Non-Linearly Embedded Visual Tracking

Cristian Sminchisescu and **Allan Jepson**

University of Toronto, Artificial Intelligence Laboratory,
Department of Computer Science, 6 King's College Road, Toronto, Canada, M5S3G4
{*crismin,jepson*}*@cs.toronto.edu, http://www.cs.toronto.edu/*{*~crismin,~jepson*}

Abstract. Many difficult visual problems like monocular human tracking require complex heuristic generative models defined over high-dimensional parameter spaces. Despite their successes, optimization with such models remains notoriously complex due to the difficulty of flexibly using prior knowledge in order to reshape an initially designed representation space. Non-linearities, inherent sparsity of high-dimensional training sets and lack of global continuity makes dimensionality reduction challenging and low-dimensional search inefficient. To address these problems, we present a sampling-based optimization framework that restricts tracking to low-dimensional spaces via non-linear embedding. The formulation leads to a layered generative model where *global continuous optimization* over the embedded manifold is made possible. Our prior flattening method allows a simple analytic treatment of boundary and manifold intrinsic curvature constraints and allows consistent iterative and closed-form solutions for embedded geodesic and sequence smoothing calculations. We analyze the structure of reduced manifold representations for a variety of human interaction activities and demonstrate that the approach gives accurate tracking and reconstruction of fast self-occluded motion in cluttered monocular video.

1 Introduction

Many successful visual tracking approaches are based on high-dimensional heuristically built non-linear generative models of shape, intensity or motion [14, 7, 8, 19, 24]. Although usually hard to construct, such models offer intuitive representations, counterpoint coherence to image clutter and offer the analytical advantage of a global coordinate system for continuous optimization or sampling. However, despite much progress, estimation with such frameworks remains notoriously difficult, mostly due to the lack of representation adaption beyond the initial design choice. This inflexibility leads to either high-dimensional, ill-conditioned parameter spaces [24] or to a lack of representational power that restricts their usage in most cases. The use of priors in the original parameter space may alleviate the problem [15, 13, 7, 19, 20] while conserving continuous representations, but still the search space dimension (*i.e*. complexity) remains unchanged. Another approach is to use forms of non-linear dimensionality reduction [5, 30, 32] but then lose the global nature of the representation [5, 30] and/or the continuity of the generative mapping [32] that makes efficient optimization possible.

University of Toronto Technical Report CSRG-477, September 2003.

To address these problems, we present a sampling-based optimization framework that restricts tracking to low-dimensional spaces via non-linear embedding [2, 27, 18]. Our formulation leads to a layered generative model where *global continuous optimization* over the embedded manifold is made possible while also respecting boundary and intrinsic curvature constraints. We describe algorithms for geodesic and smoothing calculations within the manifold. Finally, we analyze the structure of reduced manifold representations and demonstrate the approach by providing quantitative and qualitative results of tracking a variety of human activities in cluttered *monocular* video.

Related Work: There is much work involving tracking using constrained generative models [14, 5, 30] but none involving global continuous optimization over a learned non-linear manifold. Bregler & Omohundo [5] track 2D lip contours using a Gaussian Mixture Model (GMM) prior learned from training data and gradient descent. However, they still track by optimization in the original high-dimensional space and their backward/forward regularization onto/from GMM after each gradient step is not guaranteed to find a local minimum. Toyama & Blake [30] track 2D exemplars over a GMM index and Euclidean similarities using a discrete method and a set of local-coordinate system charts. Brand [3] estimates a GMM over the joint angle space and assumes known 2D silhouettes over an entire observation sequence to map to corresponding joint angle poses. The method does not produce a differentiable generative model and the coordinate system is again not global. Globally post-coordinating a local mixture representation of the manifold [4, 26] wouldn't be applicable for continuous optimization because the coordinates are uniquely defined only with respect to the considered training set. Wang *et al* [32] use an isometric embedding [28] to restrict variations of high-dimensional 2D shape coordinate sets to low-dimensions (2d in their case) and compute *local non-parametric* mappings between the intrinsic and embedding spaces. While this method, as the ones above, shares in principle the same idea of optimizing in low-dimensional spaces, it lacks a number of important desirable features for efficient manifold modeling and optimization:

(*i*) *Learning highly non-linear reduced global models* requires a dimensionality reduction method able to discover manifolds containing holes and having intrinsic curvature. These structures arise naturally in many problems and cannot be unfolded by isometric embeddings, *e.g*. physical constraints of an articulated figure or occlusion in image based representations [9]. In §2 we show that a low-dimensional representation with these properties can be built based on Laplacian, local structure preserving embeddings [2]. Estimating the intrinsic dimensionality of the model based on the Hausdorff dimension is demonstrated in §3.1.

(*ii*)*Consistent estimates. Separating sampling artifacts from intrinsic curvature* demands not only a prior on the probable regions of the embedded manifold but also a method to separate holes produced by missing sample data from genuine space curvature. The sparsity of high-dimensional training sets makes this disambiguation process inapplicable at the embedded manifold layer under unrestrictive sampling assumptions [21]. In §2.2 we propose an analytic solution that combines an embedded smoothing GMM prior with a prior flattening method that exploits the layered (hierarchical) structure of the generative model.

(*iii*)*Global continuous generative model:* Efficient continuous optimization in the embedded space requires not only a global coordinate system *but also* a global continuous generative mapping. A method for constructing such mappings between the embedded and embedding spaces is given in §2.3. The ability to do continuous search (*i.e.* compute gradients or higher-order operators of the energy surface) is a key ingredient for efficient optimization [24, 22, 25] or sampling [6, 23] in high-dimensional spaces.[1]

(*iv*)*Geodesics and Sequence Smoothing:* To obtain a powerful generative model for analysis and synthesis, interpolation and smoothing operations are also necessary. Closed-form and iterative methods for their consistent computations are given in §2.4.

