Covariance Scaled Sampling for Monocular 3D Body Tracking

Cristian Sminchisescu **Bill Triggs**
GRAVIR, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot, France
{*Cristian.Sminchisescu,Bill.Triggs*}*@inrialpes.fr, www.inrialpes.fr/movi/people/*{*Sminchisescu,Triggs*}

Abstract

We present a method for recovering 3D human body motion from monocular video sequences using robust image matching, 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: for reliable tracking at least 30 joint parameters need to be estimated, subject to highly nonlinear physical constraints; the problem is chronically ill-conditioned as about 1/3 of the d.o.f. (the depth-related ones) are almost unobservable in any given monocular image; and matching an imperfect, highly flexible, self-occluding model to cluttered image features is intrinsically hard. To reduce correspondence ambiguities we use a carefully designed robust matching-cost metric that combines robust optical flow, edge energy, and motion boundaries. Even so, the ambiguity, nonlinearity and non-observability make the parameter-space cost surface multi-modal, unpredictable and ill-conditioned, so minimizing it is difficult. We discuss the limitations of CONDENSATION*-like samplers, and introduce a novel hybrid search algorithm that combines inflated-covariance-scaled sampling and continuous optimization subject to physical constraints. Experiments on some challenging monocular sequences show that robust cost modelling, joint and self-intersection constraints, and informed sampling are all essential for reliable monocular 3D body tracking.*

Keywords: *3D human body tracking, particle filtering, high-dimensional search, constrained optimization, robust matching.*

1 Introduction

Extracting 3D human motion from natural monocular video sequences poses difficult modelling 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 joint limits and non-self-intersection constraints. (*ii*) Unlike the 2D and multi-camera 3D cases, in any given monocular image about 1/3 of the degrees of freedom are nearly unobservable (mainly motions in (relative) depth, but also rotations of near-cylindrical limbs about their axes). (*iii*) Matching a complex, imperfectly known, self-occluding model to a cluttered scene is inherently hard. These difficulties interact: minor body modelling or feature matching errors often lead to large compensatory biases in estimated depths, which eventually cause mis-prediction and tracking failure.

To appear IEEE CVPR 2001

We believe that a successful monocular 3D body tracking system must pay attention to each of these three difficulties. We control correspondence errors with a carefully designed robust matching metric that combines robust optical flow, edge energy, and motion boundaries (§3). Our system is the first to enforce both hard joint angle limits and body non-self-intersection constraints, and also includes full 3D occlusion prediction. The various ambiguities and nonlinearities make the parameter-space cost function multi-modal, ill-conditioned and highly nonlinear, so some form of non-local search is required. Existing approaches that we are aware of (§1.2) do not work well in this context, so we introduce a novel hybrid search scheme that combines covariance-scaled 'oversized' sampling with local optimization subject to joint and non-self-intersection constraints (§4). We finish with experimental results on some challenging monocular sequences, that illustrate the need for each of robust cost modelling, joint and self-intersection constraints, and well-controlled sampling plus local optimization.

1.1 High-Dimensional Search 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 Taylor models to predict good search directions [6, 23, 18, 28, 22]; **regular sampling** evaluates the cost function at a predefined pattern of points in (a slice of) parameter space, *e.g.* a local rectangular grid [11]; and **stochastic sampling** generates random sampling points according to some hypothesis distribution encoding "good places to look" [9, 25]. Densely sampling the entire parameter space would guarantee a good solution but is infeasible in more than 2–3 dimensions. In 30 dimensions any feasible sample must be extremely sparse and hence likely to miss significant cost minima. Descent methods at least (at some expense) find *local* minima, but can not guarantee global optimality. Our method tries to balance local and global effort using a combination of carefully controlled sampling and local optimization. Effective focusing of effort is the key to high-dimensional search. This is an active research area [9, 14, 7], but no existing method can guarantee a global minimum.

During tracking, the search method is applied time-

Meta 1029
p. 1

recursively, the starting point(s) for the current search being obtained from the optimized 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 realistic statistical models, Bayes-law propagation of a probability density for the true state is possible. For linearized monomodal dynamics and observation models under Gaussian noise, this leads to (Extended) Kalman Filtering. For likelihood-weighted random sampling under general multimodal observation models, CONDENSATION results. In both cases the various hyperparameters must be carefully tuned for good performance. Visual tracking usually works in the 'shotgun in the dark' regime: observation likelihoods are quite sharply peaked but multimodal, so to avoid mistracking, the dynamical noise has to be turned up until it produces a scatter of samples just big enough to cover typically-nearby peaks. In this regime there is negligible trajectory smoothing so Kalman-style covariance updating is superfluous: the previous posterior determines the locations and weights of the search regions, the dynamical noise determines their breadth, and the observation likelihood determines the location and shape of the new posterior peak(s) within each region.

