

Estimating Articulated Human Motion With Covariance Scaled Sampling

Cristian Sminchisescu, Bill Triggs

▶ **To cite this version:**

Cristian Sminchisescu, Bill Triggs. Estimating Articulated Human Motion With Covariance Scaled Sampling. The International Journal of Robotics Research, 2003, 22 (6), pp.371–391. inria-00548242

HAL Id: inria-00548242
https://inria.hal.science/inria-00548242

Submitted on 20 Dec 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Cristian Sminchisescu
Bill Triggs

INRIA Rhône-Alpes
GRAVIR-CNRS
655 avenue de l'Europe
38330 Montbonnot, France
Cristian.Sminchisescu@inrialpes.fr
Bill.Triggs@inrialpes.fr

Estimating Articulated Human Motion With Covariance Scaled Sampling

Abstract

We present a method for recovering three-dimensional (3D) human body motion from monocular video sequences based on a robust image matching metric, incorporation of joint limits and non-self-intersection constraints, and a new sample-and-refine search strategy guided by rescaled cost-function covariances. Monocular 3D body tracking is challenging: besides the difficulty of matching an imperfect, highly flexible, self-occluding model to cluttered image features, realistic body models have at least 30 joint parameters subject to highly nonlinear physical constraints, and at least a third of these degrees of freedom are nearly unobservable in any given monocular image. For image matching we use a carefully designed robust cost metric combining robust optical flow, edge energy, and motion boundaries. The nonlinearities and matching ambiguities make the parameter-space cost surface multimodal, ill-conditioned and highly nonlinear, so searching it is difficult. We discuss the limitations of CONDENSATION-like samplers, and describe a novel hybrid search algorithm that combines inflated-covariance-scaled sampling and robust continuous optimization subject to physical constraints and model priors. Our experiments on challenging monocular sequences show that robust cost modeling, joint and self-intersection constraints, and informed sampling are all essential for reliable monocular 3D motion estimation.

KEY WORDS—3D human body tracking, particle filtering, high-dimensional search, constrained optimization, robust matching

1. Introduction

Extracting three-dimensional (3D) human motion from natural *monocular* video sequences poses difficult modeling and computation problems:

(i) Even a minimal human model is very complex, with at least 30 joint parameters and many more body shape ones, subject to highly nonlinear joint limits and non-self-intersection constraints.

(ii) Matching a complex, imperfectly known, self-occluding model to a cluttered scene is inherently difficult. Typical loose clothing only complicates matters.

(iii) In contrast to simplified two-dimensional (2D) approaches (Cham and Rehg 1999; Ju, Black, and Yacoob 1996) and the multi-camera 3D case (Kakadiaris and Metaxas 1996; Gavrila and Davis 1996; Bregler and Malik 1998; Delamarre and Faugeras 1999; Plankers and Fua 2001; Drummond and Cipolla 2001), the estimation problem is extremely ill-conditioned, with at least one-third of the 30+ degrees of freedom (DOF) remaining very nearly unobservable in any given monocular image. The most important non-observabilities are motions of major body segments in depth (i.e., towards or away from the camera; these account for about one-third of the 3D DOF), but others include rotations of near-cylindrical limbs about their axes, and internal motions of compound joints such as the spine or shoulder that are difficult to observe even with 3D data.

(iv) In addition to being ill-conditioned, the monocular estimation problem is highly multimodal. In particular, for any given set of image projections of the 3D joint centers, there are typically some thousands of possible inverse kinematics solutions for the 3D body configuration.[1] Under any reasonable model-image matching cost metric, each kinematic solution produces a corresponding local minimum in configuration space, and

The International Journal of Robotics Research
Vol. 22, No. 5, May 2003, pp. xxx-xxx,
©2003 Sage Publications

1. For each body segment, for any given depth for its top (innermost) endpoint, the bottom endpoint can be aligned with its image projection either in a "sloped forwards" configuration, or in a "sloped backwards" one. A full body model contains at least ten main body segments, and hence has at least $2^{10} = 1024$ possible inverse kinematics solutions (sets of forwards/backwards segment configurations). See Lee and Chen (1985) and the empirical confirmations in Sminchisescu and Triggs (2002a,b).

correspondence ambiguities only compound this number of minima. In practice, choosing the wrong minimum rapidly leads to mistracking, so reliable tracking requires a powerful multiple hypothesis tracker capable of finding and following a significant number of minima. The development of such a tracker is one of the main contributions of this paper. Some more recent work, not reported here, further enhances tracking reliability by explicitly enumerating the possible kinematic minima (Sminchisescu and Triggs 2003).

Also note that these four difficulties interact strongly in practice. For example, minor modeling or matching errors tend to lead to large compensatory biases in hard-to-estimate depth parameters, which in turn cause mis-prediction and tracking failure. Hence, we believe that a successful monocular 3D body tracking system must pay attention to all of them.

This paper is organized as follows. In Section 1.1 we discuss several existing approaches to human articular tracking, explaining why we believe that they are not suitable for the difficult 3D-from-monocular case and informally motivating our new tracker. In Section 2 we briefly describe our 3D body model, which includes full 3D occlusion prediction, joint angle limits and body non-self-intersection constraints. In Section 3 we discuss our robust model-image matching framework, which combines robust optical flow, edge energy, and motion boundaries. In Section 4 we detail our hybrid search/tracking scheme, which combines a mixture density propagation tracker with carefully shaped cost-sensitive sampling, with robust constraint-respecting local optimization. In Section 5 we briefly describe the local optimization schedule we use to find initial 3D body poses and internal body proportions from model/image joint correspondence input. In Section 7 we detail some experiments on challenging monocular sequences. These illustrate the need for each of robust cost modeling, joint and self-intersection constraints, and well-controlled sampling plus local optimization. We end the paper with discussions of the effect of the sampling regime on search efficiency (Section 7) and approximation accuracy (Section 8), and ideas for future work.

1.1. High-Dimensional Tracking Strategies

Locating good poses in a high-dimensional body configuration space is intrinsically difficult. Three main classes of search strategies exist: **local descent** incrementally improves an existing estimate, e.g., using local Newton strategies to predict good search directions (Bregler and Malik 1998; Rehg and Kanade 1995; Kakadiaris and Metaxas 1996; Wachter and Nagel 1999); **regular sampling** evaluates the cost function at a pre-defined pattern of points in (a slice of) parameter space, e.g., a local rectangular grid (Gavrila and Davis 1996); and **stochastic sampling** generates random sampling points according to some hypothesis distribution encoding "good places to look" (e.g., Deutscher, Blake, and Reid 2000; Sidenbladh, Black, and Fleet 2000). Densely sampling the entire parameter space would in principle guarantee a good solution, but it is infeasible in more than two or three dimensions. In 30 dimensions any feasible sample must be extremely sparse, and hence likely to miss significant cost minima. Local descent does at least find a local minimum, but with multimodality there is no guarantee that the globally most representative ones are found. Whichever method is used, effective focusing is the key to high-dimensional search. This is an active research area (Deutscher, Blake, and Reid 2000; Heap and Hogg 1998; Cham and Rehg 1999; Merwe et al. 2000), but no existing method can guarantee global minima.

During tracking the search method is applied time-recursively, the starting point(s) for the current search being obtained from the results at the previous time step, perhaps according to some noisy dynamical model. To the (often limited) extent that the dynamics and the image matching cost are statistically realistic, Bayes-law propagation of a probability density for the true state is possible. For linearized unimodal dynamics and observation models under least-squares/Gaussian noise, this leads to extended Kalman filtering. For likelihood-weighted random sampling under general multimodal dynamics and observation models, bootstrap filters (Gordon, Salmond, and Smith 1993; Gordon and Salmond 1995) or CONDENSATION (Isard and Blake 1998) result. In either case various model parameters must be tuned and it sometimes happens that physically implausible settings are needed for acceptable performance. In particular, to control mistracking caused by correspondence errors, the selection of slightly incorrect inverse kinematics solutions, and similar model identification errors, visual trackers often require exaggerated levels of dynamical noise. The problem is that even quite minor errors can pull the state estimate a substantial distance from its true value, especially if they persist over several time steps. Recovering from such an error requires a state space jump greater than any that a realistic random dynamics is likely to provide, whereas using an exaggeratedly noisy dynamics provides an easily controllable degree of local randomization that often allows the mistracked estimate to jump back onto the right track. Boosting the dynamical noise does have the side effect of reducing the information propagated from past observations, and hence increasing the local uncertainty associated with each mode. But this is a small penalty to pay for reliable tracking lock, and in any case the loss of accuracy is often minor in visual tracking, where weak dynamical models (i.e., short integration times; most of the state information comes from current observations and dynamical details are unimportant) are common.

In summary, in multimodal problems, sample-based Bayesian trackers often get trapped into following incorrect local minima, and some form of explicit local (but not *too* local) search must be included to rescue them. For trackers operating in this "memoryless step and search" regime, the machinery of Bayes-law propagation is superfluous—the

dynamical model is not correct in any case—and it is simpler to think in terms of sequential local search rather than tracking and noisy dynamics. It seems that many, if not most, existing Bayesian trackers in vision operate essentially in this regime, and the current paper is no exception. Hence, we will assume only weak zeroth-order dynamical models and use the language of search rather than tracking. But this is largely a matter of terminology, and more elaborate dynamical models are trivial to incorporate if desired.

Many existing human trackers silently inflate the dynamical noise as a local search mechanism (e.g., Cham and Rehg 1999; Heap and Hogg 1998; Deutscher, Blake, and Reid 2000). But in each of these papers, it is only one component of the overall search strategy. The randomization provided by noise inflation is an effective search strategy only for relatively low-dimensional problems, where the samples can cover the surrounding neighborhood fairly densely. In high dimensions, volume increases very rapidly with radius, so any sample that is spread widely enough to reach nearby minima must necessarily be extremely sparse. Hence, the samples are most unlikely to hit the small core of low cost values surrounding another minimum. If they fall into its basin of attraction at all, they are much more likely do so at a high cost point, simply because high cost points are far more common. This is fatal for CONDENSATION-style weighted resampling; high cost points are very unlikely to be resampled, so the new minimum is almost certain to be missed even though an independent track started at the sample would eventually condense to the minimum. The moral is that, in high dimensions, random sampling alone does not suffice; some form of local optimization of the samples, or at least a delayed decision about whether they are viable or not, is essential to prevent mistracking. Cham and Rehg (1999), Heap and Hogg (1998) and the current work use explicit descent-based local optimization for this, while Deutscher, Blake, and Reid (2000) use a simulated annealing-like process (which is usually less efficient, although better aligned with the point-based sample-and-evaluate philosophy of pure particle tracking).

The 3D-from-monocular problem has characteristic ill-conditioning associated with depth DOF, whereas transverse DOF are directly observable and hence relatively well conditioned. It also has large numbers of kinematic local minima related by motions in depth, in addition to the minima in transversal directions produced by correspondence ambiguities. Hence, we would like to ensure a thorough, perhaps even a preferential, search along the hard-to-estimate depth DOF. The problem is that the two sets of directions have very different properties and scales. Precisely because they have such similar image appearances, related kinematic minima may cause confusion even if they are separated by significant distances in parameter space, whereas false-correspondence minima only cause confusion if they are relatively nearby. In other words, the natural metric for tracker confusion, and hence for the sampling distribution of the randomized local search, is perceptual image distance, not parameter space distance. This holds notwithstanding the fact that large jumps in configuration (depth) are improbable under natural human dynamics. The tracker may have been gradually misled over a period of time, and it is essential that it should be able to jump far enough to recover before tracking fails entirely.[2]

This suggests that we need to inflate the dynamical noise preferentially along the depth directions. But these depend strongly on where the model is viewed from, so no constant (configuration or camera-position independent) noise inflation suffices here. The simplest way to adapt the noise to the configuration/camera-position is to estimate the covariance of the posterior likelihood and use this for noise scaling. (In fact, we advocate inflating the *prior* covariance—the previous posterior after dynamics with *physically realistic* noise levels—i.e., there should be both realistic dynamics and some degree of deliberate random search). Evaluating covariances might be burdensome in a conventional particle tracking framework where we only had point samples of likelihoods, but we have already seen that some form of local refinement of the samples is practically essential in high dimensions, and efficient local optimizers require (and in the case of quasi-Newton style methods, even provide) information equivalent to covariance estimates.