2 Learning a Non-Linearly Embedded Continuous Generative Model

Consider a classical generative model $\mathcal{G}$ as a mapping $\mathbf{T}(\mathbf{x}^H, \mathbf{u})$ into an observation space. The mapping has parameters $\mathbf{x}^H = (\mathbf{x}^L, \mathbf{x}^S)$ in a heuristically constructed, high-dimensional parameter space $H = (L \cup S) \subset \Re^{D+c}$, subject to a prior $p^H(\mathbf{x}^H) = p^L(\mathbf{x}^L) \cdot p^S(\mathbf{x}^S)$. Also $\mathbf{u}$ is an indicator variable for an observable model element. The residual ρ between model mapped patterns $\mathbf{r}$ and matched extracted observations $\mathbf{r} \in O \subset \Re^z$ is used to define a likelihood over model configurations $\mathbf{x}^H$. To efficiently search H, gradient and Hessian operators for Newton style-optimization or hybrid MCMC sampling can be derived analytically by differentiating $\mathcal{G} : H \xrightarrow{\mathbf{T}} O \xrightarrow{\rho} \Re$.

Suppose we want to improve $\mathcal{G}$ by learning a subset of its representation[2] . Assume, restricting the high-dimensional subspace L to a low-dimensional embedded space $E \subset \Re^d$ that is able to represent the variability of a training set that is typical for the model's application domain. Consider such a set $\mathcal{T} = \{\mathbf{x}^{L(t)}\}_{t=1..N}$ and compute its global, non-linear embedding $\{\mathbf{x}^{E(t)}\}_{t=1..N}$ into a low-dimensional manifold E that preserves the local structure in the neighborhood of each training sample. Because the subspace L is generally non-linear and high-dimensional, we use a Laplacian neighborhood preserving embedding method that can reconstruct underlying manifolds E with holes and intrinsic curvature [2, 27][3]. To obtain an embedded generative model $\mathcal{E}$ and efficiently optimize or sample in $X = E \cup S$, we need to define a prior distribution p^E on the manifold, account for existing constraints or priors in L, and estimate a smooth global forward kernel regressor mapping $\mathbf{F}$ between E and L. The new representation $\mathbf{x} = (\mathbf{x}^E, \mathbf{x}^S) \in X$ can be mapped into H using $\mathbf{F}_X = [\mathbf{F}(\mathbf{x}^E)\ \mathbf{x}^S]$ and thus link with the observations. A *globally defined, continuous, generative mapping* from the learned model representation into the observation space can be derived as $\mathcal{E} : X \xrightarrow{\mathbf{F}_X} H \xrightarrow{\mathbf{T}} O \xrightarrow{\rho} \Re$ (see fig. 1(a)). Various components and operations of

[1] While we aim for dimensionality reduction here, it is likely that for many complex processes even reduced representations would have at least 10-15 dimensions.

[2] Without loss of generality and motivated by: (a) prior knowledge on the independence of subsets of parameters that makes their learning difficult or not necessary (*e.g.* translation and rotation of an object), (b) unavailability of training data for certain parameters.

[3] The same principles would hold for the construction of a generative model based on isometric embeddings [28], these only apply to more restricted classes of manifolds.

the model, including the estimation of mappings, layered priors and the calculation of geodesics and smoothing operations are given in the following sections.

2.1 Density Estimation and Propagation

We apply Bayes rule to compute the 'static' total posterior probability over the learned manifold space X: $p(\mathbf{x}|\mathbf{r}) \propto p(\mathbf{r}|\mathbf{x}) \cdot p(\mathbf{x}) = \exp\{-\sum_i e(\mathbf{r}_i|\mathbf{x})\} \cdot p(\mathbf{x})$. Here, $p(\mathbf{x})$ is the prior on the model parameters, $e(\mathbf{r}_i|\mathbf{x}) = \rho_i(\Delta\mathbf{r}_i(\mathbf{x})\Delta\mathbf{r}_i(\mathbf{x})^\top, \sigma_i)$ is the cost density associated with observation i and $\Delta\mathbf{r}_i(\mathbf{x}) = \mathbf{r}_i - \mathbf{T}(\mathbf{F}(\mathbf{x}^E), \mathbf{x}^S, \mathbf{u}_i)$ is the feature prediction error. For tracking, the prior at time t combines the previous posterior $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ and the dynamics $p_d(\mathbf{x}_t|\mathbf{x}_{t-1})$, where we have collected the observations at time t into vector $\mathbf{r}_t$ and defined $\mathbf{R}_t = \{\mathbf{r}_1, \ldots, \mathbf{r}_t\}$. The posterior at t becomes: $p(\mathbf{x}_t|\mathbf{R}_t) \propto p(\mathbf{r}_t|\mathbf{x}_t)\, p(\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p_d(\mathbf{x}_t|\mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ (In fact p_d will encode both simple dynamic rules and p_s in order to ensure the dynamics remains in the feasible region. The static prior p_s otuside the integral is also necessary to ensure moves to feasible configurations during 'static' likelihood search). Together $p_d(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p(\mathbf{x}_{t-1}|\mathbf{R}_{t-1})$ form the time t prior $p(\mathbf{x}_t|\mathbf{R}_{t-1})$ for the static Bayes equation.

2.2 Embedded and Layered Generative Priors

Optimization in the embedded space E requires a prior model that ensures the search is roughly confined within the manifold domain. This is determined by the typical training data but the boundary should be fuzzy in order to accommodate coverage and smoothness at moderate distances away. Smoothing should apply *both* to external boundaries to generalize away from a limited scope training set and to *internal* holes inside the domain. Holes in E may arise from sampling artifacts but may also be genuine, due to intrinsic curvature of L, perhaps because some of its regions are not feasible[4]. Disambiguating between sampling artifacts and intrinsic curvature in E may not be possible under unrestrictive sampling assumptions (Generally, we cannot assume that *e.g.* the training data available in L has been sampled uniformly from the unknown E, neither can we most of the time assume fine sampling granularity [21]).