Many existing methods use inflated dynamical noise as an empirical search focusing parameter [7,14,9], but we find that it produces poorly shaped search regions. An efficient high-dimensional search must adapt to the local cost surface. Rather than inflating the dynamical noise, we will argue that one should use realistic dynamics, then modestly inflate the resulting *prior* (previous posterior after dynamics) covariance to define the search region. This inflates the posterior uncertainty as well as the dynamical one, allowing far deeper sampling along the most uncertain directions (*e.g.* poorly observable depth d.o.f.), and thus preventing mistracking due to inadequate exploration of these hard-to-estimate parameter combinations. This simple change makes a huge difference in practice. For example, for the 32 d.o.f. cost spectrum in fig. 3 with 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), the uniform dynamical noise method would produce a search volume 10^{54} times larger than that of our prior-based one.

1.2 Previous Work

We will 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 [7,17] or multi-camera 3D [18,11,6,22] tracking and surprisingly few works have addressed it [9,25,28,15,5]. The main additional difficulty is the omnipresence of depth ambiguities. Every limb or body segment lying near a frontoparallel plane has a first-order observability singularity: small rotations towards or away from the camera leave the image unchanged to first order. Similarly, finite towards- and away-from-camera rotations give very similar images, so even if the segment matching cost is monomodal in the image, it is always multimodal in parameter space. To handle these difficulties, time integration or additional domain constraints such as joint limits and body non-self-intersection must be incorporated.

Deutscher [9] uses a sophisticated 'annealed sampling' strategy to speed up CONDENSATION, but for his main sequence uses 3 cameras and a black background. Sidenbladh [25] uses a similar importance sampling technique with a strong learned prior walking model to track a walking person in an outdoor sequence. Our method does not yet include a 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 [4]. Partitioned sampling [20] is another notable sampling technique for articulated models, under certain labelling assumptions [20,9].

Heap & Hogg [14] and Cham & Rehg [7] combine CONDENSATION-style sampling with local optimization, but they consider only the simpler case of 2D tracking. Cham & 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 obtained from the Hessians at the fitted optima, as in our method, but the search region widths are controlled by the traditional method of adding a large dynamical noise. This appears to work reasonably well for 2D SPM tracking, which is essentially free of observability singularities. But we find (§5) that it can not handle the much less well-conditioned monocular 3D case. One puzzling point in [7] is the presence of closely-spaced minima with overlapping peaks, which motivated Cham & Rehg to introduce their piecewise Gaussian distribution model. We do not observe such overlaps, and we suspect that they were caused by incomplete convergence in the optimizer, presumably due to either over-loose convergence criteria, or a noisy cost function (we took considerable pains to make ours smooth).

2 Human Body Model

Our human body model (fig. 1a,b) consists of kinematic 'skeletons' of articulated joints controlled by angular joint parameters covered by 'flesh' built from superquadric ellipsoids with additional tapering and bending parameters [1]. A typical model has about 30 **joint parameters** $\mathbf{x}_a$; 8 **internal proportion** parameters $\mathbf{x}_i$ encoding the positions of the hip, clavicle and skull tip joints; and 9 **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

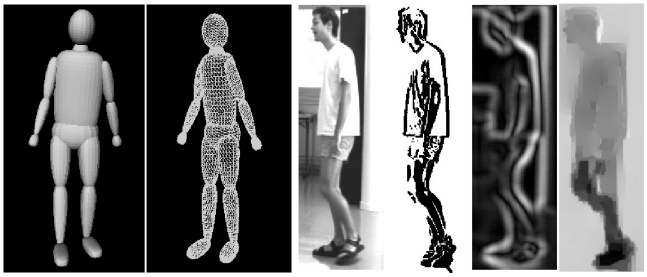

Figure 1: Two views of our human body model, and examples of our robust low-level feature extraction: original image (c), motion boundaries (d), intensity-edge energy (e), and robust horizontal flow field (f).

$\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_d, \mathbf{x}_i)$. During tracking we usually estimate only joint parameters, but our initialization method also estimates the most important internal proportions and shape parameters, subject to a soft prior based on standard humanoid dimensions from [13] updated using collected image evidence. Although far from photorealistic, this model suffices for high-level interpretation and realistic occlusion prediction, and offers 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 $\mathbf{r}_i(\mathbf{x}) = P(\mathbf{p}_i(\mathbf{x})) = P(A(\mathbf{x}_a, \mathbf{x}_i, D(\mathbf{x}_d, \mathbf{u}_i)))$, where 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 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 $\mathbf{\Delta r}_i(\mathbf{x}) = \bar{\mathbf{r}}_i - \mathbf{r}_i(\mathbf{x})$.