To emphasize how much difference covariance scaling can make, consider the 32 DOF cost spectrum in Figure 5, which has a 2000:1 range of principal standard deviations. For inflation large enough to double the sampling radius along the most uncertain direction (e.g., for a modest search for local minima along this cost valley), a scaling based on uniform dynamical noise would produce a search volume 10^{54} times larger than that of our prior-based one, and an overwhelming fraction of these samples would have extremely high cost and images implausibly different from the source image (see also Figure 1). Such wastage factors are clearly untenable. In practice, samplers based on inflating non-covariance-based dynamical noises simply cannot sample deeply enough along the most uncertain (depth) directions to find the local minima there, and frequent mistracking is the result.

Finally, given that we are including a component of covariance-scaled but inflated noise expressly as a local search mechanism, what kinds of noise distributions will give the most efficient search? Basically, we need to keep a reasonably large proportion of the samples focused on the current track, while scattering the others fairly widely in the hope of finding other good tracks. Also, volume increases very rapidly with radius in high dimensions, so (even with local optimization) we cannot hope to sample densely enough to provide effective search coverage at large inflation factors. It is preferable

2. Ideally, a subsequent smoothing process would push the corrective jump back in time to where the error first occurred (where the jump presumably becomes small). But whether or not this is done, the likelihood penalty for following an incorrect path arbitrarily far forwards in time is likely to be greater than that for any single corrective jump, bad as this may be.

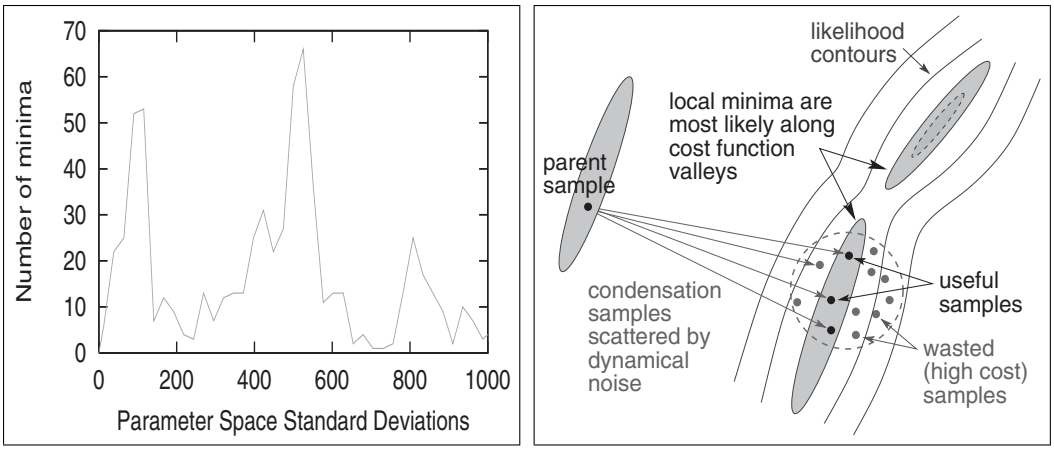


Fig. 1. (a) Typical parameter space minima distribution measured with respect to an arbitrary minimum. Notice that the minima are far from each other in parameter space so wide sampling is necessary to find them. However, boosting the dynamics by sampling from the transition prior (as in particle filtering) leads to inefficiencies (b).

to choose a moderate inflation level, even though this only provides access to relatively nearby local minima.

In summary, owing to its high dimensionality and the ill-conditioning and multimodality associated with unobservable depth DOF, we believe that reliable 3D-from-monocular human body tracking requires deliberate sampling (or some other form of local search) in a region shaped by, but significantly larger than, the local state covariance, followed by local optimization of the samples before any resampling step.

1.2. Previous Work

Below we compare our method to several existing ones, which we briefly summarize here without attempting a full literature review. 3D body tracking from monocular sequences is significantly harder than 2D (Cham and Rehg 1999; Ju, Black, and Yacoob 1996) or multi-camera 3D (Kakadiaris and Metaxas 1996; Gavrila and Davis 1996; Bregler and Malik 1998; Delamarre and Faugeras 1999; Plankers and Fua 2001; Drummond and Cipolla 2001) tracking, and surprisingly few works have addressed it (Deutscher, Blake, and Reid 2000; Sidenbladh, Black, and Fleet 2000; Wachter and Nagel 1999; Gonglaves et al. 1995; Howe, Leventon, and Freeman 1999; Brand 1999).

Deutscher, Blake, and Reid (2000) use a sophisticated "annealed sampling" strategy and a cross-over operator (Deutscher, Davidson, and Reid 2001) to speed up CONDENSATION. They report very good results for unconstrained full-body motion, but for the main sequence they use three cameras and a black background to limit the impact of the alternative minima produced by clutter and depth ambiguities. Sidenbladh, Black, and Fleet (2000) use a similar importance sampling technique with a strong learned prior walking model or a database of motion snippets (Sidenbladh, Black, and Sigal 2002) to track a walking person in an outdoor monocular sequence. Subsequent work (Sidenbladh and Black 2001) integrates flow, edge and ridge cues using Laplace-like error distributions learned from training data, and shows improved upper body tracking for a subject performing planar motion in a cluttered scene, acquired with a moving camera. Our current method uses no motion model—we optimize static poses—but it is true that when they hold, prior motion models are very effective tracking stabilizers. It is possible, but expensive, to track using a bank of motion models (Blake, North, and Isard 1999). Partitioned sampling (MacCormick and Isard 2000) is another notable sampling technique for articulated models, under certain labeling assumptions (MacCormick and Isard 2000; Deutscher, Blake, and Reid 2000).

Several authors have addressed the difficulty that the sampling-based searches of pure particle filtering converge rather slowly to modes (Pitt and Shephard 1997; Heap and Hogg 1998; Cham and Rehg 1999; Merwe et al. 2000; Choo and Fleet 2001), especially when the observation likelihood peaks deep in the tail of the prior. This is especially problematic in high dimensions, where prohibitively long sampling runs are often required for convergence. Heap and Hogg (1998), Cham and Rehg (1999), and Merwe et al. (2000) all combine CONDENSATION-style sampling with either local optimization or Kalman filtering, while Pitt and Shephard (1997) sample discretely using the current observation likelihood (and not the transition prior). The visual trackers of Heap and Hogg (1998) and Cham and Rehg (1999) combine CONDENSATION-style sampling with least-squares optimization, but they only consider the simpler (and much better conditioned) case of 2D tracking. Cham and Rehg combine their heuristic 2D scaled prismatic model (SPM) body representation with a first-order motion model and a piecewise Gaussian resampling method for the CONDENSATION step. The Gaussian covariances are estimated from the

Gauss–Newton approximation at the fitted optima,[3] but the search region widths are controlled by the traditional method of adding a large dynamical noise (Cham and Rehg 1999, Section 3.2).

Choo and Fleet (2001) use a stick model (without any shape model) for which 3D–2D joint-to-image correspondences from motion capture data are available and propose a (gradient-based) hybrid Monte Carlo sampler that is more efficient than (point-based) CONDENSATION. The method provides more efficient local descent towards the minima, but it is still prone to trapping in sub-optimal local minima.

Wachter and Nagel (1999) use articulated kinematics and a shape model built from truncated cones, and estimate motion in a monocular sequence, using edge and intensity (optical flow) information using an extended Kalman filter. Anatomical joint limits are enforced at the level of the filter prediction, but not during the update step, where they could be violated. They show experiments in an unconstrained environment for a subject wearing normal clothing, tracking motion parallel with the image plane using articulated models with 10–15 DOF.

Both Brand (1999) and Howe, Leventon, and Freeman (1999) pose 3D estimation as a learning and inference problem, assuming that some form of 2D tracking (stick 2D model positions or silhouettes) is available over an entire time series. Howe, Leventon, and Freeman (1999) learn Gaussian distributions over short "snippets" of observed human motion trajectories, then use these as priors in an EM-based Bayesian MAP framework to estimate new motions. Brand (1999) learns an HMM with piecewise linear states and solves for the MAP estimate using an entropy minimization framework. As presented, these methods are basically monomodal so they cannot accommodate multiple trajectory interpretations, and they also rely heavily on their learned-prior temporal models to stabilize the tracking. Nevertheless, they provide a powerful higher-level learning component that is complementary to the framework proposed in this paper.

2. Human Body Model

Our human body model (Figures 2(a)–(c)) consists of a kinematic "skeleton" of articulated joints controlled by angular **joint parameters** $\mathbf{x}_a$, covered by "flesh" built from superquadric ellipsoids with additional tapering and bending parameters (Barr 1984). A typical model has around 30 joint parameters, plus eight **internal proportion** parameters $\mathbf{x}_i$ encoding the positions of the hip, clavicle and skull tip joints,

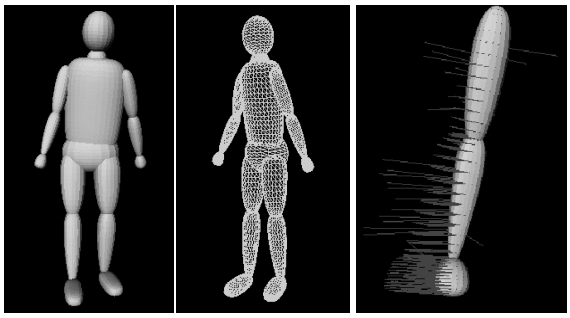


Fig. 2. Different body models using for tracking (a,b,c). In (c) the prediction errors for a model configuration are also plotted (per node, for a contour and intensity cost function, see text).

plus nine **deformable shape** parameters for each body part, gathered into a vector $\mathbf{x}_d$. A complete model can be encoded as a single large parameter vector $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_d, \mathbf{x}_i)$. During tracking we usually estimate only joint parameters, but during initialization the most important internal proportions and shape parameters are also optimized, subject to a soft prior based on standard humanoid dimensions obtained from Hanim-Humanoid Animation Working Group (2002) and updated using collected image evidence. This model is far from photorealistic, but it suffices for high-level interpretation and realistic occlusion prediction, offering a good trade-off between computational complexity and coverage.

The model is used as follows. Superquadric surfaces are discretized as meshes parametrized by angular coordinates in a 2D topological domain. Mesh nodes $\mathbf{u}_i$ are transformed into 3D points $\mathbf{p}_i = \mathbf{p}_i(\mathbf{x})$ and then into predicted image points $\mathbf{r}_i = \mathbf{r}_i(\mathbf{x})$ using composite nonlinear transformations

$$\mathbf{r}_i(\mathbf{x}) \;=\; P(\mathbf{p}_i(\mathbf{x})) \;=\; P(A(\mathbf{x}_a, \mathbf{x}_i, D(\mathbf{x}_d, \mathbf{u}_i))). \qquad (1)$$

Here D represents a sequence of parametric deformations that construct the corresponding part in its own reference frame, A represents a chain of rigid transformations that map it through the kinematic chain to its 3D position, and P represents perspective image projection. During model estimation, robust prediction-to-image matching cost metrics are evaluated for each predicted image feature $\mathbf{r}_i$, and the results are summed over all features to produce the image contribution to the overall parameter space cost function. We use both direct image-based cost metrics such as robustified normalized edge energy, and extracted feature-based ones. The latter associate the predictions $\mathbf{r}_i$ with one or more nearby image features $\bar{\mathbf{r}}_i$ (with additional subscripts if there are several matches). The cost is then a robust function of the prediction errors $\Delta\mathbf{r}_i(\mathbf{x}) \;=\; \bar{\mathbf{r}}_i - \mathbf{r}_i(\mathbf{x})$.

3. The covariance estimates of nonlinear least-squares optimizers as used by Heap and Hogg (1998) and Cham and Rehg (1999) are not robust to model/image matching errors and incorrect (i.e. biased) for natural image statistics that have highly non-Gaussian shape with high kurtosis and long tails (Zhu and Mumford 1997; Sidenbladh and Black 2001). We use an observation likelihood and a robust local continuous optimizer based on heavy tail error distributions (see Sections 3.1 and 4.1) to address these problems.