We follow a conservative approach and use a broad prior for the on-the-manifold configurations. This may violate some of the intrinsic constrains of L, but we flexibly delegate interleaved priors at subsequent generative model layers where their representation is sharper as they may have simple analytic forms. This is straightforward in our formulation since computations are modularly performed using the transformation chain of $\mathcal{E}$. Therefore, the use of priors is not only restricted to the embedded 'optimization' space but more generally applies to variables at each generative stage up to the residual layer. Since residual differentiation is the core machinery of the generative model, analytic forms for all intermediate derivatives down the chain are available. For a generative model with $n+1$ layers having variables $\mathbf{x}_i$ with priors $p_i(\mathbf{x}_i)$ and inter-layer forward mappings $\mathbf{f}_i(\mathbf{x}_i)$, with layer 0 having prior $p(\mathbf{x})$, the flattening mechanism

[4] *E.g.* in a human kinematic representation L based on joint angles $\mathbf{x}^L$, the limits of articulations or the body non-self intersection constraints exclude certain parameter combinations (see §3).

gives and equivalent absorbed prior $p(\mathbf{x}) \leftarrow p(\mathbf{x})p_1(\mathbf{f}_1(\mathbf{x}))...p_n(\mathbf{f}_n(\mathbf{f}_{n-1}...\mathbf{f}_1(\mathbf{x})))$ (see fig. 1(b)). Jacobian volume factors need to also be taken into account (see below).

In the remaining section, we show how the above mechanism is applied for the embedded/embedding layers of our generative model in order to exploit both learned representation and intrinsic curvature constraints of L.

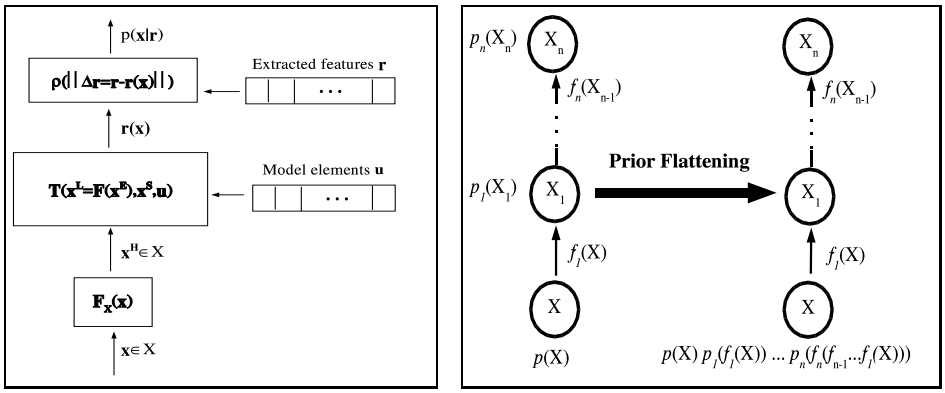

Fig. 1. (a) (left) Learned generative model allows global continuous optimization/tracking in the low-dimensional embedded space. **(b)** (right) Prior flattening mechanism allows consistent analytic treatment and optimization over manifolds with holes and intrinsic curvature.

Consider a mixture model over E obtained by k-means clustering the embedded d-dimensional training set [16] to obtain mixing proportions, centers and covariances $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1..K}$. This will be also used in section §2.4 for off-line estimation of a manifold roadmap for bootstrapping geodesic calculations. Here, we consider the mixture as a prior distribution over the manifold: $p^M(\mathbf{x}^E) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}^E, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Define now a prior on the embedded space E that combines the distribution over probable regions on the manifold with flattened priors from the embedding space L: $p^E(\mathbf{x}^E) = p^M(\mathbf{x}^E) \cdot p^L(\mathbf{F}(\mathbf{x}^E)) |\mathbf{J}_\mathbf{F}(\mathbf{x}^E)^\top \mathbf{J}_\mathbf{F}(\mathbf{x}^E)|^{1/2}$. The prior on X is thus: $p(\mathbf{x}) = p^E(\mathbf{x}^E) \cdot p^S(\mathbf{x}^S) |\mathbf{J}_\mathbf{F}(\mathbf{x}^E)^\top \mathbf{J}_\mathbf{F}(\mathbf{x})^E|^{1/2}$. Analytically differentiating $p(\mathbf{x})$ is straightforward, given that p^S is known, p^M factorizes, and there exist a parametric form for the kernel regressor mapping $\mathbf{F}$, described next.

2.3 Globally Smooth Forward and Inverse Mappings

The construction of the learned generative model requires the estimation of a forward mapping $\mathbf{F} : E(\subset \Re^d) \rightarrow L(\subset \Re^D)$ between the embedded and embedding spaces [27, 18] based on points in the training set $\mathcal{T}$ in L (stored column-wise in a matrix $\mathbf{L}$) and corresponding points in the embedded space (stored in a matrix $\mathbf{E}$). Consider a row operator $_{(i)}$ that extracts the i-th row of a matrix and $^{(i)}$ the corresponding column operator. We employ kernel regressors and estimate D mappings from $\Re^d \rightarrow \Re$.