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:

$$p(\mathbf{x}|\bar{\mathbf{r}}) \;\propto\; p(\bar{\mathbf{r}}|\mathbf{x})\, p(\mathbf{x}) \;=\; \exp\left(-\int e(\bar{\mathbf{r}}_i|\mathbf{x})\, di\right)\, p(\mathbf{x}) \qquad (1)$$

where $e(\bar{\mathbf{r}}_i|\mathbf{x})$ is the cost density associated with observation i, the integral is over all observations, and $p(\mathbf{x})$ is the prior on the model parameters. Discretizing the continuous problem, our MAP approach minimizes the negative log-likelihood for the total posterior probability:

$$f(\mathbf{x}) \;=\; -\log p(\bar{\mathbf{r}}|\mathbf{x}) - \log p(\mathbf{x}) \;=\; f_l(\mathbf{x}) + f_p(\mathbf{x})$$

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. Feature contributions are fused using robust (heavy-tailed) error distributions, *i.e.* 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.

Robust Error Distributions: MAP parameter estimation is naturally robust so long as it is based on realistic 'total likelihoods' for the combined inlier and outlier distributions of the observations. We model these as robust penalty functions $\rho_i(s_i)$ of the normalized squared errors $s_i = \|\mathbf{\Delta r}_i\|^2/\sigma_i^2$. Each $\rho_i(s)$ is an increasing sublinear function with $\rho_i(0) = 0$ and $\frac{d}{ds}\rho_i(0) = 1$, corresponding to a radially symmetric error distribution with a central peak of width σ. Here we used the 'Lorentzian' $\rho(s) = \nu \log(1 + s/\nu)$ and 'Leclerc' $\rho(s) = \nu(1 - \exp(-s/\nu))$ potentials, where ν is a strength parameter related to the frequency of outliers.

Normalizing by the number of nodes N in each mesh, the cost adopted for the ith observation is $e(\bar{\mathbf{r}}_i|\mathbf{x}) = \frac{1}{N} e_i(\mathbf{x})$, where $\mathbf{W}_i$ is a positive definite weighting matrix and:

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

The total robust observation likelihood is thus:

$$f_l(\mathbf{x}) \;=\; -\log p(\bar{\mathbf{r}}|\mathbf{x}) \;=\; f_a(\mathbf{x}) + N_{bf}\, \nu_{bf} + N_{occ}\, \nu_{occ} \qquad (2)$$

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.

Cue Integration and Assigned Image Descriptors: We use both edge and intensity features in our cost function. 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

Black's implementation [2] gives both a flow field and an associated outlier map. The outlier map conveys useful information about the motion boundaries and is used to weight the significance of edges (see fig. 1d). The motion boundaries are processed similarly to obtain a smooth image. 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, weighting them by their importance qualified by the motion boundary map. For visible nodes lying inside the object ($\mathcal{I}$), we use intensity information derived from the robust optical flow. The assigned data term (2) thus becomes:

$$f_a(\mathbf{x}) = \frac{1}{2} \sum_{i \in \mathcal{O}, e \in \mathcal{E}_i} \rho_{i_e}\left(\mathbf{\Delta r}_{i_e}(\mathbf{x})\, \mathbf{W}_{i_e}\, \mathbf{\Delta r}_{i_e}(\mathbf{x})^\top\right) + \frac{1}{2} \sum_{j \in \mathcal{I}} \rho_{j_f}\left(\mathbf{\Delta r}_{j_f}(\mathbf{x})\, \mathbf{W}_{j_f}\, \mathbf{\Delta r}_{j_f}(\mathbf{x})^\top\right)$$

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

3.2 Model Priors

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

Anthropometric data p_a: The internal proportions for a standard humanoid are collected from [13] and used effectively as a Gaussian prior, $p_a = \mathcal{N}(\boldsymbol{\mu}_a, \mathbf{\Sigma}_a)$, 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.

Parameter stabilizers p_s: Certain details are far more important than intuition would suggest. For example, it is impossible to track common turning and reaching motions unless the clavicle joints in the shoulder are modelled accurately. However, such parameters have fairly well defined equilibrium positions and leaving them unconstrained would often lead to ambiguities. We model them with Gaussian stabilizers around their equilibria, $p_s = \mathcal{N}(\boldsymbol{\mu}_s, \mathbf{\Sigma}_s)$.

Anatomical joint angle limits C_{jl}: 3D consistency requires that the values of joint angles evolve within anatomically consistent intervals. We model this with a set of inequalities of the form $C_{jl} \cdot \mathbf{x} < 0$, where C_{jl} is a constraint matrix.

Body part inter-penetration 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 each body part, $f_{pa} = \exp(-f(\mathbf{x})\,|f(\mathbf{x})|^{p-1})$ where $f(\mathbf{x})$ defines the implicit surface of the body part and p controls the decay rate.