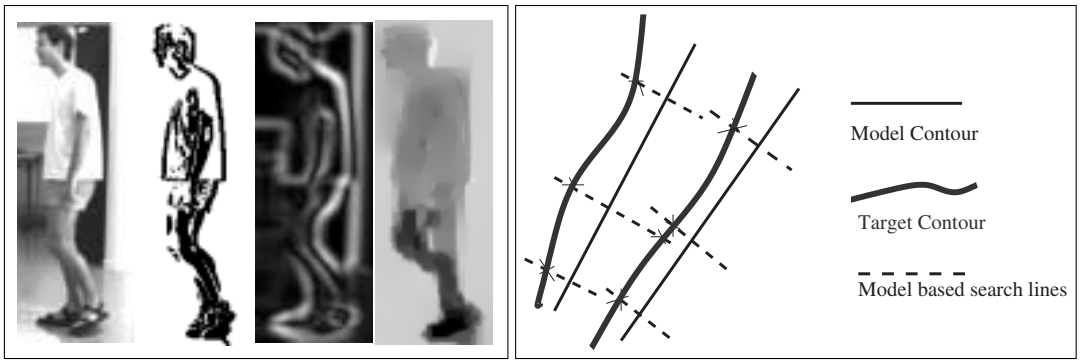


Fig. 3. Examples of our robust low-level feature extraction: (a) original image; (b) motion boundaries; (c) intensity-edge energy; (d) robust horizontal flow field; (e) the model-based edge matching process. Multiple edge matches found along individual search lines (model projected contour normals) are fused using a probabilistic assignment strategy (see text).

3. Problem Formulation

We aim towards a probabilistic interpretation and optimal estimates of the model parameters by maximizing the total probability according to Bayes rule

$$p(\mathbf{x}|\bar{\mathbf{r}}) \propto p(\bar{\mathbf{r}}|\mathbf{x})\, p(\mathbf{x}) = \exp\left(-\textstyle\sum_i e(\bar{\mathbf{r}}_i|\mathbf{x})\right) p(\mathbf{x}) \quad (2)$$

where $e(\bar{\mathbf{r}}_i|\mathbf{x})$ is the cost density associated with the observation of node i and $p(\mathbf{x})$ is a prior on model parameters. In our MAP approach, we discretize the continuous problem and attempt to minimize the negative log-likelihood for the total posterior probability, expressed as the following cost function:

$$f(\mathbf{x}) = -\log(p(\bar{\mathbf{r}}|\mathbf{x})\, p(\mathbf{x})) \quad (3)$$

$$= -\log p(\bar{\mathbf{r}}|\mathbf{x}) - \log p(\mathbf{x}) = f_o(\mathbf{x}) + f_p(\mathbf{x}). \quad (4)$$

3.1. Observation Likelihood

Whether continuous or discrete, the search process depends critically on the observation likelihood component of the parameter space cost function. Besides smoothness properties, the likelihood should be designed to limit the number of spurious local minima in parameter space. Our method employs a combination of robust edge and intensity information on top of a multiple assignment strategy based on a weighting scheme that focuses attention towards motion boundaries. Our likelihood term is also based on robust (heavy-tailed) error distributions. Note that both robustly extracted image cues and robust parameter space estimation are used; the former provides "good features to track", while the latter directly addresses the model-image association problem.

3.1.1. Robust Error Distributions

Robust parameter estimation can be viewed as the choice of a realistic total likelihood model for the combined inlier and outlier distributions for the observation. We model the total likelihood in terms of robust radial terms ρ_i, where $\rho_i(s)$ can be any increasing function with $\rho_i(0) = 0$ and $\frac{\mathrm{d}}{\mathrm{d}s}\rho_i(0) = \frac{\nu}{\sigma^2}$. These model error distributions correspond to a central peak with scale σ, and a widely spread background of outliers ν. Here we have used the "Lorentzian" $\rho_i(s,\sigma) = \nu \log(1 + \frac{s}{\sigma^2})$ and "Leclerc" $\rho_i(s,\sigma) = \nu(1 - \exp(-\frac{s}{\sigma^2}))$ robust error potentials.

The cost for the observation i, expressed in terms of the corresponding model prediction is $e(\bar{\mathbf{r}}_i|\mathbf{x}) = \frac{1}{N\nu} e_{ui}(\mathbf{x})$, where N is the total number of model nodes, $\mathbf{W}_i$ is a positive definite weighting matrix associated to the assignment i, and

$$e_i(\mathbf{x}) = \begin{cases} \frac{1}{2}\rho_i(\mathbf{\Delta r}_i(\mathbf{x})\, \mathbf{W}_i\, \mathbf{\Delta r}_i(\mathbf{x})^\top) & \text{if } i \text{ is assigned} \\ \nu_{bf} = \nu & \text{if back-facing} \\ \nu_{occ} = k\nu, \;\; k > 1 & \text{if occluded} \end{cases} . \quad (5)$$

The robust observation likelihood contribution is thus

$$f_o(\mathbf{x}) = -\log p(\bar{\mathbf{r}}|\mathbf{x}) \quad (6)$$

$$= f_a(\mathbf{x}) + N_{bf}\, \nu_{bf} + N_{occ}\, \nu_{occ} \quad (7)$$

where $f_a(\mathbf{x})$ represents the term associated with the image assigned model nodes, while N_{occ} and N_{bf} are the numbers of occluded and back-facing (self-occluded) model nodes.

Notice that occluded model predictions are not simply ignored. They contribute a constant penalty to the overall observation likelihood. This is necessary in order to build likelihoods that preserve their response properties under occlusion and viewpoint change. For instance, good fits from both frontal and side views should ideally have similar peak responses, but it is clear that the number of occluded model points is in general larger in a side view than in a frontal one. This can lead to down-weighting of peaks for side views if only the visible nodes are taken into account. An additional difficulty arises, for example, in cases where the legs pass each other (in a side view) and the model "locks" both of its legs onto the same image leg. To avoid such situations, we

include all of the model nodes when fusing the likelihood, but we slightly penalize occluded ones in order to make them less attractive. A way to choose the occlusion penalty ν is to fit the model to the data and compute an approximate error per node. By using a slightly higher value for occluded nodes, we make them more attractive than a bad fit but less attractive than other non-occluded states that can exist in the neighborhood of the parameter space. We find this heuristic gives good results in practice,[4] although a more rigorous treatment of occlusion would be desirable in the general case. At present, this is computationally too expensive, but interesting approximations can be found in MacCormick and Blake (1998).

3.1.2. Cue Integration and Assigned Image Descriptors

We use both edge and intensity features in our cost function; see Sminchisescu (2000b) for details. For edges, the images are smoothed with a Gaussian kernel, contrast normalized, and a Sobel edge detector is applied. For intensities, a robust multi-scale optical flow method based on the implementation of Black and Anandan (1996) gives both a flow field and an associated outlier map (see Figure 3(b)). The outlier map is processed similar to edges, to obtain a smooth 2D potential field S_p. It conveys useful information about the motion boundaries and is used to weight the significance of edges (see Figure 3(b)). We typically use diagonal weighting matrices $\mathbf{W}_i$, associated with the predicted feature $\mathbf{r}_i$ and corresponding matched observation $\bar{\mathbf{r}}_i$, of the form $\mathbf{W}_i(\mathbf{r}_i) = 1 - kS_p(\bar{\mathbf{r}}_i)$, where k is a constant that controls the emphasis and confidence in the motion boundary estimation. (The smoothed motion boundary image is a real image with values between 0 and 1 as in Figure 3(b). For instance, $k = 0$ will weight all the edges uniformly, while $k = 1$ will entirely exclude the edge responses that are not on motion boundaries.) In practice, we found that values of k in the range $k = 0.2$–0.4 worked well. For visible nodes on model occluding contours ($\mathcal{O}$), we perform line search along the normal and retain all possible assignments within the search window (see Figure 3(e)), weighting them by their importance qualified by the motion boundary map $\mathbf{W}$. For visible model nodes lying inside the object ($\mathcal{I}$), we use the correspondence field derived from the robust optical flow at their corresponding image prediction. This acts as a residual measurement error at each visible model node; see Sminchisescu (2002b) for details). The assigned data term (6) thus becomes

$$f_a(\mathbf{x}) = \frac{1}{2} \sum_{i \in \mathcal{O}, e \in \mathcal{E}_i} \rho_{ie}(\Delta \mathbf{r}_{ie}(\mathbf{x})\, \mathbf{W}_{ie}\, \Delta \mathbf{r}_{ie}(\mathbf{x})^{\mathrm{T}}) \qquad (8)$$

$$+ \frac{1}{2} \sum_{j \in \mathcal{I}} \rho_{jf}(\Delta \mathbf{r}_{jf}(\mathbf{x})\, \mathbf{W}_{jf}\, \Delta \mathbf{r}_{jf}(\mathbf{x})^{\mathrm{T}}) \qquad (9)$$

where the subscripts "i_e" denote multiple edges $\mathcal{E}_i$ assigned to model prediction i, and "j_f" denote the flow term assigned to model prediction j.

3.2. Model Priors

The complete prior penalty over model parameters is a sum of negative log likelihoods $f_p = f_{an} + f_s + f_{pa}$ corresponding to the following prior densities p_{an}, p_s, and p_{pa}.

3.2.1. Anthropometric Data p_{an}

The internal proportions for a standard humanoid (based on statistical measurements) are collected from Hanim-Humanoid Animation Working Group (2002) and used effectively as a Gaussian prior, $p_{an} = \mathcal{N}(\boldsymbol{\mu}_{an}, \boldsymbol{\Sigma}_{an})$, to estimate a concrete model for the subject to be tracked. Left–right symmetry of the body is assumed; only "one side" of the internal proportions parameters are estimated while collecting image measurements from the entire body.

3.2.2. Parameter Stabilizers p_s

Certain modeling details are far more important than one might think. For example, it is impossible to track common turning and reaching motions unless the clavicle joints in the shoulder are modeled accurately. However, these parameters have fairly well-defined equilibrium positions and leaving them unconstrained would often lead to ambiguities that produce nearly singular (flat) cost surfaces. We control these hard-to-estimate parameters with long-tailed "sticky prior" stabilizers scaling their Gaussian equilibria, $p_s = \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. This ensures that in the absence of strong observations, the parameters are constrained to lie near their default values, whereas stronger observations can "unstick" them from the defaults and effectively turn off the prior.

3.2.3. Anatomical Joint Angle Limits C_{bl}

The 3D consistency requires that the values of joint angles evolve within anatomically consistent intervals. Also, when estimating internal body proportions during initialization, we ensure that they remain within a certain range of deviation from the standard humanoid (typically 10%). We model this with a set of inequalities of the form $C_{bl} \cdot \mathbf{x} < 0$, where C_{bl} is a "box-limit" constraint matrix.

3.2.4. Body Part Interpenetration Avoidance p_{pa}

Physical consistency requires that different body parts do not interpenetrate during estimation. We avoid this by introducing repulsive potentials that decay rapidly outside the surface of

4. This is particularly effective when combined with the covariance scaled sampling (CSS) algorithm presented in Section 4. Loss of visibility of certain body parts leads to increased uncertainty in related parameters, and CSS automatically ensures broader sampling in those parameter space regions.

each body part, $f_{pa} = \exp(-f(\mathbf{x})\,|f(\mathbf{x})|^{p-1})$, where $f(\mathbf{x}) < 0$ defines the interior of the part and p controls the decay rate.

3.3. Distribution Representation

We represent parameter space distributions as sets of separate modes $m_i \in \mathcal{M}$, each having an associated overall probability, mean and covariance matrix $m_i = (\mu_i, \Sigma_i, c_i)$. These can be viewed as Gaussian mixtures. Cham and Rehg (1999) also use multiple Gaussians, but they had to introduce a special piecewise representation as their modes seem to occur in clusters after optimization. We believe that this is an artifact of their cost function design. In our case, as the modes are the result of robust continuous optimization, they are necessarily either separated or confounded. Our 3D-from-monocular application also requires a more effective sampling method than the 2D one of Cham and Rehg (1999), as explained in Section 4.2.