Consider a set of r representatives $\mathbf{z}_i \in E$ and place kernels $K(\mathbf{x}, \mathbf{z}_i)$ at these points[5]. For the mapping corresponding to dimension j in D, the constraint that the vectors of the training set in E map to the real values at dimension j of the corresponding vectors in L is $\mathbf{K}{\mathbf{c}^j}^\top = \mathbf{L}_{(j)}^\top$, where $\mathbf{c}^j = [c_1^j, ..., c_r^j]$ are kernel coefficients that map into dimension j and $\mathbf{K} = [K(\mathbf{E}^{(i)}, \mathbf{z}_r)]$ is the kernel matrix of size $[N \mathrm{x} r]$, where N is the dimension of the training set. Consequently, ${\mathbf{c}^j}^\top = \mathbf{K}^+ \mathbf{L}_{(j)}^\top$ and the mapping can be derived as: $\mathbf{F}(\mathbf{x}) = [\mathbf{K_x}{\mathbf{c}^1}^\top, ..., \mathbf{K_x}{\mathbf{c}^D}^\top] = [\mathbf{K_x}\mathbf{K}^+\mathbf{L}_{(1)}^\top, ..., \mathbf{K_x}\mathbf{K}^+\mathbf{L}_{(D)}^\top]$ where $\mathbf{K_x} = [K(\mathbf{x}, \mathbf{z}_1), ..., K(\mathbf{x}, \mathbf{z}_r)]$ and $\mathbf{K}^+$ is the Moore-Penrose pseudo-inverse, computed once for all D mappings. The differentiation of the generative mapping $\mathcal{E}$ to second order for continuous optimization can now be obtained using the chain rule and the straightforward derivation of the Jacobian of $\mathbf{F}_X$: $\mathbf{J}_{\mathbf{F}_X} = \left[\begin{smallmatrix} \mathbf{J_F} & \mathbf{0}_c \\ \mathbf{0}_c & \mathbf{I}_c \end{smallmatrix}\right]$ where $\mathbf{J_F} = \frac{\mathbf{dF}}{\mathbf{dx}^E}$ is the Jacobian corresponding to the embedded mapping and $\mathbf{I}_c, \mathbf{0}_c$ are identity respectively zero square matrices of dimension c corresponding to derivatives of the representation set $\mathbf{x}^S$ that is not learned. An inverse (back-projected) mapping $\mathbf{F}^i : L \rightarrow E, \mathbf{F}_X^i$ can be similarly estimated.

We also experimented with a sparse 'lasso cost' based on individual $\mathbf{c}$ components [29, 17]. In our tests, we found that this is comparable with subset selection having the same kernel set for all dimensions, in a cross-validation loop. It tends to be more predictable, but it requires iterative optimization, which is more expensive than sampling kernel subsets. The latter can select among a larger number of models.

2.4 Embedded Geodesics and Sequence Smoothing

The construction of geodesics is framed as a regularization problem [11] where we synthesize a trajectory that is smooth and preserves the internal constraints of the manifold X (this is precisely the prior p). Assume trajectory endpoints $\mathbf{y}_0, \mathbf{y}_{T+1} \in E$ and a discretization with T knots $\mathbf{y} = [\mathbf{y}_1, ...\mathbf{y}_T], \mathbf{y}_i = (\mathbf{y}_i^E, \mathbf{y}_i^S)$. The corresponding energy function is: $V_T = -\sum_{i=1}^T \log p(\mathbf{y}_i) + \lambda \mathbf{y}\mathbf{S}_{d+c}^\top \mathbf{S}_{d+c}{\mathbf{y}}^\top = -\sum_{i=1}^T \log p^E(\mathbf{y}_i^E) - \sum_{i=1}^T \log p^L(\mathbf{F}(\mathbf{y}_i^E)) + \lambda \mathbf{y}^E \mathbf{S}_d^\top \mathbf{S}_d {\mathbf{y}^E}^\top - \sum_{i=1}^T \log p^S(\mathbf{y}_i^S) + \lambda \mathbf{y}^S \mathbf{S}_c^\top \mathbf{S}_c {\mathbf{y}^S}^\top$, where λ controls amount of regularization and $\mathbf{S}_d$ is a first order difference operator consisting of band-diagonal blocks of d-dimensional identity matrices $[... -\mathbf{I}_d \mathbf{I}_d ...]$. Higher degree of smoothness can be obtained by self-multiplication, *e.g.* for second order as $\mathbf{S}_d^\top \mathbf{S}_d^\top \mathbf{S}_d \mathbf{S}_d$, *etc.* The function V_T is differentiable and can be thus be optimized for a locally optimal solution from a trivial initialization (*e.g.* points $\mathbf{y}_i$ uniformly distributed on a straight line between $\mathbf{y}_0$ and $\mathbf{y}_{T+1}$). However, we find that in practice a better initialization is desirable for manifolds of complex topology, especially for long-range geodesic calculations[6]. Floyd's algorithm is run off-line to find all shortest paths on the set of representatives $\mathbf{z}_i$ obtained from clustering E (see §2.2). This roadmap can be effectively used at geodesic query time: given known endpoints, link to the closest representative at each end and use the precomputed road. To precompute

[5] For the work here we use simple radial basis functions $K(\mathbf{x}, \mathbf{z}_i) = \mathcal{N}(\mathbf{x}, \mathbf{z}_i, \Sigma_i)$ with diagonal covariance matrices $\Sigma_i = \sigma^2 \mathbf{I}$. We also use the means obtained by clustering E as in §2.2.