3.3 Distribution Representation

We represent posterior distributions as sets of separate modes $m_i \in \mathcal{M}$, each having an associated probability, mean and covariance matrix $m_i = (c_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$. This can be viewed as a Gaussian mixture approximation. Cham & Rehg [7] use a similar model but need 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. Our modes result from running local continuous optimizations to convergence, so they are necessarily either well separated or confounded. Our sampling method is also significantly different from [7], as explained in §4.2.

3.4 Temporal Propagation

Equation (1) gives the model likelihood in a static image, under model priors but without initial state or temporal priors. Adding temporal models with observations $\mathbf{R}_t = \{\bar{\mathbf{r}}_1, \ldots, \bar{\mathbf{r}}_t\}$, the posterior distribution becomes:

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

Here $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the dynamical model and $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ is the prior distribution from $t-1$. Together they form the prior $p(\mathbf{x}_t|\mathbf{R}_{t-1})$ for the static image search (1).

4 Optimization Algorithm

Our search technique combines robust constraint-consistent local optimization and more global discrete sampling.

4.1 Robust Constrained Mode Search

The robustified gradient and Hessian corresponding to the model feature i with possible assignments $a \in \mathcal{A}$ can be derived using the model-image Jacobian $\mathbf{J}_i = \frac{\partial \mathbf{r}_i}{\partial \mathbf{x}}$:

$$\mathbf{g}_i = \mathbf{J}_i^\top \left(\sum_{a \in \mathcal{A}} \rho'_{i_a}\, \mathbf{W}_{i_a}\, \mathbf{\Delta r}_{i_a}\right)$$

$$\mathbf{H}_i \approx \mathbf{J}_i^\top \left(\sum_{a \in \mathcal{A}} \rho'_{i_a} \mathbf{W}_{i_a} + 2\rho''_{i_a} (\mathbf{W}_{i_a} \mathbf{\Delta r}_{i_a})(\mathbf{W}_{i_a} \mathbf{\Delta r}_{i_a})^\top\right) \mathbf{J}_i$$

The gradient and Hessian contributions from all observations are assembled, together with negative log prior contributions:

$$\mathbf{g} = \mathbf{g}_o + \nabla f_a + \nabla f_s + \nabla f_{pa}$$
$$\mathbf{H} = \mathbf{H}_o + \nabla^2 f_a + \nabla^2 f_s + \nabla^2 f_{pa}$$

We use a second order trust region method for local optimization. This chooses a descent direction by solving the regularized subproblem [10]:

$$(\mathbf{H} + \lambda\, \mathbf{W})\, \mathbf{\Delta x} = -\mathbf{g} \quad \text{subject to} \quad C_{jl}\, \mathbf{x} < 0$$

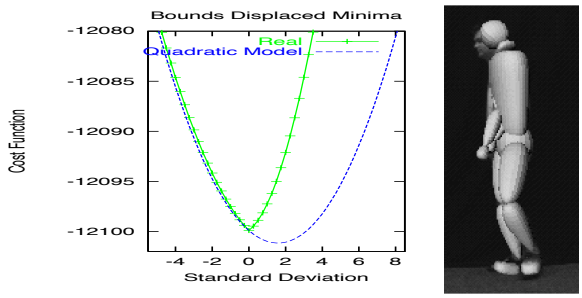

Figure 2: (a) Displaced minimum due to joint limits constraints, (b) Joint limits without body non-self-intersection constraints do not suffice for physical consistency.

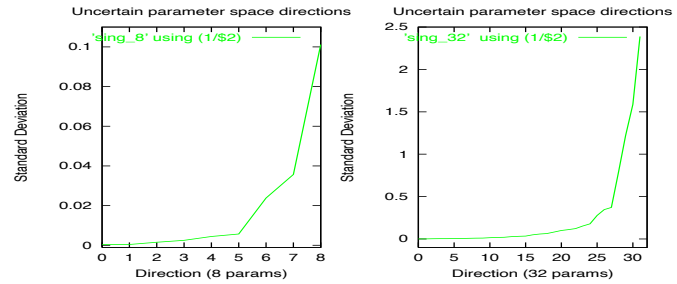

Figure 3: Typical covariance eigenvalue spectra. $\sigma_{\max}/\sigma_{\min}$ is 350 for the 8 d.o.f. arm model, 2000 for the 32 d.o.f. body one.

where $\mathbf{W}$ is a symmetric positive-definite matrix and λ is a dynamically chosen weighting factor. Joint limits C_{jl} are handled as hard bound constraints in the optimizer, by projecting the gradient onto the current active constraint set. Adding joint constraints changes the effective shape of the cost function and hence the minimum reached. Fig. 2 plots a 1D slice through the constrained cost function together with a second order Taylor expansion of the unconstrained cost. The gradient is nonzero at the constrained minimum owing to the presence of the bounds. The constrained cost gradient changes abruptly because active-set projection changes the motion direction during the slice to maintain consistency with the constraints.

