Generative Modeling for Continuous Non-Linearly Embedded Visual Inference

Cristian Sminchisescu CRISMIN@CS.TORONTO.EDU
Allan Jepson JEPSON@CS.TORONTO.EDU
University of Toronto, Department of Computer Science,
6 King's College Road, Toronto, Ontario, Canada M5S 3G4

Abstract

Many difficult visual perception problems, like 3D human motion estimation, can be formulated in terms of inference using complex generative models, defined over high-dimensional state spaces. Despite progress, optimizing such models is difficult because prior knowledge cannot be flexibly integrated in order to reshape an initially designed representation space. Nonlinearities, 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 learning and inference algorithm that restricts visual tracking to automatically extracted, non-linearly embedded, low-dimensional spaces. This formulation produces a layered generative model with reduced state representation, that can be estimated using efficient continuous optimization methods. Our prior flattening method allows a simple analytic treatment of low-dimensional intrinsic curvature constraints, and allows consistent interpolation operations. We analyze reduced manifolds for human interaction activities, and demonstrate that the algorithm *learns continuous generative models* that are useful for tracking and for the reconstruction of 3D human motion in *monocular video*.

1. Introduction

Many successful visual tracking approaches are based on high-dimensional, physically inspired, non-linear generative models of shape, intensity or motion [11, 6, 15, 18]. 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 good progress, inference in these frameworks remains difficult, mostly due to the lack of learning and representation adaption beyond the initial design choice. This inflexibility leads to either high-dimensional, ill-conditioned state spaces [18], or to a lack of representational power that restricts model usage to oversimplified scenarios. The use of priors in the original state space may alleviate this problem [10, 6, 15] while conserving continuous representations, but still the state space dimension (and search complexity) remains unchanged. Another approach is to use forms of non-linear dimensionality reduction [4, 24, 25] but then lose the global nature of the representation [4, 24] or the continuity of the generative mapping [25] that makes efficient optimization possible. In this paper, we propose an algorithm that learns reduced generative models that are global, continuous and consistent during inference. These properties are motivated as follows:

(*i*) *Learning non-linear low-dimensional global models* requires a dimensionality reduction method that recovers manifolds having intrinsic curvature (*e.g.* holes). These arise in many practical modeling settings, *e.g.* physical constraints of an articulated figure or occlusion [8]. To preserve the local manifold geometry, we use a low-dimensional representation extracted using Laplacian eigenmaps [2] (§2), but other methods with similar properties *e.g.* [14, 7, 26] would also apply. Estimating the intrinsic dimensionality of the model based on the Hausdorff dimension is demonstrated in §4.1.

(*ii*) *Continuous generative model.* Continuous optimization in the low-dimensional space requires not only a reduced global coordinate system but also a globally continuous generative mapping. Assuming the original high-dimensional model is continuous, the one obtained by reducing its dimensionality should also be. In §2.1, we estimate a smooth mapping between the learned and the original model state space, based on kernel regression. Smoothness allows the use of efficient continuous methods for high-dimensional optimization [5, 20, 18, 19]. While we aim at dimensionality reduction, it is likely that for many complex processes, even reduced representations would still have rather large dimensionality (*e.g.* 10–15).

(*iii*) *Consistent estimates* require not only a prior on the probable regions of the low-dimensional manifold, as pre-

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

dicted by the typical training data density, but also separating holes produced by insufficient sampling from genuine intrinsic space curvature. The inherent sparsity of high-dimensional training sets makes this disambiguation difficult. (An analysis based on the training data distribution usually requires restrictive sampling assumptions [16]). In §2.2 we propose an analytic solution that combines a smoothing Gaussian mixture, and a prior flattening method. This exploits the layered structure of our learned generative model, in order to push down sharp curvature constrains in the low-dimensional space.

(*iv*) *Geodesics for Interpolation:* To obtain a complete low-dimensional generative model for analysis and synthesis, interpolation is also necessary. A 'geodesic' cost function for this computation is given in §2.3.

Related Work: There is important work involving tracking using constrained generative models [11, 4, 24], but we are not aware of algorithms that allow continuous optimization over a learned non-linear manifold. Bregler & Omohundo [4] track 2D lip contours using a high-dimensional Gaussian Mixture prior (GMM) learned from training data and gradient descent. They optimize in the original high-dimensional space, and regularize the estimates using GMM projection. Toyama & Blake [24] track 2D exemplars over a GMM index and Euclidean similarities using a discrete method and a set of local-coordinate system charts. Globally post-coordinating a local mixture representation of the manifold [21] would not be applicable for continuous optimization because the coordinates are uniquely defined only w.r.t. the considered training set. Thus, the coordinates of new configurations sampled during optimization may not be unique. Wang *et al* [25] use isometric embeddings [22] to restrict variations of high-dimensional 2D shape coordinate sets to low-dimensions (2d in their case) and compute local non-parametric, not necessarily continuous mappings, between their intrinsic and embedding spaces.