[6] Applies only to $\mathbf{y}^E$. For $\mathbf{y}^S$ use trivial straight line initialization (no training data available).

a smooth road between representatives, assume the shortest path traverses s representatives $\sigma_1, ...\sigma_s$. Take $\boldsymbol{\mu} = [\boldsymbol{\mu}_{\sigma_1}, ...\boldsymbol{\mu}_{\sigma_s}]$ and $\boldsymbol{\Sigma} = \mathbf{diag}[\boldsymbol{\Sigma}_{\sigma_1}, ..., \boldsymbol{\Sigma}_{\sigma_s}]$. An approximate energy function, that does not preserve the internal constraints of L can be derived as: $\bar{V_s^E} = (\mathbf{y}^E - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}^E - \boldsymbol{\mu}) + \lambda \mathbf{y}^E \mathbf{S}_d^\top \mathbf{S}_d {\mathbf{y}^E}^\top$ and the trajectory $\mathbf{y}$ can be solved in closed form by differentiating $\bar{V_s^E}$ to give: ${\mathbf{y}^E}^\top = (\boldsymbol{\Sigma}^{-1} + \lambda \mathbf{S}_d^\top \mathbf{S}_d)^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^\top$. This coarse trajectory can be re-parameterized, based on curvature into T pieces and estimated using the full energy function V_T. Interpolation in the embedding space L for problems with missing data (*e.g.* visual optimization in the presence of occlusion) is also possible by back-projecting the nearest neighbor corresponding to the present data (known indicies of $\mathbf{x}^L$) using $\mathbf{F}^i$ followed by geodesic computations in the embedded space as above, and forward projection using $\mathbf{F}$ (generalizations over k nearest neighbors are straightforward but require sampling from multiple geodesic paths).

Optimal estimates over an entirely tracked observation sequence (*smoothing*) are obtained in a similar manner as geodesic calculations, with the following differences:(*i*) The entire observation sequence $\mathbf{R}_t = \{\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_t\}$ is used; (*ii*) The local modes of a tracked trajectory are initial estimates for $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_t\}$, and (*iii*) Trajectory smoothness is controlled by the dynamical model $p_d(\mathbf{y}_t|\mathbf{y}_{t-1})$. Under these assumptions, the corresponding energy function is: $V_S = -\sum_{i=1}^{t} \log\{p(\mathbf{r}_i|\mathbf{y}_i) \cdot p(\mathbf{y}_i) \cdot p_d(\mathbf{y}_i|\mathbf{y}_{i-1})\}$ and can be optimized efficiently for $\mathbf{y}$ using sparse non-linear optimization methods[10, 31].

3 Human Representation Learning for Visual Tracking

Representation Learning is based on a heuristic 3D body representation that consists of a kinematic 'skeleton' of articulated joints controlled by angular joint parameters, covered by a 'flesh' built from superquadric ellipsoids with deformations. The model has 29 joint parameters $\mathbf{x}^L$, 6 global rigid parameters $\mathbf{x}^S$ and additional internal body and shape parameters $\mathbf{x}^D$. The complete model is encoded in a single parameter vector $\mathbf{x}^H = (\mathbf{x}^L, \mathbf{x}^S)$, ($\mathbf{x}^D$ are held fixed here and do not count for optimization). We learn a low-dimensional representation $\mathbf{x}^E \in E$ for $\mathbf{x}^L$ using manifold embedding on a set of training joint angle data obtained with a motion capture system (courtesy of the motion capture database at the CMU graphics laboratory [1]). We estimate a mixture model for E using k-means clustering on the d embedded eigenvectors to build the prior $p^M(\mathbf{x}^E)$ and compute a forward mapping $\mathbf{F}$ into the original joint angle space using a radial basis function approximation. During tracking and static pose estimation we estimate the parameters $\mathbf{x} = (\mathbf{x}^E, \mathbf{x}^S)$ of the global rigid motion + the embedded coordinate. In use, model superquadric surfaces are discretized into 2D meshes and the mesh nodes $\mathbf{u}$ are mapped to 3D points using knowledge of the kinematic parameters predicted at configuration $\mathbf{x}^L$ by $\mathbf{F}(\mathbf{x}^E)$. These map to each body kinematic chain and then project to predicted image points $\mathbf{r}_i(\mathbf{x})$ using perspective image projection (transformations that are encoded into $\mathbf{T}(\mathbf{x}^H, \mathbf{u})$). **The Edge and Intensity-based Observation Model** is based on sums of predicted-to-image matching likelihoods (and their gradient and Hessian metrics) evaluated for each model feature $\mathbf{r}_i$. As image features, we use a robust combination of intensity-based alignment metrics and robustified normalized edge distances [24]. **Flattened Embedded Priors** consist of soft joint angle limits and body non self-intersection constraints [24].

3.1 Experiments

The experiments we show include image-based visual tracking of human activities in *monocular* video. This underlines the importance of using prior knowledge because often the motion of subsets of body limbs is unobserved for long periods, *e.g.* when a tracked subject is sideways or not facing the camera. However, information about unobserved variables *is present* indirectly in the observed ones and this constrains their probability distribution. Learning a global, non-linear, low-dimensional representation, produces a model that couples the state variables. We derive models based on various training datasets, including walking, running and human interaction (gestures in conversations).

Analysis of the walking manifold involves a corpus of 2500 frames coming from 5 subjects, and thus contains significant variability. Fig. 2 shows walking data analysis and various structures necessary for optimization. Fig. 2(a) (left) gives estimates of the data intrinsic dimensionality based on the Hausdorff dimension $d = \lim_{r \to 0} \frac{\log N(r)}{\log(1/r)}$, where r is the radius of a sphere centered at each point, and $N(r)$ are the number of points in that neighborhood (the plot is averaged over many nearby points). The slope of the curve in the linear domain $0.01 - 1$ corresponds roughly to a 1d hypothesis. Fig. 2(b) plots the embedding distortion, computed as the normalized Euclidean SSE over each neighborhood in the training set graph. Notice its stability across different neighborhood sizes, and contrast it with the larger distortion of more variate training sets, in fig. 5(c). Fig. 2(c) and fig. 2(d) show embeddings into 2d and 3d. The latter representation is more flexible, and allows more variability. The results correspond to spherical neighborhood sizes of $r = 0.35$ and Gaussian standard deviation $\sigma = 1.25$. The figures show the embedded manifold as defined by the GMM prior $p_E(\mathbf{x})$ (3 stdev). Notice the shape has similarities with the position-velocity plot of a harmonic oscillator. Fig. 2(d) shows the spatial decomposition of the data based on oriented bounding boxes OBB [12]. This is used for fast nearest-neighbor queries in geodesic calculations (§2.4). The embedded generative model used for tracking is based on a forward mapping $\mathbf{F}$ (§2.3) that has 500 kernels.