4.2 Covariance Scaled Sampling

Although representations based on propagating multiple modes, hypotheses or samples do tend to increase the robustness of model estimation, the great difficulty with high-dimensional distributions is finding a sampleable proposal density that 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. Local optimization gives us not only local modes, but also their (robust, constraint consistent) Hessians and hence estimates of their local parameter estimation uncertainties. 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 modelling imperfections, 3D nonlinearities and constraints have the most influence, 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 unobservable motion-in-depth d.o.f. Some examples of such multimodal behaviour along high covariance eigen-directions are given in fig. 3. Also, it is seldom enough to sample at the scale of the estimated covariance — significantly deeper sampling is needed to capture nearby but non-overlapping modes lying further up the valley. Hence, we sample according to rescaled covariances, typically scaling up by a factor of around 10. One can sample either randomly or according to a regular pattern. Our current implementation samples regularly, in fact only along the lines corresponding to the lowest few covariance eigen-directions. Although this gives an exceedingly sparse sample, we find that it works well in practice.

Proposal Density for Mode $m_i = (c_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
1. Eigen-decompose $\boldsymbol{\Sigma}_i$, select its k most uncertain eigen-directions $\mathbf{v}_j$, and reconstitute the subspace covariance matrix $\boldsymbol{\Sigma}'_i = \sum_{j=1}^{k} \lambda_j \, \mathbf{v}_j \, \mathbf{v}_j^\top$.
2. The proposal density is $p_i \sim \mathcal{N}(\boldsymbol{\mu}_i, s\,\boldsymbol{\Sigma}'_i)$. The stretching factor s is 8–14 in our experiments.

Covariance Scaled Sampler
Until the desired number of samples are obtained:
1. Choose a mode m_i with probability c_i
2. Sample from m_i's proposal density p_i

Multiple-Mode Tracker
For each time-frame:
1. Starting from the above samples, generate samples $\mathbf{x}_i$ from $p(\mathbf{x}_t|\mathbf{R}_{t-1})$.
2. Refine each sample $\mathbf{x}_i$ using continuous optimization (§4.1) to obtain $(c_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Prune redundant samples converging to the same minimum.
3. Weight the samples by their prior likelihoods, assuming that they came from the closest (most probable *a posteriori*) prior mode. Prune to keep the best k modes, and renormalize the weights to compute c_i.

We have empirically studied the shape of the cost surface by sampling along uncertain directions for various model configurations. With our carefully selected image descriptors, the cost surface is smooth and our local optimizer reliably finds a local minimum. Multiple modes occur for certain configurations, as in fig. 4, which shows the two most uncertain modes of the fig. 6 human tracking sequence at times 0.8 s and 0.9 s.

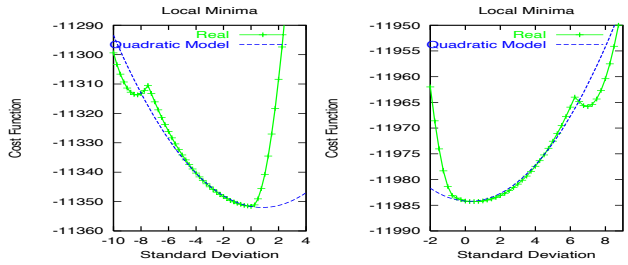

Figure 4: Multimodal behaviour along highest uncertainty eigen-directions (0.8 and 0.9 s in cluttered body tracking sequence).

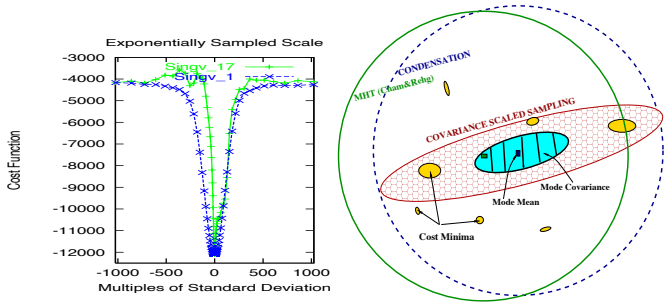

Figure 5: (a) Cost function slices at large scales, (b) Comparison of sampling methods: (1) CONDENSATION (dashed circle coverage) randomizes each sample by dynamic noise, (2) MHT ([7], solid circle) samples within covariance support (dashed ellipse) and applies the same noise policy as (1), finally, our (3) *Covariance Scaled Sampling* (pattern ellipse) targets good cost minima (flat filled ellipses) by inflating the highly uncertain subspace of the current sample robust covariance estimation (dashed ellipse))

We have also studied the cost surface at much larger scales in parameter space — see fig. 5a. Note that we recover the expected robust shape of the distribution, with some but not too many spurious local minima. Hence, the combination of our robust cost function and informed search is likely to be comparatively efficient computationally.