2. Learning a Non-Linearly Embedded Continuous Generative Model

Consider a generative model (fig. 1a)

$$\mathbf{T}_{\boldsymbol{\lambda}} : H(\subset \mathbb{R}^D) \rightarrow O(\subset \mathbb{R}^z) \tag{1}$$

representing smooth non-linear transformations $\mathbf{T}_{\boldsymbol{\lambda}}$ that reproduce the variability, but also the strong correlations, encountered in some observation domain O. The model is defined over an original state space $\mathbf{x}^H \in H$, subject to prior $p_H(\mathbf{x}^H)$, and has additional parameters $\boldsymbol{\lambda}$.[1]

A common difficulty with many intuitive, physically inspired generative models like $\mathbf{T}_{\boldsymbol{\lambda}}$, is that they usually have too general, high-dimensional state spaces, that are difficult to estimate and prior knowledge cannot be flexibly used during the model state inference. An additional difficulty (in many vision problems) is caused by the non-linearity and non-convexity of the original representation space. This may be produced, *e.g.* by (physical) domain constraints, present in the model.

To learn a consistent reduced model, we use Laplacian Eigenmaps [2], a non-linear embedding method that can, in principle reconstruct low-dimensional manifolds $E \subset \mathbb{R}^d$ ($d < D$), having intrinsic curvature (methods like [14, 7, 26] could also be used). These algorithms recover embeddings that minimally distort the local geometry of a typical distribution from $H \subset \mathbb{R}^D$. The geometry is approximated based on a training set $\mathcal{T}^H = \{\mathbf{x}^{H(t)}\}_{t=1..N}$, and the resulting embedded set of coordinates is $\mathcal{T} = \{\mathbf{x}^{(t)}\}_{t=1..N} \subset \mathbb{R}^d$. If the reduced manifold were convex, alternative embeddings that preserve the global geometry would also apply [22]. An advantage of spectral embeddings [22, 14, 2] is their good generalization [3].

A *continuous embedded generative model* (fig. 1a)

$$\mathcal{E}_{(\boldsymbol{\theta},\boldsymbol{\lambda})} : E \xrightarrow{\mathbf{F}_{\boldsymbol{\theta}}} H \xrightarrow{\mathbf{T}_{\boldsymbol{\lambda}}} O \tag{2}$$

can be obtained by learning the parameters $\boldsymbol{\theta}$ of a global smooth mapping $\mathbf{F}_{\boldsymbol{\theta}}$ between $\mathcal{T}$ and $\mathcal{T}^H$ and by constructing a prior $p(\mathbf{x}), \mathbf{x} \in E$ on the embedded manifold (fig. 1a). For consistent inference in E, the prior $p(\mathbf{x})$ has to reflect the data density in the training set $\mathcal{T}$, but also intrinsic curvature induced by existing priors at other layers in the generative model (H and beyond). Details are given in the following sections.

2.1. Globally Smooth Generative Mappings

The construction of the learned generative model requires the estimation of a forward mapping $\mathbf{F}_{\boldsymbol{\theta}} : E(\subset \mathbb{R}^d) \rightarrow H(\subset \mathbb{R}^D)$ between the embedded and embedding spaces based on points in the training set $\mathcal{T}^H$ in H (stored column-wise in a matrix $\mathbf{H}$) and corresponding points $\mathcal{T}$ 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 a sparse kernel regressor and estimate D mappings from $\mathbb{R}^d \rightarrow \mathbb{R}$. Sparsity and good generalization are important for efficient low-dimensional generative models. Consider r representatives $\mathbf{z}_l \in \mathbb{R}^d$, $l = 1...r$, and kernels $K(\mathbf{x}, \mathbf{z}_l)$ at these points.[2] The constraint that the vec-

[1]For example, consider a possible articulated generative human modeling: $\mathbf{x}^H$ are the rotational state parameters for skeleton articulations, $\boldsymbol{\lambda}$ are various internal body, shape and surface color parameters, $\mathbf{T}_{\boldsymbol{\lambda}}$ are transformations that construct the body limbs, position them through the skeletal kinematic chains and project the resulting body into the image space, O. Also, $p_H(\mathbf{x}^H)$ could be 'physical' priors that penalize states that are implausible according to anatomical constraints, *e.g.* limbs penetrating the body.