3.4. Temporal Propagation

Equation 2 reflects the search for the model parameters in a static image, under likelihood terms and model priors but without a temporal or initialization prior. For temporal observations $\mathbf{R}_t = \{\bar{\mathbf{r}}_1, \bar{\mathbf{r}}_2 ..., \bar{\mathbf{r}}_t\}$, and sequence of states $\mathbf{X}_t = \{\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_t\}$, the posterior distribution over model parameters becomes

$$p(\mathbf{x}_t|\mathbf{R}_t) \;\propto\; p(\bar{\mathbf{r}}_t|\mathbf{x}_t)\; p(\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1})\; p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) \tag{10}$$

where $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is a dynamical prior and $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ is the prior distribution from $t-1$. Together they form the temporal prior $p(\mathbf{x}_t|\mathbf{R}_{t-1})$ for initializing the static image search (2).[5]

4. Search Algorithm

Our parameter search technique combines robust constraint-consistent local optimization with a more global discrete sampling method.

4.1. Mode Seeking using Robust Constrained Continuous Optimization

The cost function is a negative log likelihood. In order to optimize a sample $\mathbf{x}$ to find the center of its associated likelihood peak, we employ an iterative second-order robust constrained local optimization procedure. At each iteration, the log-likelihood gradients and Hessians of the observations and the soft priors[6] are assembled from eq. (3):

$$\mathbf{g} = \frac{\mathbf{d}f}{\mathbf{dx}} = \mathbf{g}_o + \nabla f_{an} + \nabla f_s + \nabla f_{pa} \tag{11}$$

$$\mathbf{H} = \frac{\mathbf{d}^2 f}{\mathbf{dx}^2} = \mathbf{H}_o + \nabla^2 f_{an} + \nabla^2 f_s + \nabla^2 f_{pa}. \tag{12}$$

For local optimization we use a second-order trust region method, where a descent direction is chosen by solving the regularized subproblem (Fletcher 1987)

$$(\mathbf{H} + \lambda \mathbf{W})\,\delta\mathbf{x} \;=\; -\mathbf{g} \quad \text{subject to} \quad C_{bl} \cdot \mathbf{x} < 0 \tag{13}$$

where $\mathbf{W}$ is a symmetric positive definite damping matrix and λ is a dynamically chosen weighting factor. Joint limits C_{bl} are handled as hard bound constraints in the optimizer, by projecting the gradient onto the currently active (i.e., currently unlimited) variables. The joint constraints change the character of the cost function and the minima reached very significantly. Figure 4 plots a one-dimensional (1D) slice through the constrained cost function together with a second-order Taylor expansion of the unconstrained cost. Owing to the presence of the bounds, the cost gradient is nonzero (orthogonal to the active constraints) at the constrained minimum. The unconstrained cost function is smooth, but the constrained one changes gradient abruptly when a constraint is hit, essentially because the active-set projection method changes the motion direction to maintain the constraint.

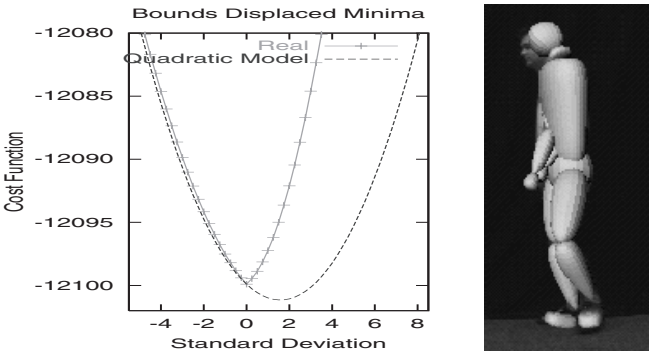


Fig. 4. (a) Displaced minimum due to joint limits constraints. (b) Joint limits without body non-self-intersection constraints do not suffice for physical consistency.

4.2. Covariance Scaled Sampling

Although representations based on propagating multiple modes, hypotheses or samples tend to increase the robustness of model estimation, the great difficulty with high-dimensional distributions is finding a proposal density that

5. In practice, at any given time step we work on a negative log-likelihood "energy" function that is essentially static, being based on both the current observation likelihood and the parameter space priors, as in eq. (3). The samples from the temporal prior $p(\mathbf{x}_t|\mathbf{R}_{t-1})$ are used as initialization seeds for local energy minimization. The different minima found will represent the components of the posterior mixture representation.

6. "Soft" means that these terms are part of the cost surface, whereas "hard" constraints such as joint limits restrict the range of variation of their corresponding parameters.

can be sampled, which often hits their **typical sets**—the areas where most of their probability mass is concentrated. Here we develop a proposal density based on local parameter estimation uncertainties.[7] The local sample optimizations give us not only local modes, but also their (robust, constraint consistent) Hessians and hence estimates of the local posterior parameter estimation uncertainty at each mode.[8]

The main insight is that alternative cost minima are most likely to occur along local valleys in the cost surface, i.e., along highly uncertain directions of the covariance. It is along these directions that cost-modeling imperfections and noise, and 3D nonlinearities and constraints, have the most likelihood of creating multiple minima, as the cost function is shallowest and the 3D movements are largest there. This is particularly true for monocular 3D estimation, where the covariance is unusually ill-conditioned owing to the many poorly observable motion-in-depth DOF. Some examples of such multimodal behavior along high covariance eigen-directions are given in Figure 7. Also, it is seldom enough to sample at the scale of the estimated covariance. Samples at this scale almost always fall back into the same local minimum, and significantly deeper sampling is necessary to capture nearby but non-overlapping modes lying further up the valley.[9] Hence, we sample according to rescaled covariances, typically scaling by a factor of eight or so. Finally, we can sample either randomly, or according to a regular pattern.[10] For the experiments shown here, we use random sampling using CSS with Gaussian tails. Figure 6 summarizes the resulting covariance-scaled search method.

Given the explanations above, we must implement the following steps:

(i) Generate fair samples from a prior with known modes. This is easy. In our case we propagate Gaussian mixtures[11] which can be used as importance sampling distributions, and correction weighting is readily performed. Mixtures provide a compact, explicitly multimodal representation and accurate localization, advantages emphasized by Heap and Hogg (1998) and Cham and Rehg (1999); see Section 5.1. However, both papers use sampling stages based on the unmodified process model (i.e., dynamics with fixed, near-spherical noise), which therefore have trapping and sample wastage problems analogous to CONDENSATION.

(ii) Recover new modes of a distribution for which only *some* of the modes are known. This is significantly more difficult. A priori, the distribution of unknown modes is not available, nor are the boundaries of the basins of attraction of the existing modes (in order to find their neighbors). Also, such likelihood peaks are often well separated in configuration space (e.g., the forwards/backwards flipping ambiguities for human pose, or the cascades of incorrect matches when a model limb is assigned to the incorrect side of an image limb). For typical distributions of minima in parameter space and in cost, see Figure 13 and the results in Table 1. For well-separated peaks, sampling based purely on the known (and potentially incomplete) ones is inadequate, as most of the samples will simply fall back into the peaks they arose from.[12] So broader sampling is necessary, but it is also important to focus the samples in relatively low cost regions (see also Figure 1). To achieve this we propose to use the local cost surface to shape a broad sampling distribution. As expected on theoretical grounds, this turns out to give significantly improved results for CSS (for sample cost median, number of minima found, their cost) than competing methods based on either pure prior-based sampling or prior-based sampling plus spherical "dynamical" noise (see Table 1).

(iii) Sample a prior under dynamic observations but without making restrictive assumptions on the motion of its peaks. In this case the modes from time $t-1$ are available, and it is critical that the sampling procedure cover the peaks of the observation likelihood in the next time step t. This means that samples should be generated in

7. Related variable metric ideas can be found in global optimization, in the context of continuous annealing (Vanderbilt and Louie 1984) and have been applied by Black (1992) to low-dimensional (2D) optical flow computation.

8. A sample is optimized to convergence to obtain the corresponding mode. The Hessian matrix at the convergence mode gives the principal curvature directions and magnitude around the mode and its inverse gives the covariance matrix, reflecting the cost local uncertainty structure. The Hessian is estimated by the algorithm (Section 4.1) during optimization (using eq. (11)), and the covariance is readily obtained from there.

9. In part this is due to imperfect modeling, which easily creates biases greater than a few standard deviations, particularly in directions where the measurements are weak. Also, one case in which multiple modes are likely to lie so close together in position and cost that they cause confusion is when a single mode fragments due to smooth evolutions of the cost surface. In this case, singularity ("catastrophe") theory predicts that generically, exactly two modes will arise (bifurcation) and that they will initially move apart very rapidly (at a speed proportional to $1/\sqrt{t}$). Hence, it is easy for one mode to get lost if we sample too close to the one we are tracking.

10. For efficiency purposes, an implementation could sample regularly, in fact only along lines corresponding to the lowest few covariance eigen-directions. Although this gives a very sparse sampling indeed, this is an avenue that can be explored in practice.

11. There are at least two ways to obtain a mixture. One is by clustering a set of posterior samples generated, e.g., by CONDENSATION updates. This may produce centers that are not necessarily well separated, and that may not actually reflect the true modes of the posterior owing to sampling artifacts. Another possibility, followed here, is to optimize the samples locally. In this case the modes found are true local peaks that are, necessarily, either separated or confounded.

12. Several metrics exist for assessing the efficiency of particle filters (Liu 1996; MacCormick and Isard 2000). The "survival diagnostic" (also called "effective sample size") measures how many particles will survive a resampling operation. If the weights are unbalanced very few may survive, thus reducing search diversity. But balanced weights do not imply that all peaks have been well explored; samples trapped in a single mode have reasonably well-balanced weights. The same criticism applies to the "survival rate". This tries to characterize the ratio of the volume of support of the posterior to that of the prior. Low values suggest that the filter may produce inaccurate density estimates, but again trapping leaves the survival rate reasonably high.

Table 1. Quantitative Results for the Distribution of Minima Found

Method	Scale	Number of Minima	Parameter Distance Median		Standard Deviations Median		Cost Median	
			Unopt	Opt	Unopt	Opt	Unopt	Opt
CSS	1	8	1.148	2.55242	10.9351	47.6042	116.951	8.49689
CSS	4	59	3.21239	2.9474	35.2918	55.3163	1995.12	6.98109
CSS	8	180	4.969	3.34661	75.1119	109.813	16200.8	7.09866
SS	1	0	0.199367	—	24.5274	–	273.509	–
SS	4	11	0.767306	2.04928	96.1519	39.0745	4291.12	6.28014
SS	8	42	1.47262	2.54884	188.157	56.8268	16856.1	6.96481

Note that CSS finds more minima and places raw samples at lower cost than SS.

the basins of attraction of the density peaks *after* applying the dynamical update. In the absence of knowledge about the peaks' motion (i.e., known system dynamics), we exploit the local uncertainty structure in the distribution, and shape the search region based on it. Again, broader sampling is necessary, as the tracked object moves between frames. Also, as explained above, the mode tracking process is not one-to-one. New modes might emerge or split under the effect of increased uncertainty, and it is important that the sampling process does not miss such events by sampling too close to a given mode core, which may both move and split between two temporal observations. Our quantitative results in Section 6 directly support such findings, e.g., for mode splitting reflecting bi-modality generated by locally-planar versus in-depth motion explanations (see below).

In this paper, we have not used specific motion models as we want to be able to track general human motions (see, for instance, the sequences given in Figures 9–11). For the experiments shown in the next section, we used trivial driftless diffusion dynamics, so CSS has to account for local uncertainty and sample widely enough to cover moving peaks. We could also use constant velocity dynamics, or more sophisticated learned motion models such as walking (Rohr 1994; Deutscher, Blake, and Reid 2000; Sidenbladh, Black, and Fleet 2000). When they hold, such models can significantly stabilize tracking, but note that they often turn out to be misleading, e.g., when the subject makes unexpected motions like turning or switching activities.

To build up intuition about the shape of our cost surface, we studied it empirically by sampling along uncertain covariance directions (in fact, eigenvectors of the covariance matrix), for various model configurations. With our carefully selected image descriptors, the cost surface is smooth apart from the apparent gradient discontinuities caused by active-set projection at joint constraint activation points. Hence, our local optimizer reliably finds a local minimum. We find that multiple modes do indeed occur for certain configurations, usually separated by cost barriers that a classical (uninflated) sampling strategy would have difficulty crossing. For example, Figure 7 shows the two most uncertain modes of the Figure 9 human tracking sequence at times 0.8 and 0.9 s. (These are minima only within the sampled slice of parameter space, but they do lie in the attraction zones of full parameter space minima.) Secondary minima like those shown here occur rather often, typically for one of two reasons. The first is incorrect registration and partial loss of track when both edges of a limb model are attracted to the same image edge of the limb. This is particularly critical when there is imperfect body modeling and slightly misestimated depth. The second occurs when the character of a motion in depth is misinterpreted. Image registration is maintained until the incorrect 3D interpretation becomes untenable, at which point recovery is difficult. This situation occurs in Figure 7 (see also Figure 12). Identifying and tracking such ambiguous behaviors is critical, as incorrect depth interpretations quickly lead to tracking failure.