5 Experiments

To illustrate our method we show results for an 8 second arm tracking sequence and two full body ones (1.2 s and 4 s). All of these sequences contain both self-occlusion and significant relative motion in depth. The first two (fig. 6) were shot at 25 frames (50 fields) per second against a cluttered, unevenly illuminated background. The third (fig. 7) 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 evaluating the cost function. The figures overlay the current best candidate model on the original images.

Cluttered background sequences: These sequences explore 3D estimation behaviour with respect to image assignment and depth ambiguities, for a bending rotating arm under an 8 d.o.f. model and a pivoting full-body motion under a 30 d.o.f. 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 frontoparallel motions in moderate clutter, although it gradually slips out of registration when the arm passes the strong edges of the white pillar (0.5 s 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 frontoparallel motions reasonably well even in clutter, but quickly loses track during in-depth motions, which it tends to misinterpret as frontoparallel ones. In the arm tracking sequence, shoulder motion towards the camera is 'explained' as frontoparallel 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-frontoparallel 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 multi-mode tracker* correctly estimates the motion of the entire arm and body sequence. We retain just the 3 best modes found by sampling along the 3 most uncertain directions for the arm sequence, and the 7 best modes from the 6 most uncertain directions for the full human body sequence. As discussed in §4.2, multimodal behaviour occurs mainly during significantly non-frontoparallel motions, between 2.2–4.0 s for the arm sequence, and over nearly the whole full body sequence (0.2–1.2 s). For the latter, the modes mainly reflect the ambiguity between true pivoting motion and its incorrect "frontoparallel explanation".

We also compared our sampling method with a 3D version of Cham & Rehg's MHT [7] for the body turn sequence. (But the original method used non-robust optimization and did not incorporate physical constraints or model priors). We used 10 modes to represent the distribution in our 30 d.o.f. 3D model, as [7] used 10 for their 38 d.o.f. 2D SPM model. Our first set of experiments used a nonrobust SSD image matching metric and a Levenberg-Marquardt routine for local sample optimization, as in [7] (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-frontoparallel.

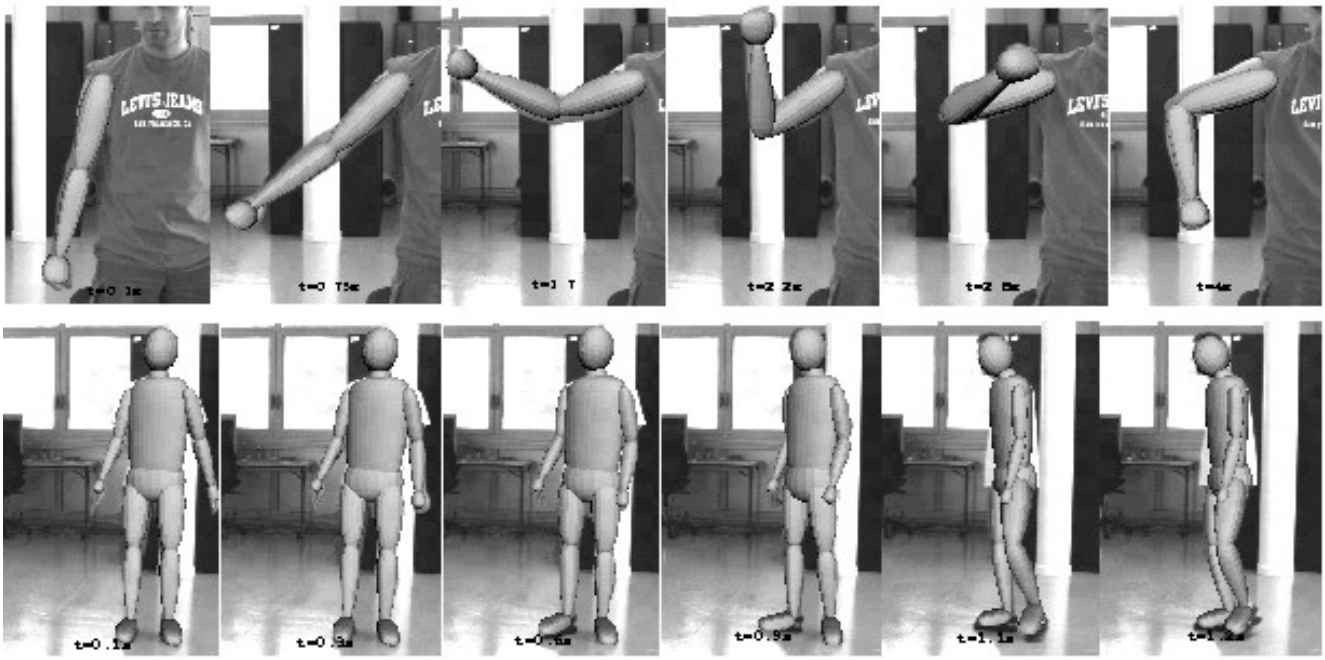

Figure 6: Arm tracking and full body tracking against a cluttered background.