[2]Here, we use Gaussian kernels with means $\mathbf{z}_l$ and diagonal covariances $\Sigma_l = \sigma^2\mathbf{I}$. As representatives, we subsample and cross-validate the means obtained from clustering E (§2.2).

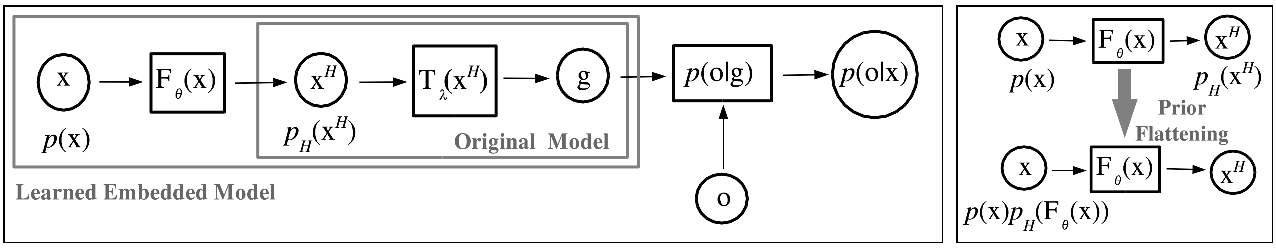

Figure 1. **(a)** (left) Learned generative model allows continuous optimization in the low-dimensional embedded space. Enclosing solid boxes label functions and circles label variables. The embedded model state $\mathbf{x}$ (or the original model state $\mathbf{x}^H$) is inferred based on input observations (data) $\mathbf{o}$. **(b)** (right) Prior flattening mechanism allows consistent optimization over manifolds with intrinsic curvature.

tors in E map to the dimension j in H is $\mathbf{K}\boldsymbol{\theta}_j^\top = \mathbf{H}_{(j)}^\top$, where $\boldsymbol{\theta}_j = [\theta_j^1, ..., \theta_j^r]$ map into dimension j and $\mathbf{K} = [K(\mathbf{E}^{(i)^\top}, \mathbf{z}_l)]$, $i = 1...N$, $l = 1...r$ is the kernel matrix of size $[N\text{x}r]$, where N is the dimension of the training set. The parameter vector is thus $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$. Consequently, $\boldsymbol{\theta}_j{}^\top = \mathbf{K}^+\mathbf{H}_{(j)}^\top$ and the mapping can be derived as: $\mathbf{F}_{\boldsymbol{\theta}}(\mathbf{x}) = [\mathbf{K}_\mathbf{x}\boldsymbol{\theta}_1{}^\top, \ldots, \mathbf{K}_\mathbf{x}\boldsymbol{\theta}_D{}^\top] = [\mathbf{K}_\mathbf{x}\mathbf{K}^+\mathbf{H}_{(1)}^\top, \ldots, \mathbf{K}_\mathbf{x}\mathbf{K}^+\mathbf{H}_{(D)}^\top]$ where $\mathbf{K}^+$ is the damped pseudo-inverse of $\mathbf{K}$, computed once for all D mappings and $\mathbf{K}_\mathbf{x} = [K(\mathbf{x}, \mathbf{z}_1), \ldots, K(\mathbf{x}, \mathbf{z}_r)]$.[3] Differentiation of the generative mapping $\mathcal{E}_{(\boldsymbol{\theta},\boldsymbol{\lambda})}$ to second order for continuous optimization can be obtained using the chain rule and the derivation of the Jacobian of $\mathbf{F}_{\boldsymbol{\theta}}$: $\mathbf{J}_{\mathbf{F}_{\boldsymbol{\theta}}}(\mathbf{x}) = \frac{\mathbf{dF}_{\boldsymbol{\theta}}(\mathbf{x})}{\mathbf{dx}}$.

2.2. Embedded and Layered Generative Priors

Consistent inference in the embedded space $\mathbb{R}^d$ requires a prior over the probable regions of the low-dimensional manifold E, determined by the training data density. Here we use a mixture prior $p_E(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{G}(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\mathcal{G}$ are Gaussian functions with parameters obtained by k-means clustering the embedded training set [12]. [4]

Sampling artifacts and problem domain constraints may interact in a way that is difficult to separate in E. In particular, the constraints may generate unfeasible regions having intrinsic curvature. Geometrically these will be holes, in both H and in E. For human kinematic representations based on joint angles $\mathbf{x}^H$, the intrinsic curvature is produced by the limits of articulations and by the body non-self intersection constraints. These exclude certain state variable combinations (see also §4). While for many domain models, analytic characterizations of unfeasible regions may be available (in H), directly separating sampling artifacts from intrinsic curvature in E is nearly impossible, under general, unrestrictive, sampling assumptions. The reason is that one cannot assume that *e.g.* the training data available in H has been sampled uniformly and / or densely from the unknown E [16], and the prior p_E is simply blind to such effects (*i.e.* it smooths them). In fact, it may assign unfeasible regions a moderately high probability, especially if these are surrounded by densely sampled zones.