Figure 8(a) shows some typical slices along cost eigendirections at much larger scales in parameter space. Note that we recover the expected robust shape of the matching distribution, with some but not too many spurious local minima. This is crucial for efficiency and robustness, as the tracker can only follow a limited number of possible minima.

5. Model Initialization

Our tracker starts with a set of initial hypotheses produced by a model initialization process. Correspondences need to be specified between model joint locations and approximate joint positions of the subject in the initial image, and a nontrivial optimization process is run to estimate certain body dimensions and the initial 3D joint angles. Previous approaches to single-view model initialization (Taylor 2000; Barron and Kakadiaris 2000) do not fully address the generality and consistency problems, failing to enforce the joint limit constraints, and assuming either restricted camera models or restricted

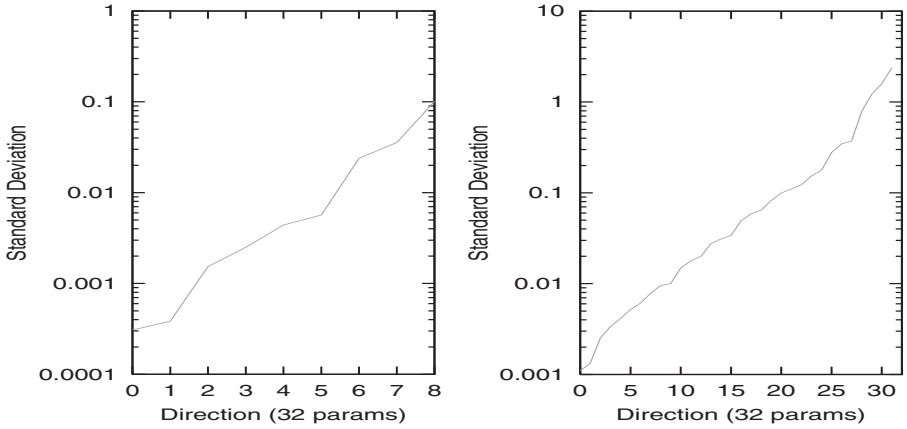


Fig. 5. Typical covariance eigenvalue spectra plotted on a logarithmic scale, for a local minimum. $\sigma_{\max}/\sigma_{\min}$ is 350 for the 8-DOF arm model, and 2000 for the 32-DOF body one.

From the "old" mixture prior $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1}) = \sum_{i=1}^{K} \pi_i^{t-1} \mathcal{N}(\mu_i^{t-1}, \Sigma_i^{t-1})$, at time $t-1$, build "new" mixture posterior $p(\mathbf{x}_t|\mathbf{R}_t) = \sum_{i=1}^{K} \pi_i^t \mathcal{N}(\mu_i^t, \Sigma_i^t)$, at time t, as follows:

---

1. Build covariance scaled proposal density $p_{t-1}^* = \sum_{i=1}^{K} \pi_i^{t-1} \mathcal{P}(\mu_i^{t-1}, \Sigma_i^{t-1})$. For the experiments we have used Gaussian tails. The covariance scaled Gaussian component proposals are $\mathcal{P} = \mathcal{N}(\mu_i^{t-1}, s\Sigma_i^{t-1})$ with $s = 4$–14 in our experiments.

---

2. Generate components of the posterior at time t by sampling from p_{t-1}^* as follows. Iterate over $j = 1...N$ until the desired number of samples N are generated:
2.1. Choose component i from p_{t-1}^* with probability π_i^{t-1}.
2.2. Sample from $\mathcal{P}(\mu_i^{t-1}, \Sigma_i^{t-1})$ to obtain $\mathbf{s}_j$.
2.3. Optimize $\mathbf{s}_j$ over the observation likelihood at time t, $p(\mathbf{x}|\bar{\mathbf{r}}_t)$ defined by eq. (2), using the local continuous optimization algorithm (Section 4.1). The result is the parameter space configuration at convergence μ_j^t, and the covariance matrix $\Sigma_j^t = \mathbf{H}(\mu_j^t)^{-1}$. If the μ_j^t mode has been previously found by a local descent process, discard it (For notational clarity, without any loss of generality, consider all the modes found are different.)

---

3. Construct an un-pruned posterior for time t as: $p_t^u(\mathbf{x}_t|\mathbf{R}_t) = \sum_{j=1}^{N} \pi_j^t \mathcal{N}(\mu_j^t, \Sigma_j^t)$ where $\pi_j^t = \frac{p(\mu_j^t|\bar{\mathbf{r}}_t)}{\sum_{j=1}^{N} p(\mu_j^t|\bar{\mathbf{r}}_t)}$.

---

4. Prune the posterior p_t^u to keep the best K components with highest probability π_j^t (rename indices $j = 1...N$ into the set $k = 1...K$) and renormalize the distribution as follows: $p_t^p(\mathbf{x}_t|\mathbf{R}_t) = \sum_{k=1}^{K} \pi_k^t \mathcal{N}(\mu_k^t, \Sigma_k^t)$ where $\pi_k^t = \frac{p(\mu_k^t|\bar{\mathbf{r}}_t)}{\sum_{j=1}^{K} p(\mu_j^t|\bar{\mathbf{r}}_t)}$.

---

5. For each mixture component $j = 1...K$ in p_t^p, find the closest prior component i in $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$, according to a Bhattacharyya distance $\mathcal{B}_{ij}(\mu_i^{t-1}, \Sigma_i^{t-1}, \mu_j^t, \Sigma_j^t) = (\mu_i^{t-1} - \mu_j^t)^{\mathrm{T}} \left[\frac{\Sigma_i^{t-1} + \Sigma_j^t}{2}\right]^{-1} (\mu_i^{t-1} - \mu_j^t) + \frac{1}{2} \log \frac{\left|(\Sigma_i^{t-1} + \Sigma_j^t)/2\right|}{\sqrt{|\Sigma_i^{t-1}||\Sigma_j^t|}}$.
Recompute $\pi_j^t = \pi_j^t * \pi_i^{t-1}$. Discard the component i of $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ from further consideration.

---

6. Compute the posterior mixture $p(\mathbf{x}_t|\mathbf{R}_t) = \sum_{k=1}^{K} \pi_k^t \mathcal{N}(\mu_k^t, \Sigma_k^t)$ where $\pi_k^t = \frac{\pi_k^t}{\sum_{j=1}^{K} \pi_j^t}$.

Fig. 6. The steps of our covariance-scaled sampling algorithm.

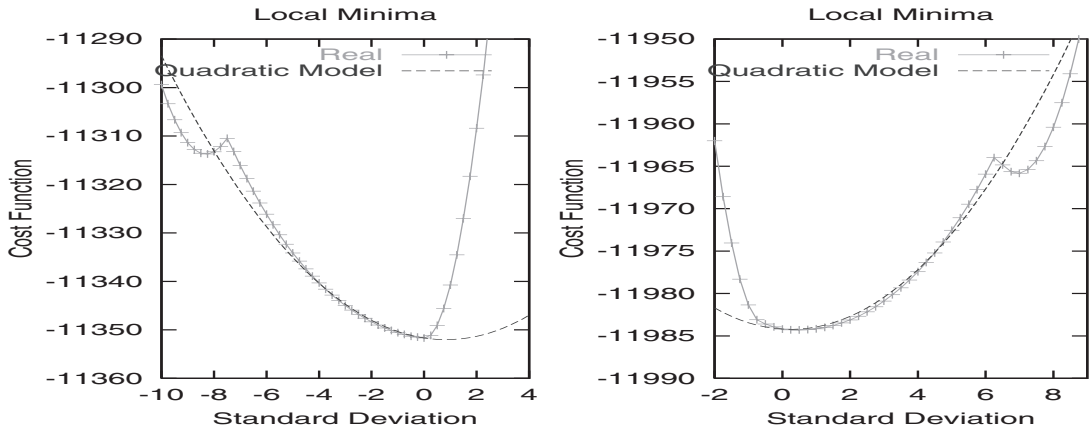


Fig. 7. Multimodality along several uncertain eigen-directions (0.8 and 0.9 s in cluttered body tracking sequence).

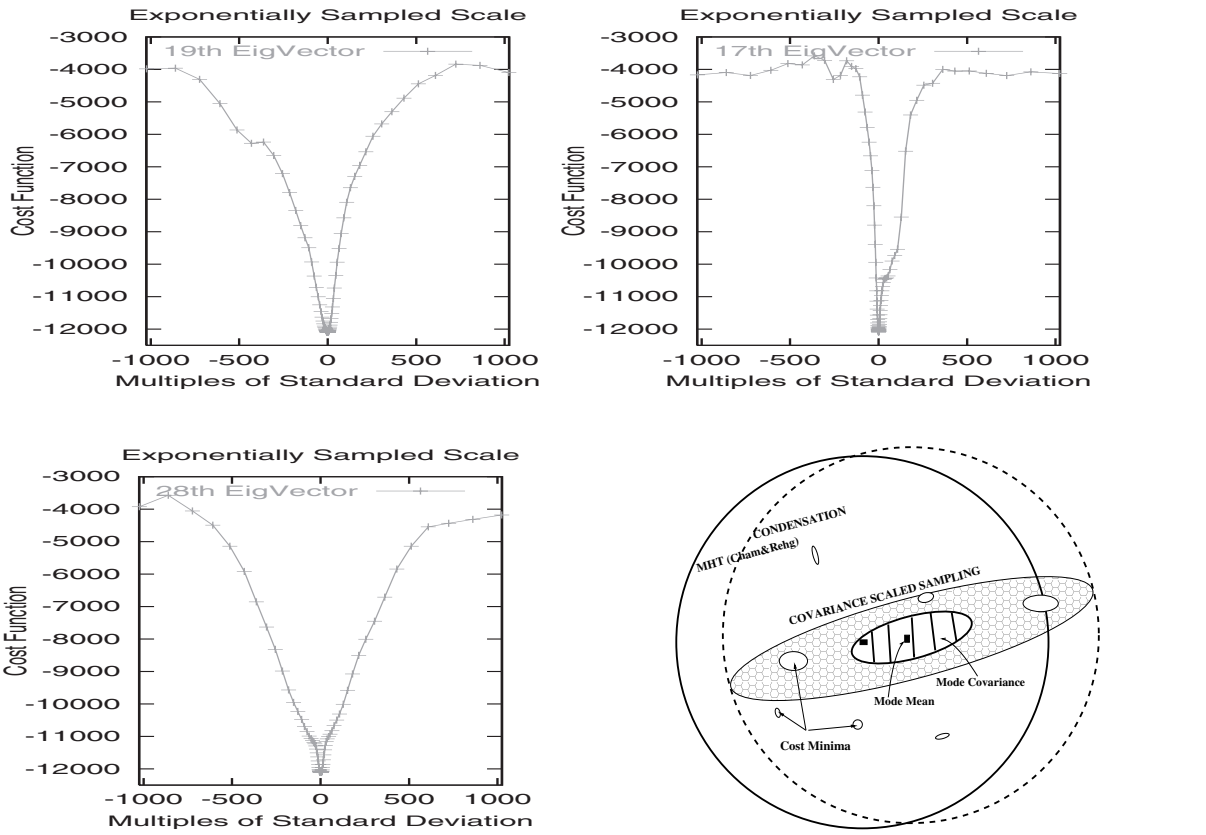


Fig. 8. (a,b,c) Cost function slices at large scales. (d) Comparison of sampling methods: (1) CONDENSATION (dashed circle coverage) randomizes each sample by dynamic noise; (2) MHT, solid circle (Cham and Rehg 1999, Section 3.2, p 3) samples within covariance support (dashed ellipse) and applies the same noise policy as (1); and finally, (3) our CSS (pattern ellipse) targets good cost minima (flat filled ellipses) by inflating or heavy tail sampling the local robust covariance estimation (dashed ellipse)).

human poses in the image. An algorithm like the one we propose could also probably be bootstrapped using estimates of 2D joint positions derived from learned models of silhouette appearance (Rosales and Sclaroff 2000).

