the prior. Learning such priors is a problem for many techniques, especially for typical regressors, which can only model functions. Kernel density estimation, however, handle such situations in a very natural way without any need to adapt the methodology. Hence, tracking the jogging motion in Figure 6 with such more general training data is not a problem.

Table 1 compares several experimental settings of the jogging sequence quantitatively by showing the mean error of the results. Ground truth has been provided by parallel tracking with a marker-based system. Tracking with the more general prior is almost as good as with the special running prior. Interestingly, the isotropic kernel density estimator yields a higher accuracy than the anisotropic density estimator in this sequence. The arm pose in all training patterns does not fit well to the tracked arm motion. Since the anisotropic kernel leads to more concise density estimates, it also tends to a stronger prior. This explains the better result of the isotropic estimator in this case, as we kept the weighting between image and prior data fixed. The large error of the result with the walking prior emerges from the large impact of the wrong elbow angles. For the knee joints alone, one obtains an average error of only 5.29°.

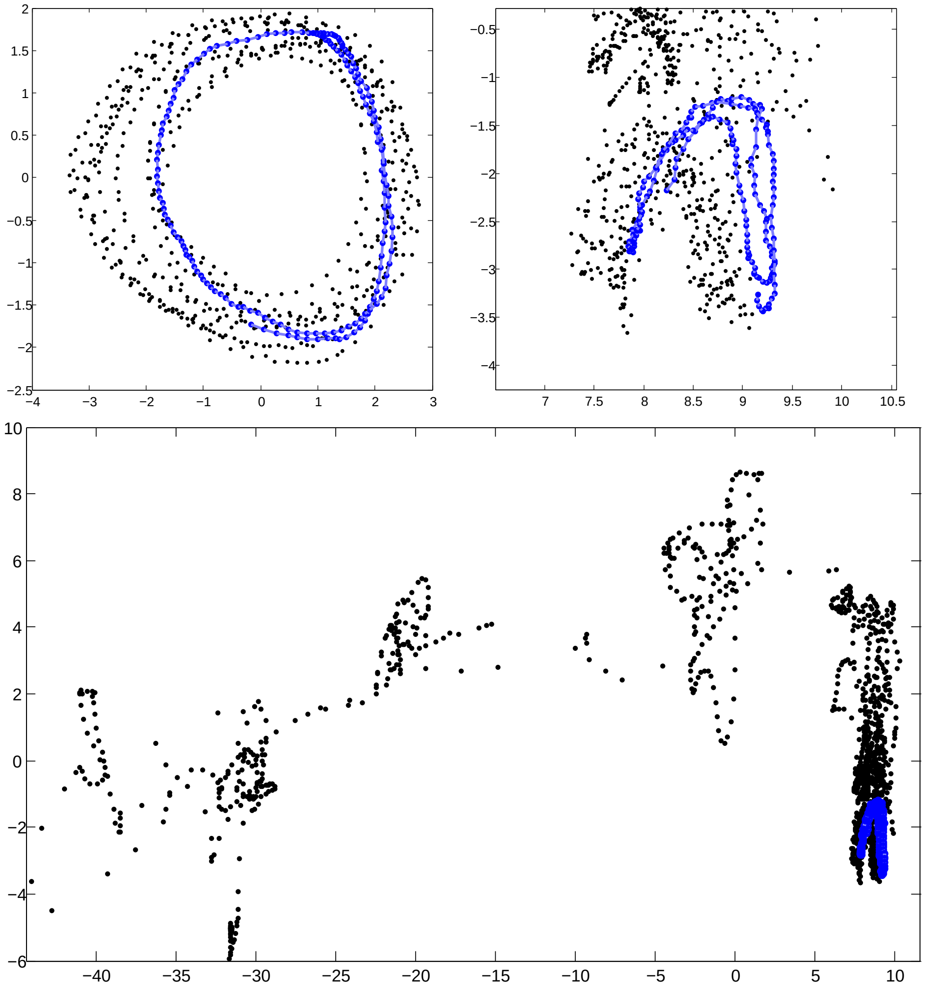

Fig. 8. Training samples (black dots) and tracked pose (blue dots) for the jogging sequence. Points have been projected to 2D via multidimensional scaling (MDS). **Top Left:** Training set consists only of running and jogging motions. **Bottom:** Other motions, such as walking, jumps, leaps, cardwheels, flips, and break dance, have been added to the training set. The sample distribution becomes very irregular in the 2D projection. **Top Right:** Zoom into the part of the more general prior that is relevant for tracking the jogging motion.

Finally, Figure 9 shows a highly dynamic handspring sequence. Without the ability to predict the rough pose at a successive frame, such a motion is hard to track. With the anisotropic kernel density estimate, however, we obtain a rather accurate result.