Because the learned model $\mathcal{E}_{(\boldsymbol{\theta},\boldsymbol{\lambda})}$ is layered, sharper curvature constraints may be induced in the embedded space by existing priors in the original representation space, where these may be available in simple analytic form. For a layered continuous generative model $\mathcal{E}_{(\boldsymbol{\theta},\boldsymbol{\lambda})}$, one can exploit the modular structure of its forward transformation chain. Since evaluation and differentiation of $\mathcal{E}_{(\boldsymbol{\theta},\boldsymbol{\lambda})}$ with respect to its state variables is the main computational machinery of the model, analytic forms for intermediate function values and derivatives on the generative transformation chain are available. For a two-layer embedded-embedding model slice $E \xrightarrow{\mathbf{F}_{\boldsymbol{\theta}}} H$ with $\mathbf{x} \in E$, $\mathbf{x}^H (= \mathbf{F}_{\boldsymbol{\theta}}(\mathbf{x})) \in H$ and priors $p_E(\mathbf{x})$ and $p_H(\mathbf{x}^H)$ respectively, we combine the distribution over probable regions in E with flattened priors from the embedding space H: $p(\mathbf{x}) \propto p_E(\mathbf{x}) \cdot p_H(\mathbf{F}_{\boldsymbol{\theta}}(\mathbf{x})) \cdot |\mathbf{J}_{\mathbf{F}_{\boldsymbol{\theta}}}(\mathbf{x})^\top \mathbf{J}_{\mathbf{F}_{\boldsymbol{\theta}}}(\mathbf{x})|^{1/2}$ (see fig. 1b). Notice that the resulting prior is not normalized and it requires a state-dependent Jacobian scaling factor. Analytically differentiating $p(\mathbf{x})$ is possible, given p_E, and the parametric form of the mapping $\mathbf{F}_{\boldsymbol{\theta}}$, from §2.1. The mechanism allows consistent inference in the embedded space E (see §3). Priors at subsequent layers can be discarded, being already absorbed in $p(\mathbf{x})$.

2.3. Geodesics for Interpolation

The construction of geodesics can be framed as optimal inference where we synthesize a trajectory that is smooth and consistent with the prior p on the manifold E. Assume a trajectory with endpoints $\mathbf{x}^0, \mathbf{x}^{T+1} \in E$, and its discretization with T knots $\bar{\mathbf{x}} = [\mathbf{x}^1, \ldots, \mathbf{x}^T]$. The energy function for geodesics can be written as: $V_g(\bar{\mathbf{x}}) = -\sum_{i=1}^T \log p(\mathbf{x}^i) + \bar{\mathbf{x}}\mathbf{S}^\top\mathbf{S}\bar{\mathbf{x}}^\top$, where $\mathbf{S}$ is a first order difference operator square matrix of dimension $[T\text{x}d]$ consisting of T band-diagonal blocks of d-dimensional identity matrices $[... - \mathbf{I}_d\mathbf{I}_d...]$. Priors encoding higher degree

[3] We also experimented with a sparse 'lasso cost' based on individual $\boldsymbol{\theta}$ components [23, 13]: $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{j=1}^D ||\mathbf{K}\boldsymbol{\theta}_j^\top - \mathbf{H}_{(j)}^\top||^2$ with constraint $\sum_{j=1}^D \sum_{l=1}^N |\theta_j^l| \leq \alpha(r)$, and full-dimensional $\mathbf{K}$, $\boldsymbol{\theta}$. 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.

[4] The mixture centers will also be used in §2.3 for off-line estimation of a roadmap for initializing geodesic calculations.

of smoothness can be obtained by self-multiplication, *e.g.* for second order as $\mathbf{S}^\top\mathbf{S}^\top\mathbf{S}\mathbf{S}$, *etc.* The function V_g is differentiable and can be sampled or optimized for a local MAP solution from a trivial initialization (*e.g.* points $\mathbf{x}^i$ uniformly distributed on a straight line between $\mathbf{x}^0$ and $\mathbf{x}^{T+1}$). To avoid unrepresentative local optima, we initialize using Floyd's dynamic programming algorithm (DP). This is run off-line to find all shortest paths on the set of mixture centers $\boldsymbol{\mu}_i$ obtained from clustering E (see §2.2). This roadmap can be effectively used at geodesic query time: given known endpoints, link to the closest mixture component at each end and use the precomputed road (see fig. 2(d) for an oriented bounded box decomposition used in nearest neighbor queries). The DP trajectory is then refined using the consistent geodesic function V_g.