For stability, parameters are initialized in three stages, each based on the formulation described in Section 4.1. Hard joint limits are enforced at all stages by the constrained optimization procedure, and corresponding parameters on the left and right sides of the body are held equal, whereas measurements are collected from the entire body (see below). The first stage estimates joint angles $\mathbf{x}_a$, internal proportions $\mathbf{x}_i$ and a few simple shape $\mathbf{x}_d$ parameters, subject to the given 3D–2D joint correspondences and prior intervals on the internal proportions and body part sizes. The second stage uses both the given joint correspondences and the local contour signal from image edges to optimize the remaining volumetric body parameters (limb cross-sections and their tapering parameters $\mathbf{x}_d$) while holding the other parameters fixed. Finally, we refine the full model ($\mathbf{x}$) using similar image information to the second stage. The covariance matrix corresponding to the final estimate is used to generate an initial set of hypotheses, which are propagated in time using the algorithm described in Section 4. While the process is heuristic, it gives a balance between stability and flexibility. In practice, we find that enforcing the joint constraints, mirror information and prior bounds on the variation of body parameters gives far more stable and satisfactory results. However, with monocular images, the initialization always remains ambiguous and highly uncertain in some parameter space directions, especially under 3D–2D joint correspondence data. In our case, we employ a suitable coarse pose initialization and use the above process for fine refinement but, if available, we could fuse pose information from multiple images.

6. Experiments

For the experiments shown here we use an edge and intensity based cost function and a body model incorporating priors and constraints as explained in Sections 3.1 and 3.2. We use Gaussian tails for CSS. A quantitative evaluation of different Gaussian scalings appears in Table 1.

To illustrate our method we show results for an 8 s arm tracking sequence and two full body ones (3.5 and 4 s). All three sequences contain both self-occlusion and significant relative motion in depth. The first two (Figure 9) were shot at 25 frames (50 fields) per second against a cluttered, unevenly illuminated background. The third (Figure 11) is at 50 non-interlaced frames per second against a dark background, but involves a more complex model and motions. In our unoptimized implementation, a 270 Mhz SGI O2 required about 5 s per field to process the arm experiment and 180 s per field for the full body ones, most of the time being spent in cost function evaluation. The figures show the current best candidate model overlayed on the original images. We also explore the characteristic failure modes of various tracker components, as follows. By a *Gaussian single mode tracker* we mean a single hypothesis tracker performing local continuous optimization based on Gaussian error distributions and without enforcing any physical constraints. A *robust single mode tracker* improves this by using robust matching distributions. A *robust single mode tracker with joint limits* also enforces physical constraints. For multimodal trackers, the sampling strategy can be either CONDENSATION-based or CSS-based, as introduced in previous sections.

6.1. Cluttered Background Sequences

These sequences explore 3D estimation behavior with respect to image assignment and depth ambiguities, for a bending rotating arm under an 8-DOF model and a pivoting full-body motion under a 30-DOF one. They have cluttered backgrounds, specular lighting and loose fitting clothing. In the arm sequence, the deformations of the arm muscles are significant and other imperfections in our arm model are also apparent.

The *Gaussian single mode tracker* manages to track 2D fronto-parallel motions in moderate clutter, although it gradually slips out of registration when the arm passes the strong edges of the white pillar (0.5 and 2.2 s for the arm sequence and 0.3 s for the human body sequence). Any significant motion in depth is untrackable.

The *robust single mode tracker* tracks fronto-parallel motions reasonably well even in clutter, but quickly loses track during in-depth motions, which it tends to misinterpret as fronto-parallel ones. In the arm tracking sequence, shoulder motion towards the camera is misinterpreted as fronto-parallel elbow motion, and the error persists until the upper bound of the elbow joint is hit at 2.6 s and tracking fails. In the full body sequence, the pivoting of the torso is underestimated, being partly interpreted as quasi-fronto-parallel motion of the left shoulder and elbow joints. Despite the presence of anatomical joint constraints, the fist eventually collapses into the body if non-self-intersection constraints are not present.

The *robust joint-limit-consistent CSS multimode tracker* tracks the motion of the entire arm and body sequence without failure. We retain just the three best modes for the arm sequence and the seven best modes for the full human body sequence. As discussed in Section 4.2, multimodal behavior occurs mainly during significantly non-fronto-parallel motions, between 2.2–4.0 s for the arm sequence, and over nearly the entire full body sequence (0.2–1.2 s). For the latter, the modes mainly reflect the ambiguity between true pivoting motion and its incorrect "fronto-parallel explanation".

We also compared our method with a 3D version of that of Heap and Hogg (1998) and Cham and Rehg (1999). These methods were developed for 2D tracking and we were interested in how well they would behave in the far less well

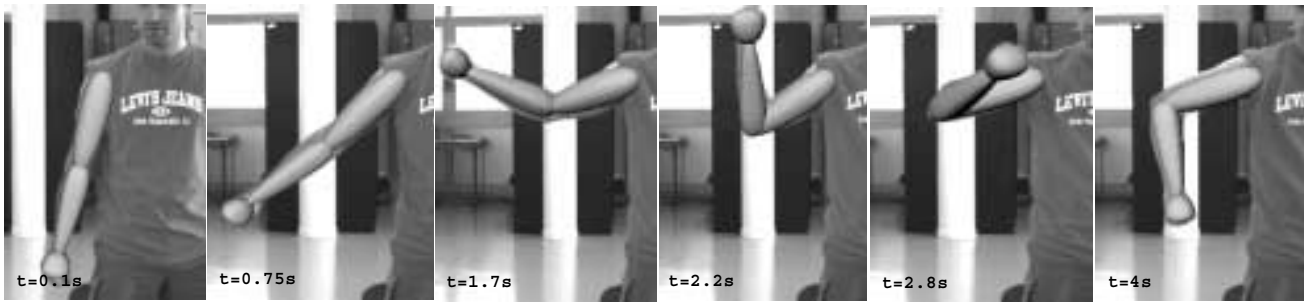


Fig. 9. Arm tracking against a cluttered background.

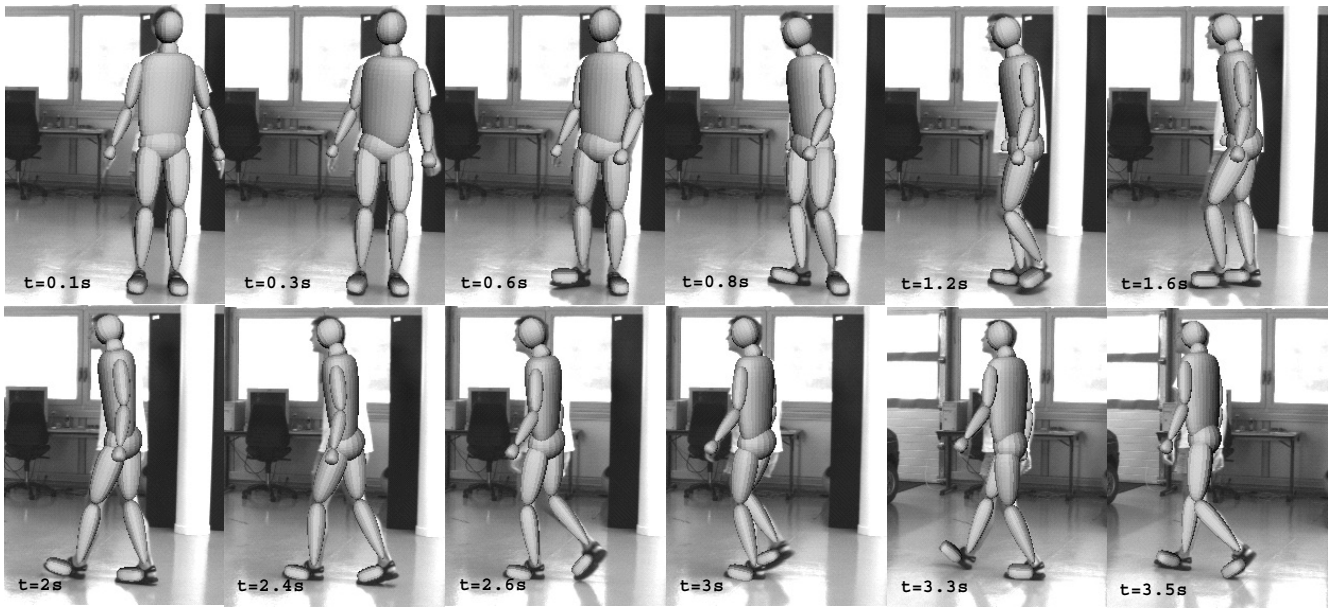


Fig. 10. Human tracking against a cluttered background. See Figure 14 for details.

controlled monocular 3D case. We used a parametric Gaussian mixture representation, local descent for mode refinement, as in Heap and Hogg (1998) and Cham and Rehg (1999) and a process model based on constant velocity plus dynamical noise sampling as in Cham and Rehg (1999; Section 3.2, p. 3), on the cluttered full body tracking sequence. However, note that unlike the original methods, ours uses robust (rather than least-squares) image matching and robust optimization by default, and also incorporates physical constraints and model priors. We used ten modes to represent the distribution over our 30-DOF 3D configurations, whereas Cham and Rehg (1999) used ten for their 38-DOF 2D SPM model. Our first set of experiments used a non-robust SSD image matching metric and a Levenberg–Marquardt routine for local sample optimization, as in Cham and Rehg (1999), except that we use analytical Jacobians. With this cost function, we find that outliers cause large fluctuations, bias and frequent convergence to physically invalid configurations. Registration is lost early in the turn (0.5 s), as soon as the motion becomes significantly non-fronto-parallel. Our second experiments used our robust cost function and optimizer, but still with sampling as in Cham and Rehg (1999). The track survived further into the turn, but was lost at 0.7 s when the depth variation became larger. As expected, we find that a dynamical noise large enough to provide sufficiently deep sampling along uncertain in-depth directions produces much too deep sampling along well-controlled transversal ones, so that most of the samples are lost on uninformative high-cost configurations. Similar arguments apply to standard CONDENSATION, as can be seen in the monocular 3D experiments of Deutscher, Blake, and Reid (2000).

6.2. Black Background Sequence

In this experiment we focus on 3D errors, in particular depth ambiguities and the influence of physical constraints and

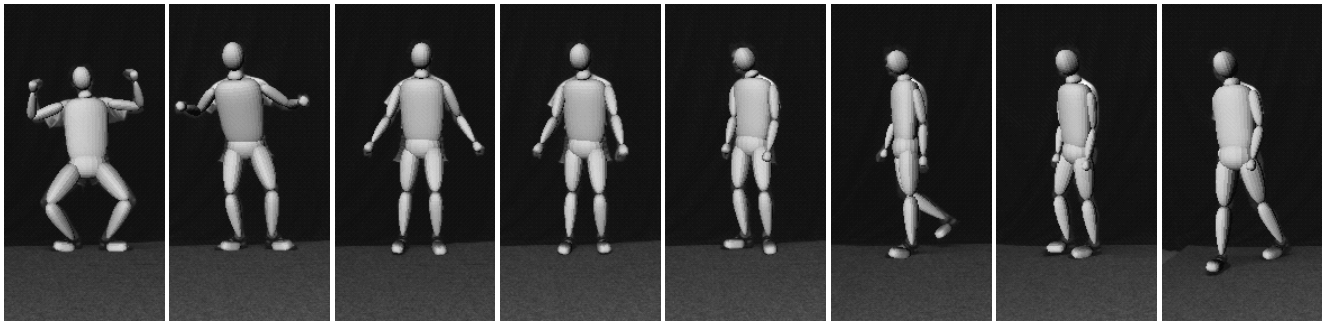


Fig. 11. Human tracking under complex motion. See Figure 15 for details.