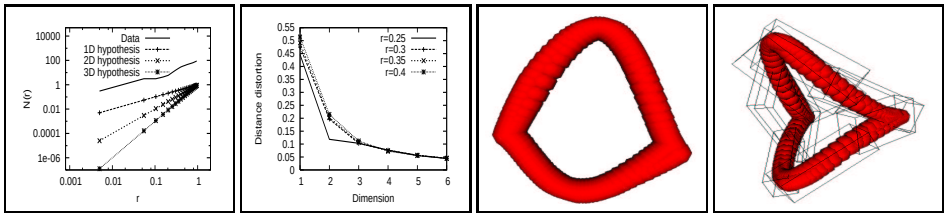

Fig. 2. Analysis of walking data. **(a)** estimates intrinsic dimensionality based on the Hausdorff dimension. **(b)** plots average local geometric embedding distortion vs. neighborhood size (notice its stability). Figures **(c)** and **(d)** show embeddings of a large 2500 walking data set in 2d and 3d and the manifold mixture prior p_E. **(d)** shows the spatial decomposition of the data used for nearest-neighbor queries in geodesic calculations (see text).

The image based tracking of walking is based on 2s of video of a subject moving against a cluttered background in a monocular sequence (fig. 3). We use a 9d state model consisting of a 3d embedded coordinate (for the 2500 walking dataset above) ($\mathbf{x}$) + 6d rigid motion ($\mathbf{x}^R$). and track using CSS with 5 hypotheses. Aside from clutter, the sequence is difficult due to the self-occlusion of the left side of the body. This occasionally makes the state variables associated to the invisible limbs close to singular. While singularity can be artificially resolved with stabilization priors, the more serious problem is that without prior knowledge, the related state variables would be mistracked, thus making recovery from failure extremely unlikely. Also notice the elimination of timescale dependence present in classical dynamic predictive models. The manifold is traversed at a speed driven by image evidence, as opposed to a prespecified one.

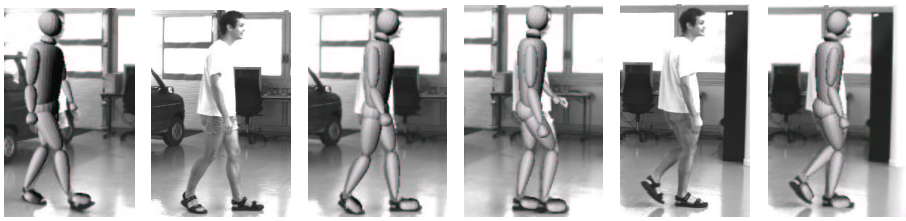

Fig. 3. Tracking a 2s *monocular* video sequence of a walking subject using optimization over a mixed 9d state space $(\mathbf{x}, \mathbf{x}^R)$ consisting of embedded 3d coordinate (from 29d walking data) + 6d (rigid motion). In this way the search complexity is significantly reduced and can tolerate missing observations (*e.g.* an occluded limb in a monocular side view).

Embedded vs. original model comparison for walking in fig. 4 is based on 60 frames of *left out* test motion capture data, synthesized using the articulated 3D model. We select 15 (3D) joint positions (shoulders, hips, elbows, *etc.*), perturb them with 1cm spherical noise to simulate modeling errors and project them onto a virtual monocular camera image plane (440x358 pixels). This input data is used to define a SSD reprojection error (Gaussian likelihood), for body joints. We track with 2 hypotheses, using both the 35d original model (having joint angle limit and body non self-intersection priors) and the 9d embedded walking model. The left and middle figures 4(a), (b) show the average pixel reprojection error per joint, whereas fig. 4(c) gives the average joint angle error with respect to ground truth (for the embedded model we plot the estimated 0.014 radians $\approx 1^o$, average range of uncertainty of the kernel regressor $\mathbf{F}$ with errorbars). Both models maintain track, but the original one overfits the data, leading to low reprojection errors, but larger variance in joint angle estimates. This is caused by tracks that follow equivalent class (monocular reflective) neighboring minima w.r.t. ground truth, more clearly noticeable at the beginning and the end of the sequence. The region between the frames 40-60 corresponds to moments where the model puppet is situated sideways in straight-stand positions with respect to the camera ray of sight. The accuracy of the original model improves during this period, perhaps because some of the depth ambiguities are eliminated due to physical constraints. The embedded model is biased for walking and has thus larger reprojection error but significantly smaller 3D

variance, having the error rather uniformly distributed among its joint angles. The average error in fig. 4(c) is about 1.4^o, and the maximum error during tracking was 4.3^o in one left hip joint angle. The original model tends to have large localized errors caused by reflective ambiguities at particular limbs. The average error in fig. 4(c) is about 2^o, but the maximum error was 35.6^o in one right shoulder joint angle. For the limited computational resources used, and for the limited walking task, the learned embedded model is clearly more accurate.