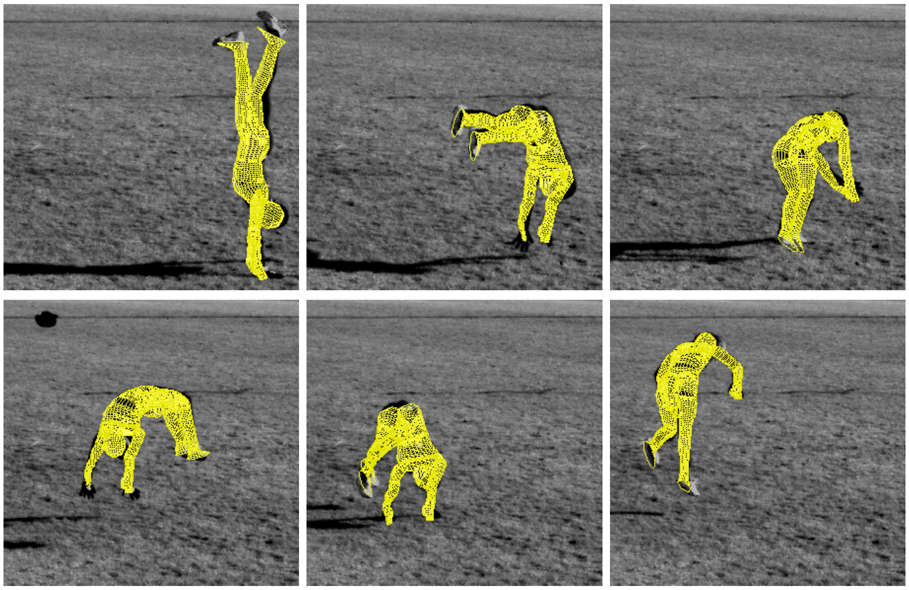

Fig. 9. Tracking of a handspring with four camera views. See the video in the supplementary material for the whole sequence.

5 Conclusions

We have introduced the use of kernel density estimation as a transparent way to model motion priors in human tracking. In order to cope with the high-dimensional nature of the input space, we proposed density estimation with anisotropic kernels. They are especially appropriate when the density concentrates on a low-dimensional subspace. We suggested a Bayesian tracking framework that makes use of such an anisotropic density estimator by combining the prior density with observation probabilities derived from the image data. A broad experimental evaluation showed the main properties of such a tracking technique. In particular, the prior density is able to support the tracking in case of missing or corrupted image data, even when there are only few, mildly fitting training samples. Moreover, as it is a nonparametric technique, it can easily model multiple motions. Future work will concentrate on appropriate data structures that allow for an efficient sublinear computation of densities from many thousand training samples.

References

1. H. Akaike. An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6:127–132, 1954.

2. T. Brox, B. Rosenhahn, U. Kersting, and D. Cremers. Nonparametric density estimation for human pose tracking. In K. Franke et al., editor, *Pattern Recognition*, volume 4174 of *LNCS*, pages 546–555, Berlin, Germany, September 2006. Springer.
3. CMU. Carnegie-Mellon Motion Capture Database. `http://mocap.cs.cmu.edu`.
4. D. Cremers, T. Kohlberger, and C. Schnörr. Shape statistics in kernel space for variational image segmentation. *Pattern Recognition*, 36(9):1929–1943, 2003.
5. K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popović. Style-based inverse kinematics. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*, volume 23, pages 522–531, 2004.
6. L. Herda, R. Urtasun, and P. Fua. Implicit surface joint limits to constrain video-based motion capture. In T. Pajdla and J. Matas, editors, *Proc. European Conference on Computer Vision*, volume 3022 of *LNCS*, pages 405–418, Prague, Czech Republic, May 2004. Springer.
7. N. Howe, M. Leventon, and W. Freeman. Bayesian reconstruction of 3D human motion from single-camera video. In *Proc. Neural Information Processing Systems*, pages 820–826, 2000.
8. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.
9. T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, 2006.
10. E. Parzen. On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.
11. R. Rosales and S. Sclaroff. Learning body pose via specialized maps. In *Proc. Neural Information Processing Systems*, December 2001.
12. F. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.
13. B. Rosenhahn, T. Brox, and J. Weickert. Three-dimensional shape knowledge for joint image segmentation and pose tracking. *International Journal of Computer Vision*, 73(3):243–262, July 2007.
14. S. R. Sain. Multivariate locally adaptive density estimation. *Computational Statistics & Data Analysis*, 39(2):165–186, 2002.
15. D. Scott. *Multivariate Density Estimation.* Wiley, 1992.
16. H. Sidenbladh, M. J. Black, and L. Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proc. European Conference on Computer Vision*, volume 2353 of *LNCS*, pages 784–800. Springer, 2002.
17. C. Sminchisescu and A. Jepson. Generative modeling for continuous non-linearly embedded visual inference. In *Proc. International Conference on Machine Learning*, 2004.
18. C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning joint top-down and bottom-up processes for 3D visual inference. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 1743–1752, 2006.
19. C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22(6):371–391, 2003.
20. R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *Proc. International Conference on Computer Vision and Pattern Recognition*, pages 238–245. IEEE Computer Society Press, 2006.
21. P. Vincent and Y. Bengio. Manifold parzen windows. In *Proc. Neural Information Processing Systems*, volume 15, pages 825–832, 2003.