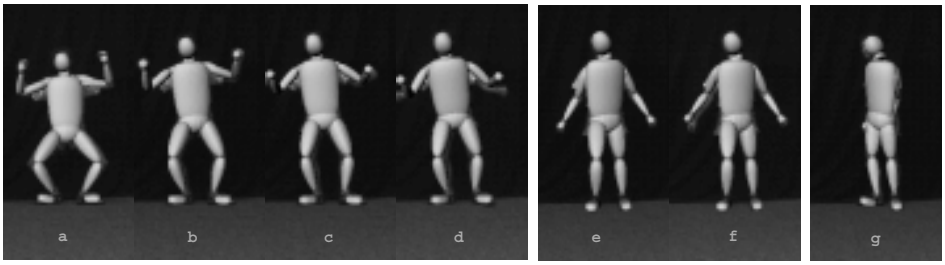


Fig. 12. Failure modes of various components of the tracker (see text).

parameter stabilization priors. We use an improved body model with 34 DOF. The four extra parameters control the left and right clavicle joints in the shoulder complex, which we find to be essential for following many arm motions. Snapshots from the full 4 s sequence are shown in Figure 11, and various failures modes in Figure 12.

The *Gaussian single mode tracker* manages to follow near-fronto-parallel motions fairly reliably owing to the absence of clutter, but it eventually loses track after 0.5 s (Figures 12(a)–(d)). The *robust single mode tracker* tracks the non-fronto-parallel motion slightly longer (about 1 s), although it significantly misestimates the depth (Figures 12(e) and (f); the right leg and shoulder are pushed much too far forward and the head is pushed forward to match subject contour, cf. the "correct" pose in Figure 11). It eventually loses track during the turn. The *robust multimode tracker with joint limits* is able to track quite well but, as body non-self-intersection constraints are not enforced, the modes occasionally converge to physically infeasible configurations (Figure 12(g)) with terminal consequences for tracking. Finally, the *robust fully constrained multimode tracker* is able to deal with significantly more complex motions and tracks the full sequence without failure (Figure 11).

7. Sampling Distributions

We also ran some more quantitative experiments aimed at studying the behavior of the different sampling regimes, particularly the efficiency with which they locate minima or low-cost regions of parameter space. We are interested in how the sampling distribution, as characterized by the shape of its core and the width of its tails, impacts the search efficiency. For the study here we used the simple, but still highly multimodal, 3D joint to image joint likelihood surface that we use for initializing our 34-DOF articulated model. We only estimated joint parameters, not body dimensions. We ran experiments involving CSS and spherical sampling (SS) for Gaussian distributions with scalings 1, 2 and 8. To allow a fair comparison, at each scale we kept the volume of the sphere (proportional to R^n) equal to the volume of the corresponding rescaled unit covariance CSS ellipsoid (proportional to $\lambda_1 \ldots \lambda_n$, the product of eigenvalues). Also note that the final sampling distributions are not exactly Gaussian—in fact, they are often noticeably multimodal—because our sampler preserves the physical constraints by projecting inadmissible samples back onto the constraint surface. Once made, the samples are locally optimized subject to the physical constraints using the method of Section 4.1. In Table 1, we report on the number of minima found by each method, and the medians and standard deviations of their parameter space distances and cost differences. Figure 13 shows distributions of numbers of samples and minima versus parameter space distance, standard deviation and cost, for scaling 8. Note that CSS finds significantly more minima, and also places samples at positions of significantly lower cost, than SS. We can also see the large cost difference between optimized and unoptimized samples. SS appears to

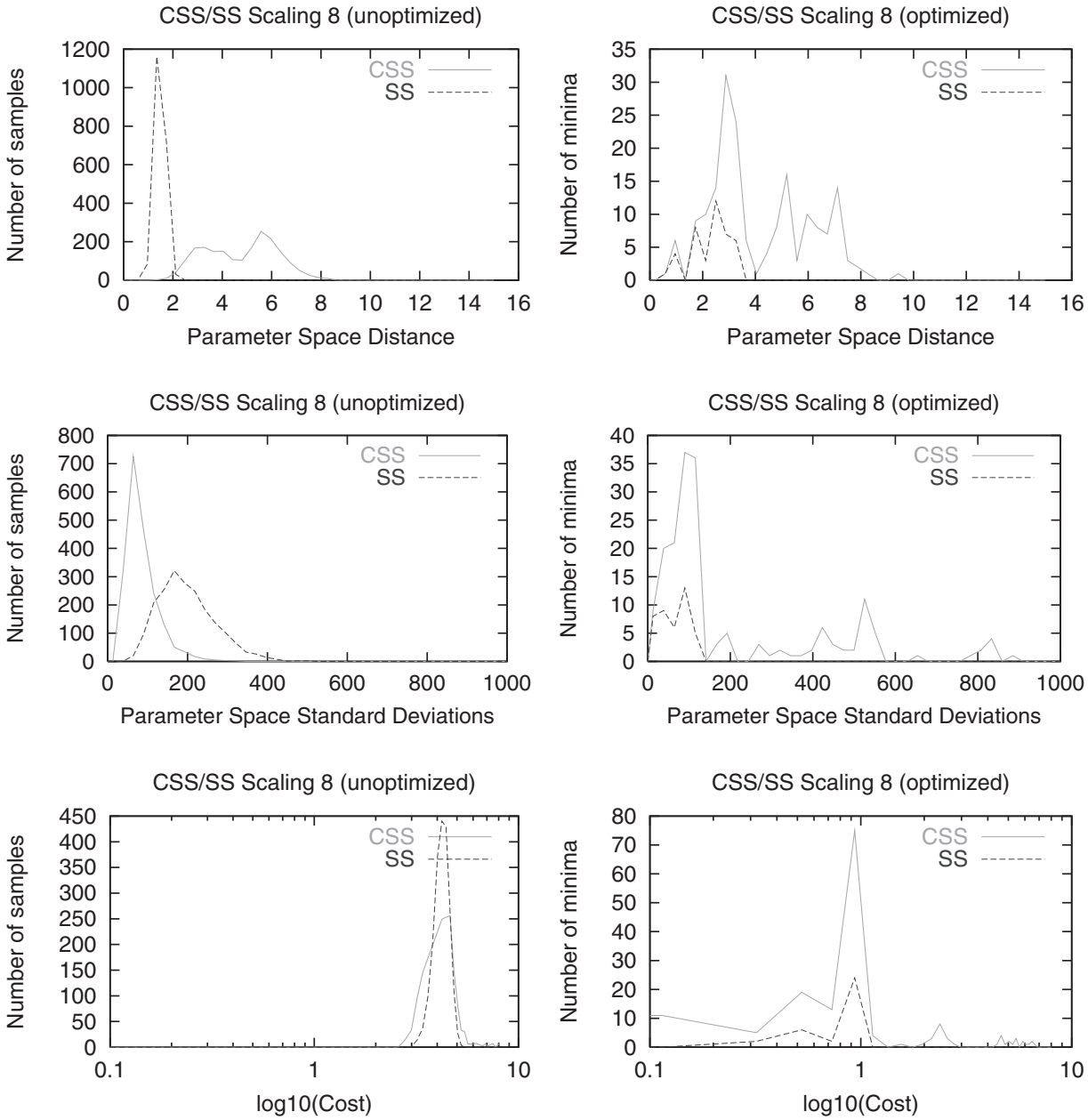


Fig. 13. Optimized and unoptimized sample statistics for SS and CSS with scaling factor 8 and runs of 2000 samples. Note the significantly larger number of minima found by CSS than by SS, and also that the samples are placed at much lower cost.

find minima of slightly lower median cost than CSS, but this is misleading. CSS still finds the few minima found by SS, but it also finds many other more distant ones, which, being further away, tend to increase its median cost.

8. Approximation Accuracy

The tracking experiments in Section 6 have illustrated the practical behavior and failure modes of some of the components of the CSS algorithm, and in Section 7 we have presented a more quantitative evaluation. Now we turn to more technical points.

The CSS algorithm involves both local continuous optimization and more global covariance-scaled sampling. It therefore has a natural mechanism to trade off speed and robustness. When tracking fails, both the number of modes used to represent the distribution and the number of samples produced in the sampling stage can be increased. This increases the computational cost, but it may allow the tracker to follow



Fig. 14. Clutter human tracking sequence detailed results for the CSS algorithm in Section 6.

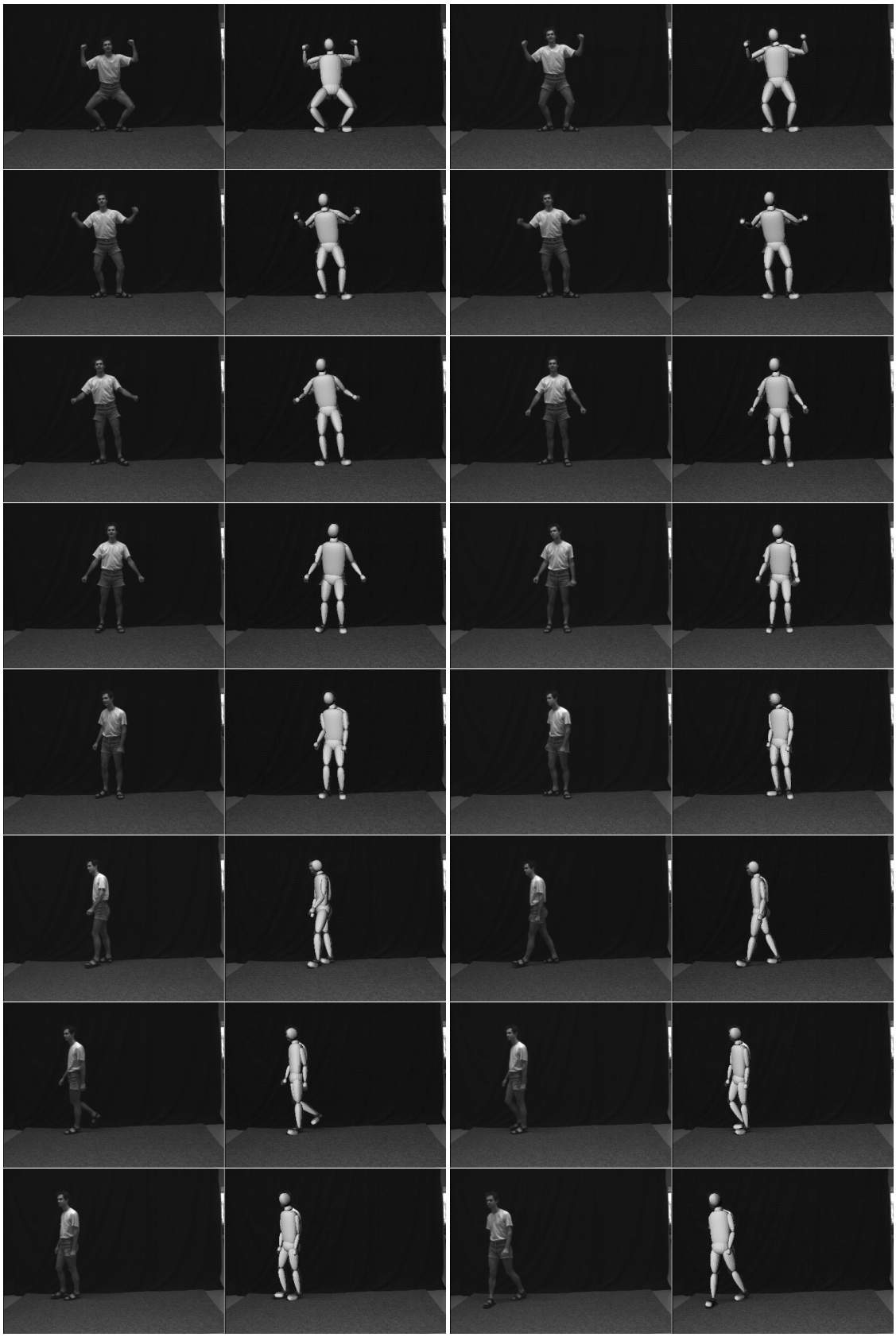


Fig. 15. Complex motion tracking sequence detailed results for the CSS algorithm in Section 6.

more difficult portions of the image sequence. In principle, a sufficiently long run of any sampling method would visit every region of the parameter space, so that the basin of attraction of each mode was sampled and all minima were found. It has been argued that mixed continuous/discrete trackers (Heap and Hogg 1998; Cham and Rehg 1999) will "diverge" if the visual information is ambiguous and converge to a "best" mode when the target in the image is easily detectable. However, this kind of divergence is not that important here. We are working with likelihood surfaces that have multiple peaks with individual probabilities. Local optimization methods can converge to any of these peaks and sampling methods will eventually "condense" near them if they use enough samples. Given the sampling/dynamics stage, both methods have a chance of jumping between peaks (i.e., escaping spurious ones), although this may be a very slow process. The method presented here is designed to address the problems of more efficient and systematic multimodal exploration. Note also that CSS can be viewed as an importance sampling distribution and correction weighting for fair sample generation can be performed with respect to the true prior.