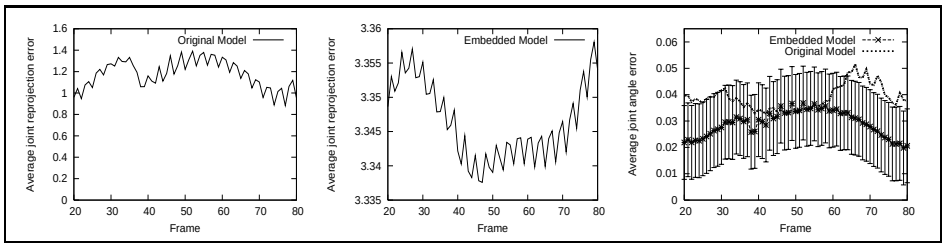

Fig. 4. Embedded (9d) vs. original (35d) model comparison for walking. **(a)** and **(b)** show the average joint reprojection error (in pixels). **(c)** plots joint angle angle error vs.ground truth (within 0.014 radians $\approx 1^o$, average uncertainty range for the map $\mathbf{F}$). The original model overfits the data (low reprojection errors, larger 3D variance estimates). The embedded model has higher bias (larger reprojection error) but also superior 3D accuracy. The original model has about 2^o average error, but the maximum error was 35.6^o in one of the right shoulder joints. The embedded one has about 1.4^o average error, but the maximum was 4.3^o in one of the left hip joints.

Analysis of the running, walking and human interaction manifold is illustrated in fig. 5 where we show a 600 point training set consisting of samples drawn from an activity set consisting of walks, runs and conversations. Left plots in fig. 5(a),(b) show 3d projections of neighborhood graphs ($r = 0.35$) for 6d and 5d embeddings onto their 3 leading Laplacian eigenvectors. Note that the the submanifolds of these activities mix, therefore pathways between these are probable (this can be also qualitatively checked by connected component analysis in the training set graph). Circular structures related to periodic walks and runs are less observable for 5d embeddings but are more clearly visible for 6d ones. The plot in fig. 5(c) confirms that the embedded neighborhood distortion decreases monotonically with increasing dimension. In practice, the stability of optimization in the embedded space becomes satisfactory beginning at about 5-6d, ruling out the use of very low-dimensional 2-4d models. The performance of the optimizer is based on both the latent space structure, and the accuracy of the mapping $\mathbf{F}$. Indeed, we found that the constrained topology of low-dimensional spaces (2-4d) collapses data from embedded runs and walks into nearly overlapping cycles (not shown), and this leads to estimation instability. In fig. 5(d) we show the good accuracy of a mapping $\mathbf{F}$ (based on 100 kernels) from the 6d embedded data in fig. 5(a) into the original 29d training set.

Tracking of human activities is exemplified in fig. 6 where we analyze a 5s video using a 12d model consisting of 6d rigid state + 6d embedded coordinate obtained from

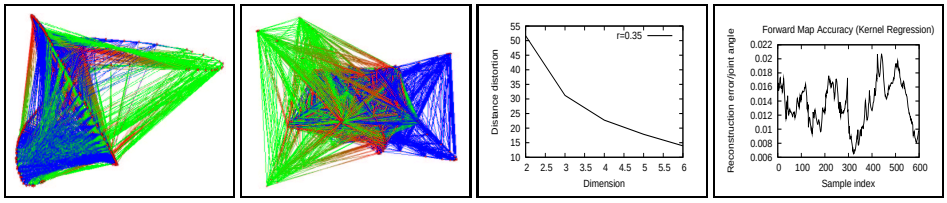

Fig. 5. Analysis for a 600 sample dataset consisting of mixed walking, running and conversation samples, best viewed in color (light red, green and blue local graph neighborhood connections originate at points in each set respectively). Left **(a)** and **(b)** show 3d projections of 6d and 5d embeddings respectively. **(c)** shows the neighborhood distortion plot for dimension range 2-6 and **(d)** plots the good average joint angle accuracy of a 6d-29d map $\mathbf{F}$, in radians (maximum $\approx 1.3^o$) (see text).

a 9000 element training set consisting of 2000 walking, 2000 running and 5000 human interaction samples. The 6d-29d mapping $\mathbf{F}$ is based on 900 kernels. Fig. 6 shows snapshots from the original sequence together with image-based tracking and monocular 3D reconstructions of the most probable configurations rendered from a synthetic scene viewpoint. The algorithm tracks and reconstructs 3D motion with good accuracy using 7 hypotheses. Missing data resulting from frequent occlusion / disocclusion of limbs would make monocular tracking with quasi-global cost sensitive search [24] or optima enumeration methods [25], *alone* difficult without prior-knowledge, or at least a sophisticated image-based limb detector. On the other hand, the presence of multiple activities and complex scenarios of human interaction demands a flexible learned representation, and makes dedicated dynamic predictors (*e.g.* walking, running) [7, 20] difficult to apply. In fig. 7 we show various components failure modes. Fig. 7(a),(b) shows the behavior of the system in a run that does not use the flattened embedded priors for physical constraints. Indeed, these are useful – notice unfeasible configurations of the right hand inside the back and right upper-arm inside the torso. The effects of missing training data on tracking behavior are explored in fig. 7(c)-(f) where an embedded model computed without conversation training data is used to track the sequence. The model tracks the first part of the sequence and the beginning of the conversation, but eventually looses lock of the arms when the gestures deviate significantly from the training set.

4 Conclusion

We have presented a sampling-based optimization framework that restricts tracking to low-dimensional spaces via non-linear embedding. Because existing approaches to optimization over learned, constrained generative representations are based on only locally valid models, they do not succeed in exploiting both the convenience of low-dimensional models and the one of efficient continuous search. Therefore they operate either discretely or in hybrid non-convergent regimes. To address these difficulties, we introduce a layered generative model where *global continuous optimization* over the embedded manifold becomes possible. Manifold boundaries and intrinsic curvature constraints are automatically accounted for and geodesic and smoothing computations can be efficiently performed in the embedded space. We analyze the structure of reduced

Fig. 6. Tracking a 5s *monocular* video sequence of mixed running, walking and conversational activities over a 12d state space. **Top row:** original sequence. **Middle row:** most probable 3D model configuration (wireframe) projected onto image at given time-step. **Bottom row:** reconstructed 3D poses rendered from a synthetic scene viewpoint. Although clutter, motion variation and missing data resulting from frequent self-occlusion / disocclusion makes monocular tracking difficult, motion tracking and reconstruction have good accuracy. Without prior knowledge, the occluded limbs can't be reliably estimated.