3. Temporal Inference

We apply Bayes rule to compute the 'static' total posterior probability over the learned manifold space E given (data) observation $\mathbf{o}$: $p(\mathbf{x}|\mathbf{o}) \propto p(\mathbf{o}|\mathbf{x}) \cdot p(\mathbf{x})$. Here, $p(\mathbf{x})$ is the prior on the model state space and $p(\mathbf{o}|\mathbf{x})$ is the observation likelihood, that can be computed in terms of $p(\mathbf{o}|\mathbf{g}(\mathbf{x}) = \mathbf{T}_{\boldsymbol{\lambda}}(\mathbf{F}_{\boldsymbol{\theta}}(\mathbf{x})))$, the probability of observation $\mathbf{o}$ as predicted by the generative model feature $\mathbf{g}$ at configuration $\mathbf{x}$ (see fig. 1a). For tracking using dynamic observations, the prior at time t combines the previous posterior $p(\mathbf{x}_{t-1}|\mathbf{O}_{t-1})$ and the dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, where we have collected the observations at time t into vector $\mathbf{o}_t$ and defined $\mathbf{O}_t = \{\mathbf{o}_1, \ldots, \mathbf{o}_t\}$. The posterior at t becomes: $p(\mathbf{x}_t|\mathbf{O}_t) \propto p(\mathbf{o}_t|\mathbf{x}_t) \cdot p(\mathbf{x}_t|\mathbf{O}_{t-1})$, where $p(\mathbf{x}_t|\mathbf{O}_{t-1}) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{O}_{t-1})$.[5] Together, $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and $p(\mathbf{x}_{t-1}|\mathbf{O}_{t-1})$ form the time t prior $p(\mathbf{x}_t|\mathbf{O}_{t-1})$ for the static Bayes equation. To approximate the propagating density, we use Covariance Scaled Sampling (CSS) [18]. This probabilistic method represents the posterior distribution of hypotheses in state space $p(\mathbf{x}_t|\mathbf{O}_t)$, as a Gaussian mixture, whose weights, centers and covariances are obtained as follows. Random samples are generated from the temporal prior $p(\mathbf{x}_t|\mathbf{O}_{t-1})$, and each is optimized by nonlinear local optimization (respecting any prior constraints, *etc.*) to maximize the local posterior likelihood encoded by $p(\mathbf{o}_t|\mathbf{x}_t)$. The optimized likelihood value and position gives the weight and center of a new component, and the inverse Hessian of the log-likelihood gives a scale matrix that is well adapted to the contours of the cost function, even for very ill-conditioned problems like monocular human tracking. The likelihood and temporal prior distributions are then composed and pruned to a maximum number of mixture components, in order to produce the posterior $p(\mathbf{x}_t|\mathbf{O}_t)$ for the current timestep (see [18] for details).

[5] Here $p(\mathbf{x}_t|\mathbf{x}_{t-1}) \propto p_s(\mathbf{x}_t)p_d(\mathbf{x}_t|\mathbf{x}_{t-1})$ will encode both simple dynamic rules p_d and a prior p_s in order to ensure the dynamics remains inside the feasible manifold region. We use the prior p on the manifold (§2.2) as p_s.

4. Human Representation Learning for Visual Tracking

Representation Learning is based on a physically inspired 3D body model that consists of a kinematic 'skeleton' of articulated joints controlled by angular joint variables, covered by a 'flesh' built from superquadric ellipsoids with deformations. The model has internal proportions, shape and surface color parameters $\boldsymbol{\lambda}$. The state space consists of 29 joint angle variables (for shoulder, elbow, hip, knee joints, *etc.*) and 6d global rigid motion variables $\mathbf{x}^R$, encoded in the state $\mathbf{x}^H$. We learn a low-dimensional representation $\mathbf{x} \in E$ for training vector slices of $\mathbf{x}^H$, that do not include the rigid components $\mathbf{x}^R$, using manifold embedding on a set of body joint angle training 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 by k-means clustering the embedded eigenvectors, to build the prior $p_E(\mathbf{x})$. We also learn the parameters $\boldsymbol{\theta}$ of a forward mapping $\mathbf{F}_{\boldsymbol{\theta}}$ into the original joint angle space using Gaussian kernel regression. In use, model superquadric surfaces are discretized into 2D meshes and the mesh nodes (and their colors, updated after each tracked image, *e.g.* by texture mapping) are mapped to 3D points using knowledge of the kinematic state variables predicted at configuration $\mathbf{x}^H$ by $\mathbf{F}_{\boldsymbol{\theta}}(\mathbf{x})$. These map to each body kinematic chain and then predict image positions and pixel colors, using perspective image projection, transformations that are all encoded in $\mathbf{T}_{\boldsymbol{\lambda}}(\mathbf{x}^H)$ [6]. **The Observation Model** is based on sums of predicted-to-image matching likelihoods (and their gradient and Hessian metrics) evaluated for each model feature prediction $\mathbf{g}$. As image features, we use a robust combination of intensity-based alignment metrics, silhouettes and robustified normalized edge distances [18]. **Flattened Embedded Priors** consist of soft joint angle limits and body non self-intersection constraints [18]. For the experiments here, we work with the negative log-likelihood energy function in §3 and the prior is not normalized and not scaled. For **temporal state inference** (tracking), we use CSS [18], as explained in §3.