Our second experiments used our robust cost function and optimizer, but still with MHT-style sampling. 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 usefully deep sampling along uncertain directions produces much too deep sampling along well-controlled ones, so that most of the samples are wasted on uninformative high-cost configurations. Similar arguments apply to standard CONDENSATION, as can be seen in the monocular 3D experiments of Deutscher [9].

Black background sequence: In this experiment we focus on 3D errors, in particular depth ambiguities and the influence of physical constraints and parameter stabilization priors. We use an improved body model with 34 d.o.f. 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 fig. 7, and various failures modes in fig. 8.

The *Gaussian single mode tracker* manages to follow near-frontoparallel motions fairly reliably owing to the absence of clutter, but it eventually loses track after 0.5 s (fig. 8 a-d). The *robust single mode tracker* tracks the non-frontoparallel motion somewhat longer (about 1 s, fig. 8 e-f), although it significantly mis-estimates the depth — the right leg, shoulder and head are too far forward compared to the "correct" pose in fig. 7 — and eventually loses track during the turn. The *robust multi-mode tracker with joint-limits* manages to track quite well, but as body non-self-intersection constraints are not enforced the modes eventually converge to physically infeasible configurations (fig. 8 g) with terminal consequences for tracking. Finally, the *robust fully constrained multi-mode tracker* is able to deal with significantly more complex motions and tracks the full sequence without failure (fig. 7).

6 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 Covariance Scaled Sampling, 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 combining the posterior at the previous time step (represented as a Gaussian mixture defined by the observed cost minima and their Hessians / covariances) and the assumed dynamics to find the current-time-step prior, then inflating the prior covariances to sample more broadly. Our experiments on real sequences show that this is significantly more effective than using inflated dynamical noise estimates as in previous approaches.

Future work will compare stochastic and regular sampling CSS and variant covariance scaled hypothesis distributions such as longer-tailed or coreless distributions. It should also be possible to extend the benefits of CSS to CONDENSATION by using inflated (diluted weight) posteriors and dynamics for

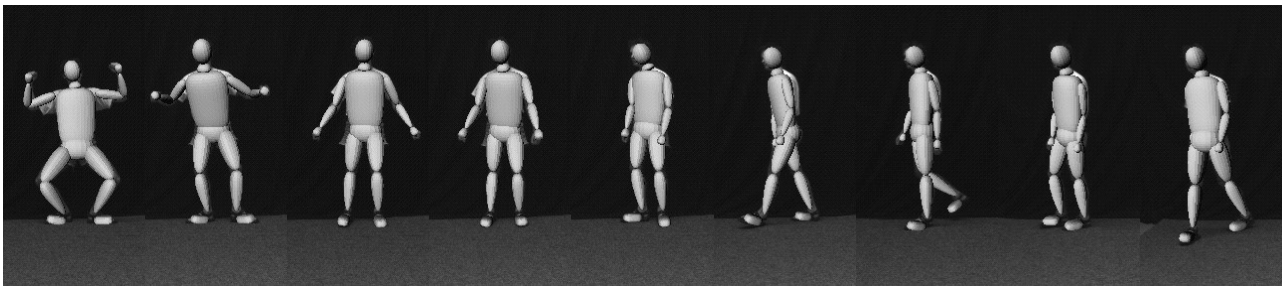

Figure 7: Human tracking under complex motion

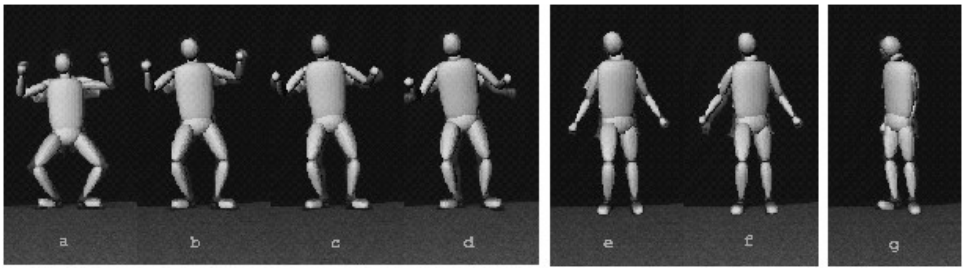

Figure 8: Various components failure modes

sample generation, then reweighting the results, *c.f.* [9]. Our human tracking work will focus on incorporating better pose and motion priors.