A second issue concerns the algorithm's efficiency versus its bias behavior. In tracking, assuming temporal coherence, we may want to confine the search to the neighborhood of the configurations propagated from the previous tracking step. This can be done implicitly by designing a likelihood surface that emphasizes local responses,[13] or by tuning the search process for locality, using short-range dynamics. In either case, a global estimate of the posterior distribution is too expensive, both for sampling and optimization, while restricting attention to nearby states carries the risk of missing distant but significant peaks.

A third issue concerns the approximation accuracy of a Gaussian mixture for arbitrary multimodal distributions. The mixture model is likely to be inaccurate away from the mode cores, and this may affect the accuracy of statistical calculations based on it. However, for tracking and localization applications we are mainly interested in the modes themselves, not the low-probability regions in their remote tails. Sampling methods are non-parametric so in principle they do not have this limitation, but in practice so few samples fall deep in the tails that noisiness of the estimates is a problem. Pure sample-based representations also provide little insight into the structure of the uncertainty and the degree of multimodality of the likelihood surface. In any case, the issue of approximation accuracy in low-probability regions is not the main concern here. Provided initial seeds are available in the individual mode's basins of attraction, sampling methods can generate fair samples from the modes and optimization methods can precisely identify their means and covariances by local descent. The two techniques can be used interchangeably, depending on the application. It is the process of finding the initial seeds for each mode that represents the major difficulty for high-dimensional multimodal distributions.

A fourth and important practical issue concerns the properties of the likelihood function. For many complex models a good likelihood is difficult to build, and that used may be a poor reflection of the desired observation density. In these situations, the strength of true and spurious responses is similar and this may affect the performance of the tracking algorithm, irrespective how much computational power is used. In such contexts, it can be very difficult to identify the true tracked trajectory in a temporal flow of spurious responses. This is a particularly complex problem, since many likelihoods commonly used in vision degrade ungracefully under occlusion/disocclusion events and viewpoint change. At present, we do not have good mechanisms for detecting disocclusion events in complex backgrounds. The CSS algorithm has an elegant mechanism that accounts for high-uncertainty if particular DOF are not observed (such as occluded limbs, etc.) and it will automatically sample broadly there. However, for sub-sequences with long occlusion events, it is still likely to attach occluded limbs to background clutter, rather than maintaining them as occluded. Global silhouettes, or a human contour detector (Papageorgiu and Poggio 1999), or higher-order matching consistency (Sminchisescu 2002a) may help here. As an indication of the potential benefits of this, we currently use foreground/background segmentation and the motion boundaries from the robust optical flow computation to weight the importance of contours, and this significantly improves the results in the sequences we have analyzed.

9. Conclusions and Future Work

We have presented a new method for monocular 3D human body tracking, based on optimizing a robust model-image matching cost metric combining robustly extracted edges, flow and motion boundaries, subject to 3D joint limits, non-self-intersection constraints, and model priors. Optimization is performed using CSS, a novel high-dimensional search strategy based on sampling a hypothesis distribution followed by robust constraint-consistent local refinement to find a nearby cost minima. The hypothesis distribution is determined by propagating the posterior at the previous time step (represented as a Gaussian mixture defined by the observed cost minima and their Hessians/covariances) through the assumed dynamics (here trivial) to find the prior at the current time step, then inflating the prior covariances and resampling to scatter samples more broadly. Our experiments on real sequences show that this is significantly more effective than using inflated dynamical noise estimates as in previous approaches, because it concentrates the samples on low-cost points, rather than points that are simply nearby irrespective

13. For example, an optical flow correspondence field, like that described in Section 3.1 but based on least-squares brightness matching, can behave as a locality prior, forcing local image velocity explanations and pruning away remote, potentially "objective" multimodality.

of cost. In future work, it should also be possible to extend the benefits of CSS to CONDENSATION by using inflated (diluted weight) posteriors and dynamics for sample generation, then re-weighting the results. Our human tracking work will focus on incorporating better pose and motion priors as well as designing likelihoods that are better adapted for human localization in the image.

Acknowledgments

This work was supported by an EIFFEL doctoral grant and the European Union under FET-Open project VIBES. We would like to thank Alexandru Telea for stimulating discussions and implementation assistance and Frédérick Martin for helping with the video capture and posing as a model.

References

Barr, A. 1984. Global and local deformations of solid primitives. *Computer Graphics* 18:21–30.

Barron, C., and Kakadiaris, I. 2000. Estimating anthropometry and pose from a single image. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 669–676.

Black, M. 1992. Robust Incremental Optical Flow. PhD thesis, Yale University.

Black, M., and Anandan, P. 1996. The robust estimation of multiple motions: parametric and piecewise smooth flow fields. *Computer Vision and Image Understanding* 6(1):57–92.

Blake, A., North, B., and Isard, M. 1999. Learning multi-class dynamics. *Advances in Neural Information Processing Systems* 11:389–395.

Brand, M. 1999. Shadow puppetry. In *IEEE International Conference on Computer Vision*, pp. 1237–1244.

Bregler, C., and Malik, J. 1998. Tracking people with twists and exponential maps. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Cham, T., and Rehg, J. 1999. A multiple hypothesis approach to figure tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 239–245.

Choo, K., and Fleet, D. 2001. People tracking using hybrid Monte Carlo filtering. In *IEEE International Conference on Computer Vision*.

Delamarre, Q., and Faugeras, O. 1999. 3D articulated models and multi-view tracking with silhouettes. In *IEEE International Conference on Computer Vision*.

Deutscher, J., Blake, A., and Reid, I. 2000. Articulated body motion capture by annealed particle filtering. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Deutscher, J., Davidson, A., and Reid, I. 2001. Articulated partitioning of high dimensional search spacs associated with articulated body motion capture. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Deutscher, J., North, B., Bascle, B., and Blake, A. 1999. Tracking through singularities and discontinuities by random sampling. In *IEEE International Conference on Computer Vision*, pp. 1144–1149.

Drummond, T., and Cipolla, R. 2001. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *IEEE International Conference on Computer Vision*.

Fletcher, R. 1987. Practical Methods of Optimization, Wiley, New York.

Gavrila, D., and Davis, L. 1996. 3D model based tracking of humans in action: a multiview approach. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 73–80.

Gonglaves, L., Bernardo, E., Ursella, E., and Perona, P. 1995. Monocular tracking of the human arm in 3D. In *IEEE International Conference on Computer Vision*, pp. 764–770.

Gordon, N., and Salmond, D. 1995. Bayesian state estimation for tracking and guidance using the bootstrap filter. *Journal of Guidance, Control and Dynamics*.

Gordon, N., Salmond, D., and Smith, A. 1993. Novel approach to non-linear/non-Gaussian state estimation. *IEE Proceedings F*.

Hanim-Humanoid Animation Working Group. 2002. Specifications for a Standard Humanoid. http://www.h-anim.org/Specifications/H-Anim1.1/.

Heap, T., and Hogg, D. 1998. Wormholes in shape space: tracking through discontinuities changes in shape. In *IEEE International Conference on Computer Vision*, pp. 334–349.

Howe, N., Leventon, M., and Freeman, W. 1999. Bayesian reconstruction of 3D human motion from single-camera video. *Neural Information Processing Systems*.

Isard, M., and Blake, A. 1998. CONDENSATION—Conditional density propagation for visual tracking. *International Journal of Computer Vision*.

Ju, S., Black, M., and Yacoob, Y. October 1996. Cardboard people: a parameterized model of articulated motion. In *2nd Int. Conf. on Automatic Face and Gesture Recognition*, pp. 38–44.

Kakadiaris, I., and Metaxas, D. 1996. Model-based estimation of 3D human motion with occlusion prediction based on active multi-viewpoint selection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 81–87.

Lee, H.J., and Chen, Z. 1985. Determination of 3D human body postures from a single view. *Computer Vision, Graphics and Image Processing*, 30:148–168.

Liu, J. 1996. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6.

MacCormick, J., and Blake, A. 1998. A probabilistic contour discriminant for object localisation. In *IEEE International Conference on Computer Vision*.

MacCormick, J., and Isard, M. 2000. Partitioned sampling, articulated objects, and interface-quality hand tracker. In *European Conference on Computer Vision*, Vol. 2, pp. 3–19.

Merwe, R., Doucet, A., Freitas, N., and Wan, E. May 2000. The unscented particle filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University, Department of Engineering.

Papageorgiu, C., and Poggio, T. 1999. Trainable pedestrian detection. In *International Conference on Image Processing*.

Pitt, M., and Shephard, N. 1997. Filtering via simulation: auxiliary particle filter. *Journal of the American Statistical Association*.

Plankers, R., and Fua, P. 2001. Articulated soft objects for video-based body modeling. In *IEEE International Conference on Computer Vision*, pp. 394–401.

Rehg, J., and Kanade, T. 1995. Model-based tracking of self occluding articulated objects. In *IEEE International Conference on Computer Vision*, pp. 612–617.

Rohr, K. 1994. Towards model-based recognition of human movements in image sequences. *Computer Vision, Graphics and Image Processing* 59(I):94–115.

Rosales, R., and Sclaroff, S. 2000. Inferring body pose without tracking body parts. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 721–727.

Sidenbladh, H., and Black, M. 2001. Learning image statistics for Bayesian tracking. In *IEEE International Conference on Computer Vision*.

Sidenbladh, H., Black, M., and Fleet, D. 2000. Stochastic tracking of 3D human figures using 2D image motion. In *European Conference on Computer Vision*.

Sidenbladh, H., Black, M., and Sigal, L. 2002. Implicit probabilistic models of human motion for synthesis and tracking. In *European Conference on Computer Vision*.

Sminchisescu, C. 2002a. Consistency and coupling in human model likelihoods. In *IEEE International Conference on Automatic Face and Gesture Recognition*, Washington DC, pp. 27–32.

Sminchisescu, C. July 2002b. Estimation algorithms for ambiguous visual models—three-dimensional human modeling and motion reconstruction in monocular video sequences. PhD thesis, Institute National Politechnique de Grenoble (INRIA).

Sminchisescu, C., Metaxas, D., and Dickinson, S. 2001. Improving the scope of deformable model shape and motion estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Hawaii, Vol. 1, pp. 485–492.

Sminchisescu, C., and Telea, A. 2002. Human pose estimation from silhouettes. A consistent approach using distance level sets. In *WSCG International Conference for Computer Graphics, Visualization and Computer Vision*, Czech Republic.

Sminchisescu, C., and Triggs, B. 2001a. A robust multiple hypothesis approach to monocular human motion tracking. Technical Report RR-4208, INRIA.

Sminchisescu, C., and Triggs, B. 2001b. Covariance-scaled sampling for monocular 3D body tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Hawaii, Vol. 1, pp. 447–454.

Sminchisescu, C., and Triggs, B. 2002a. Building roadmaps of local minima of visual models. In *European Conference on Computer Vision*, Copenhagen, Vol. 1, pp. 566–582.

Sminchisescu, C., and Triggs, B. 2002b. Hyperdynamics importance sampling. In *European Conference on Computer Vision*, Copenhagen, Vol. 1, pp. 769–783.

Sminchisescu, C., and Triggs, B. 2003. Kinematic jump processes for monocular 3D human tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Taylor, C.J. 2000. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 677–684.

Triggs, B., McLauchlan, P., Hartley, R., and Fitzgibbon, A. 2000. Bundle adjustment—A modern synthesis. In *Vision Algorithms: Theory and Practice*, Springer-Verlag, Berlin.

Vanderbilt, D., and Louie, S.G. 1984. A Monte Carlo simulated annealing approach over continuous variables. *Journal of Computational Physics* 56.

Wachter, S., and Nagel, H. 1999. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding* 74(3):174–192.

Zhu, S. C., and Mumford, D. 1997. Learning generic prior models for visual computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(11).