4.1. 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

[6] The 6d global rigid state representation $\mathbf{x}^R$ is **not** learned using embedding because people can move in any directions and can be seen from any viewpoint, so it is restrictive to learn preferential subspaces for global translation or rotation. This implies that this slice of variables, although part of the inferred state, is mapped by $\mathbf{T}_{\boldsymbol{\lambda}}$, and not by $\mathcal{E}_{(\boldsymbol{\theta},\boldsymbol{\lambda})}$. This is simply a technicality and we avoided making it explicit for notational simplicity. In practice, we do inference over an augmented hidden state $(\mathbf{x}, \mathbf{x}^R)$ (embedded coordinate + global rigid motion) and therefore need to add a trivial identity component to $\mathbf{F}_{\boldsymbol{\theta}}$ for the map $\mathbf{g} = T_{\boldsymbol{\lambda}}(\mathbf{x}^H = (\mathbf{F}_{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{x}^R))$.

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 $d = \lim_{r\rightarrow 0} \frac{\log N(r)}{\log(1/r)}$, 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 [9]. This is used for fast nearest-neighbor queries in geodesic calculations (§2.3). The embedded generative model used for tracking is based on a forward mapping $\mathbf{F}_{\boldsymbol{\theta}}$ (§2.1) that has 500 kernels.

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.

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 $\mathbf{F}_{\boldsymbol{\theta}}$ 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.

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

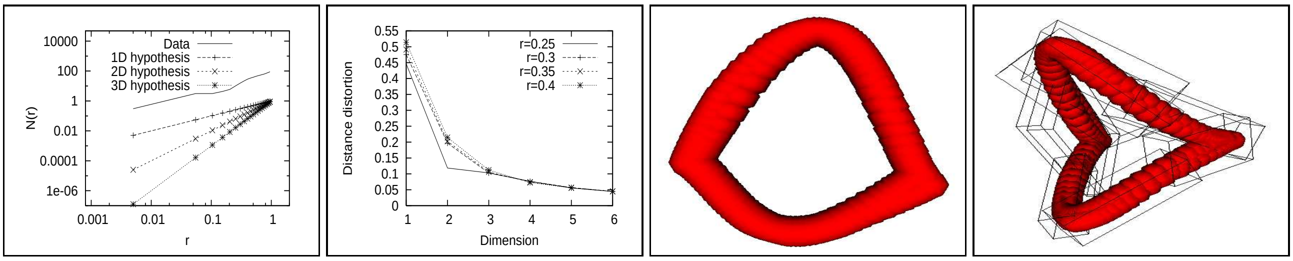

Figure 2. Analysis of walking data. **(a)** estimates intrinsic dimensionality based on the Hausdorff dimension. **(b)** plots average local geometric embedding distortion vs. neighborhood size (notice its stability). Figures **(c)** and **(d)** show embeddings of a large 2500 walking data set in 2d 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).