Acknowledgements

This work was supported by an EIFFEL doctoral grant and European Union 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

[1] A.Barr, "Global and local deformations of solid primitives," *Computer Graphics*, 18:21-30, 1984.

[2] M.Black and P.Anandan, "The Robust Estimation of Multiple Motions:Parametric and Piecewise Smooth Flow Fields," *CVIU*, Vol.63, No.1, pp.75-104, 1996.

[3] M.Black, A.Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *IJCV*, Vol. 19, No. 1, pp. 57-92, July, 1996.

[4] A.Blake, B.North and M.Isard, "Learning multi-class dynamics", *ANIPS 11*, pp.389-395, 1999.

[5] M.Brand, "Shadow Puppetry," *ICCV*, pp.1237-1244, 1999.

[6] C.Bregler and J.Malik, "Tracking People with Twists and Exponential Maps," *CVPR*, 1998.

[7] T.Cham and J.Rehg, "A Multiple Hypothesis Approach to Figure Tracking", *CVPR*, Vol.2, pp.239-245, 1999.

[8] J.Deutscher, B.North, B.Bascle, A.Blake, "Tracking through Singularities and Discontinuities by Random Sampling," *ICCV*, pp.1144-1149, 1999.

[9] J.Deutscher, A.Blake, I.Reid, "Articulated Body Motion Capture by Annealed Particle Filtering," *CVPR*, 2000.

[10] R.Fletcher, "Practical Methods of Optimization," *John Wiley*, 1987.

[11] D.Gavrila and L.Davis, "3-D Model Based Tracking of Humans in Action:a Multiview Approach," *CVPR*, pp. 73-80, 1996.

[12] D.Gavrila, "The Visual Analysis of Human Movement:A Survey," *CVIU*, Vol.73, No.1, pp.82-98, 1999.

[13] Hanim-Humanoid Animation Working Group, "Specifications for a standard humanoid", available at *http://www.h-anim.org/Specifications/H-Anim1.1/*

[14] T.Heap, D.Hogg, "Wormholes in Shape Space: Tracking through discontinuities changes in shape", *ICCV*, pp.334-349, 1998.

[15] N.Howe, M.Leventon, W.Freeman, "Bayesian Reconstruction of 3D Human Motion from Single-Camera Video", *ANIPS*, 1999.

[16] N.Jojic, M.Turk, T.Huang, "Tracking Self-Occluding Articulated Objects in Dense Disparity Maps," *ICCV*, pp.123-130, 1999.

[17] S.Ju, M.Black, Y. Yacoob, "Cardboard people: A parameterized model of articulated motion", *2nd Int.Conf.on Automatic Face and Gesture Recognition*, pp.38-44, Oct 1996

[18] I.Kakadiaris and D.Metaxas, "Model-Based Estimation of 3D Human Motion with Occlusion Prediction Based on Active Multi-Viewpoint Selection," *CVPR*, pp. 81-87, 1996.

[19] J.Kuch and T.Huang "Vision-based modeling and tracking for virtual teleconferencing and telecollaboration," *ICCV*, pp.666-671, 1995.

[20] J.MacCormick and M.Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracker", *ECCV*, Vol.2, pp.3-19, 2000.

[21] D.Morris and J.Rehg, "Singularity Analysis for Articulated Object Tracking", *CVPR*, pp. 289-296, 1998.

[22] R.Plankers and P.Fua, "Articulated Soft Objects for Video-based Body Modeling," *ICCV*, pp. 394-401, 2001.

[23] J.Rehg and T.Kanade "Model-Based Tracking of Self Occluding Articulated Objects," *ICCV*, pp.612-617, 1995.

[24] R.Rosales and S.Sclaroff, "Inferring Body Pose without Tracking Body Parts," *CVPR*, pp.721-727, 2000.

[25] H.Sidenbladh, M.Black, D.Fleet, "Stochastic Tracking of 3D Human Figures Using 2D Image Motion," *ECCV*, 2000.

[26] C.J.Taylor, "Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image," *CVPR*, pp.677-684, 2000.

[27] B.Triggs, P.McLauchlan, R.Hartley, A.Fitzgibbon, "Bundle Adjustment - A Modern Synthesis," *Vision Algorithms: Theory and Practice*, Springer-Verlag, LNCS 1883, 2000.

[28] S.Wachter and H.Nagel, "Tracking Persons in Monocular Image Sequences," *CVIU*, 74(3):174-192, 1999.

[29] C.Wren and A.Pentland, "DYNAMAN; A Recursive Model of Human Motion," *MIT Media Lab Tech. Report*, No. 451.