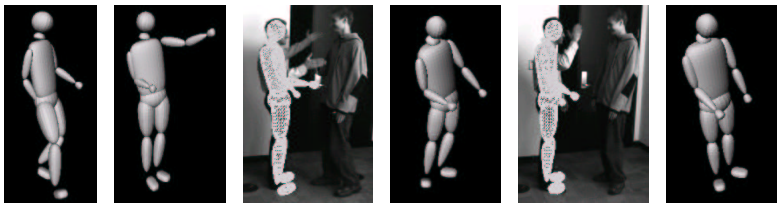

Fig. 7. Exploring system component failure modes. Left **(a), (b)** shows unfeasible configurations (right hand inside the back and right upper-arm inside the torso) from a run that does not use the flattened embedded priors for physical constraints. Middle **(c),(d)** and right **(e), (f)** show two pairs of image projection and 3D configurations when tracking with an embedded model computed without conversation data. The model tracks the beginning of the conversation but eventually looses lock of the arms when the gestures deviate significantly from the training set.

manifold representations and demonstrate the approach by providing quantitative and qualitative results of tracking a variety of human walking, running and conversational activities in cluttered monocular video.

Future and ongoing work will explore the use of low-dimensional representations for shape and appearance and multiple activity recognition.

Acknowledgments We thank Alexandru Telea for feedback on AMG software and advice on the implementation of data visualization methods. We also kindly thank Kyros Kutulakos and Nigel Morris for help with the video capture and N.M. for also posing as a model.

References

1. CMU Human Motion Capture DataBase. Available online at http://mocap.cs.cmu.edu/search.html, 2003.
2. M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, 2002.
3. M. Brand. Shadow Puppetry. In *IEEE International Conference on Computer Vision*, pages 1237–44, 1999.
4. M. Brand. Charting a Manifold. In *Advances in Neural Information Processing Systems*, 2002.
5. C. Bregler and S. Omohundro. Non-linear Manifold Learning for Visual Speech Recogntion. In *IEEE International Conference on Computer Vision*, 1995.
6. K. Choo and D. Fleet. People Tracking Using Hybrid Monte Carlo Filtering. In *IEEE International Conference on Computer Vision*, 2001.
7. J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2000.
8. J. Deutscher, A. Davidson, and I. Reid. Articulated Partitioning of High Dimensional Search Spacs associated with Articulated Body Motion Capture. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2001.
9. D. Donoho and C. Grimes. When Does ISOMAP Recover the Natural Parameterization of Families of Articulated Images? Technical report, Dept. of Statistics, Stanford University, 2003.
10. R. Fletcher. Practical Methods of Optimization. In *John Wiley*, 1987.
11. F. Girosi, M. Jones, and T. Poggio. Priors, Stabilizers and Basis Functions; From Regularization to Radial, Tensor and Additive Splines. Technical Report 1430, M.I.T, 1993.
12. S. Gottschalk, M. Lin, and D. Manocha. OBBTree: A Hierarchical Structure for Rapid Interference Detection. In *SIGGRAPH*, 1996.
13. N. Howe, M. Leventon, and W. Freeman. Bayesian Reconstruction of 3D Human Motion from Single-Camera Video. *Advances in Neural Information Processing Systems*, 1999.
14. M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *International Journal of Computer Vision*, 1998.
15. M. Leventon and W. Freeman. Bayesian Estimation of 3-d Human Motion from an Image Sequence. Technical Report TR-98-06, MERL, 1998.
16. A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*, 2001.
17. M. Osborne, B. Presnell, and B. Turlach. On the Lasso and its Dual. *J.Comput.Graphical Statist*, 9:319–337, 2000.
18. S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 2000.
19. H. Sidenbladh, M. Black, and D. Fleet. Stochastic Tracking of 3D Human Figures Using 2D Image Motion. In *European Conference on Computer Vision*, 2000.
20. H. Sidenbladh, M. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *European Conference on Computer Vision*, 2002.

21. V. Silva and G. Tenenbaum. Global versus Local Methods in Nonlinear Dimensionality Reduction. In *Advances in Neural Information Processing Systems*, 2002.
22. C. Sminchisescu and B. Triggs. Building Roadmaps of Local Minima of Visual Models. In *European Conference on Computer Vision*, volume 1, pages 566–582, Copenhagen, 2002.
23. C. Sminchisescu and B. Triggs. Hyperdynamics Importance Sampling. In *European Conference on Computer Vision*, volume 1, pages 769–783, Copenhagen, 2002.
24. C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *International Journal of Robotics Research*, 22(6):371–393, 2003.
25. C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 69–76, Madison, 2003.
26. Y. Teh and S. Roweis. Automatic Alignment of Hidden Representations. In *Advances in Neural Information Processing Systems*, 2002.
27. J. Tenenbaum. Mapping a Manifold to Perceptual Observations. In *Advances in Neural Information Processing Systems*, 1998.
28. J. Tenenbaum, V. Silva, and J. Langford. A Global Geometric Framewok for Nonlinear Dimensionality Reduction. *Science*, 2000.
29. R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. Roy. Statist.Soc*, B58(1):267–288, 1996.
30. K. Toyama and A. Blake. Probabilistic Tracking in a Metric Space. In *IEEE International Conference on Computer Vision*, 2001.
31. B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle Adjustment - A Modern Synthesis. In Springer-Verlag, editor, *Vision Algorithms: Theory and Practice*, 2000.
32. Q. Wang, G. Xu, and H. Ai. Learning Object Intrinsic Structure for Robust Visual Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2003.