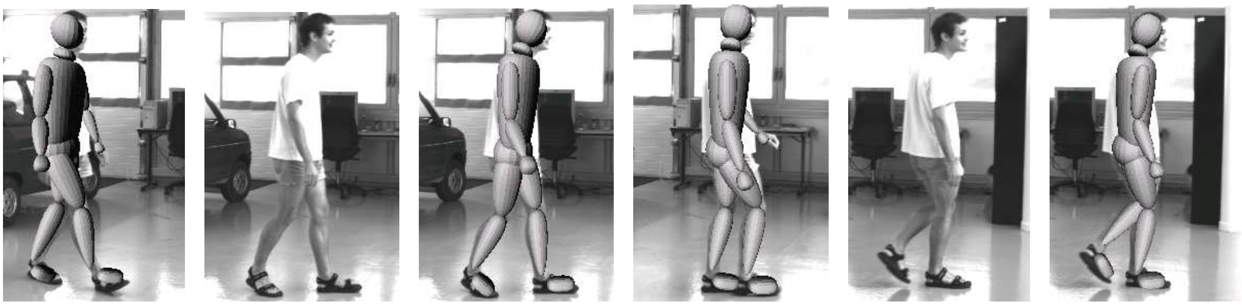

Figure 3. Tracking a 2s *monocular* video sequence of a walking subject using optimization over a mixed 9d state space $(\mathbf{x}, \mathbf{x}^R)$ consisting of embedded 3d coordinate (from 29d walking data) + 6d (rigid motion). In this way the search complexity is significantly reduced and can tolerate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

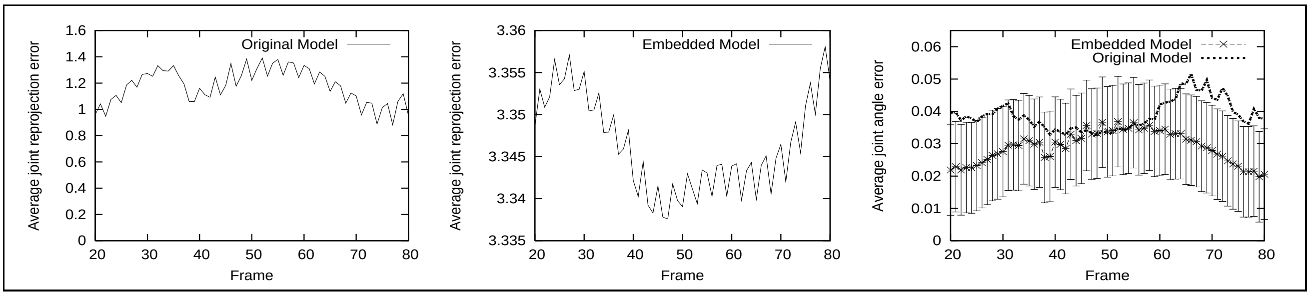

Figure 4. Embedded (9d) vs. original (35d) model comparison for walking. **(a)** and **(b)** show the average joint reprojection error (in pixels). **(c)** plots joint angle angle error vs.ground truth (within 0.014 radians $\approx 1^o$, average uncertainty range for the map $\mathbf{F}_{\boldsymbol{\theta}}$). 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.

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 $\mathbf{F}_{\boldsymbol{\theta}}$. 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 $\mathbf{F}_{\boldsymbol{\theta}}$ (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 a 9000 element training set consisting of 2000 walking, 2000 running and 5000 human interaction samples. The 6d-29d mapping $\mathbf{F}_{\boldsymbol{\theta}}$ 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 [18] or optima enumeration methods [19], *alone* difficult without prior-knowledge, or at least a sophis-

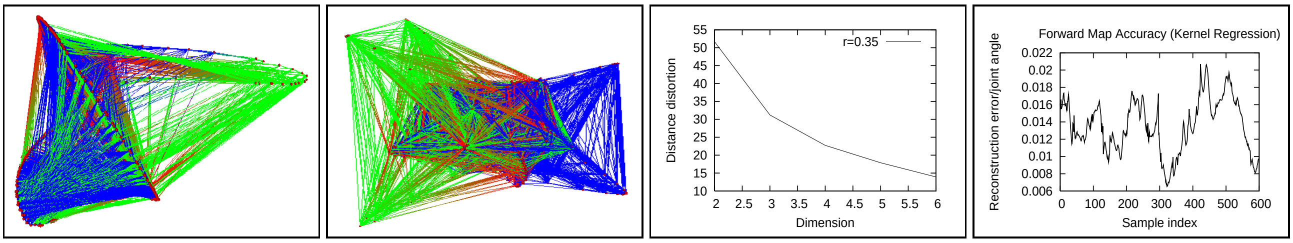

Figure 5. Analysis for a 600 sample dataset consisting of mixed walking, running and conversation samples, best viewed in color (light red, green and blue local graph neighborhood connections originate at points in each set respectively). Left **(a)** and **(b)** show 3d projections of 6d and 5d embeddings 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 $\mathbf{F}_{\boldsymbol{\theta}}$, in radians (maximum $\approx 1.3^o$) (see text).

ticated 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) [6, 15] 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.

5. Conclusion

We have presented a learning and inference framework that reduces visual tracking to low-dimensional spaces computed using non-linear embedding. Because existing approaches to optimization over learned, constrained generative representations are based on only locally valid models, they can't easily exploit both the convenience of low-dimensional modeling and the one of efficient continuous search. Therefore they may operate either discretely or in hybrid non-convergent regimes. To address these difficulties, we introduce a layered generative model having learned, embedded representation, that can be estimated using efficient continuous optimization methods. We analyze the structure of reduced manifold representations for a variety of human walking, running and conversational activities, and demonstrate the algorithm by providing quantitative and qualitative results of human tracking and 3D motion reconstruction based on learned low-dimensional models, in monocular video.

Future and ongoing work will explore the construction of flexible dynamic predictors for tracking, low-dimensional shape representations, and activity recognition.

Acknowledgments Special thanks to Kyros Kutulakos and Nigel Morris for generous help with video capture.

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 *NIPS*, 2002.

[3] J. Bengio, J. Paiement, and P. Vincent. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps and Spectral Clustering. In *NIPS*, 2003.

[4] C. Bregler and S. Omohundro. Non-linear Manifold Learning for Visual Speech Recogntion. In *ICCV*, 1995.

[5] K. Choo and D. Fleet. People Tracking Using Hybrid Monte Carlo Filtering. In *ICCV*, 2001.

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

[7] D. Donoho and C. Grimes. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-dimensional Data. *Proc. Nat. Acad. Arts and Sciences*, 2003.

[8] 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.

[9] S. Gottschalk, M. Lin, and D. Manocha. OBBTree: A Hierarchical Structure for Rapid Interference Detection. In *SIGGRAPH*, 1996.

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

[11] M. Isard and A. Blake. CONDENSATION – Conditional Density Propagation for Visual Tracking. *IJCV*, 1998.

[12] A. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *NIPS*, 2001.

[13] M. Osborne, B. Presnell, and B. Turlach. On the Lasso and its Dual. *J.Comput.Graphical Statist*, 9:319–337, 2000.

[14] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 2000.

[15] H. Sidenbladh, M. Black, and L. Sigal. Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In *ECCV*, 2002.

[16] V. Silva and G. Tenenbaum. Global versus Local Methods in Nonlinear Dimensionality Reduction. In *NIPS*, 2002.

[17] C. Sminchisescu and A. Jepson. Variational Mixture Smoothing for Non-Linear Dynamical Systems. In *CVPR*, Washington D.C., 2004.

Figure 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.

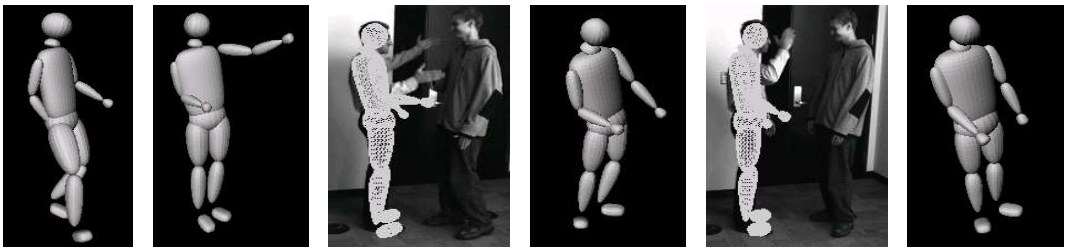

Figure 7. Exploring system component failure modes. Left **(a), (b)** shows unfeasible configurations (right hand inside the back and right upper-arm inside the torso) from a run that does not use the flattened embedded priors for physical constraints. Middle **(c),(d)** and right **(e), (f)** show two 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.

[18] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *IJRR*, 22(6):371–393, 2003.

[19] C. Sminchisescu and B. Triggs. Kinematic Jump Processes for Monocular 3D Human Tracking. In *CVPR*, volume 1, pages 69–76, Madison, 2003.

[20] C. Sminchisescu, M. Welling, and G. Hinton. A Mode-Hopping MCMC Sampler. Technical Report CSRG-478, University of Toronto, submitted to *Machine Learning Journal*, September 2003.

[21] Y. Teh and S. Roweis. Automatic Alignment of Hidden Representations. In *NIPS*, 2002.

[22] J. Tenenbaum, V. Silva, and J. Langford. A Global Geometric Framewok for Nonlinear Dimensionality Reduction. *Science*, 2000.

[23] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. Roy. Statist.Soc*, B58(1):267–288, 1996.

[24] K. Toyama and A. Blake. Probabilistic Tracking in a Metric Space. In *ICCV*, 2001.

[25] Q. Wang, G. Xu, and H. Ai. Learning Object Intrinsic Structure for Robust Visual Tracking. In *CVPR*, 2003.

[26] K. Weinberger and L. Saul. Unsupervised Learning of Image Manifolds by Semidefinite Programming. In *CVPR*, 2004.