Contributions

⇒ Learning a low-dimensional, non-linear generative model that is global and allows consistent inference using efficient continuous methods
⇒ Can be a restriction of a general domain model (e.g. predicting human body motions) to specific tasks (e.g. running, walking or motions during conversations)
⇒ Applies to difficult inference problems (non-linear dynamical systems)
  • complex high-dimensional models, ambiguous image evidence, etc.
  • tracking, structure from motion
⇒ Preserves the power of generative modeling
  • Arbitrary multimodal dynamics and hidden state conditional distributions
  • Analytic representation for transformations that are restrictive or expensive to learn (e.g. global translation, rotation)
⇒ Existing domain constraints are preserved during inference
⇒ Efficient continuous optimization or sampling methods directly apply

Approach

⇒ Use unsupervised learning to compute non-linear embedding of original model hidden state space into low-dimensional latent space, given typical training data
⇒ Compute smooth forward mapping between latent and original hidden space using kernel regression (simple map because embedding preserves local geometry)
⇒ Construct latent space prior (data density + existing domain constraints)
⇒ Perform robust visual inference using mixture density propagation in latent space

Conclusions

1. Low-dimensional 9-12 d.o.f. models useful for tracking human activities
2. Improved stability (less variance), good accuracy, can tolerate missing data

Learning the Generative Model

⇒ Use unsupervised learning to compute non-linear embedding of the original model hidden state space into low-dimensional latent space
  • we use Laplacian Eigenmaps but Hessian or LLE may as well recover non-linear, intrinsically curved manifolds
  • intrinsic curvature may be produced by model domain constraints (e.g. for 3D human models, joint angle limits or body non-self-intersection)
  • estimate model intrinsic dimensionality based on the Hausdorff dimension
⇒ Compute smooth forward mapping \( F_o \) between latent and original hidden space using kernel regression
  • select a sparse set of inputs by clustering in latent space; use dumped, cross-validated estimator for improved generalization
⇒ Construct latent space prior \( p(x) \)
  • model training data distribution as a Gaussian mixture \( p_M \) (clustering in latent space is convenient due to reduced dimensionality)
  • prior flattening transfers existing constraints (expressed as priors \( p_H \) on the original hidden state) into latent space. This helps separate sampling artifacts from intrinsic curvature

Visual Inference

⇒ For temporal inference, solve filtering recursion in latent space:

\[
p(x_t|O_t) \propto p(o_t|x_t) \int p(x_{t+1}|x_t) \cdot p(x_{t+1}|O_{t+1}) \, dx_{t+1}
\]

where \( O_t = (o_1, \ldots, o_t) \) is the joint observation vector up to time \( t \), \( p(x_t|x_{t-1}) \) is the dynamic transition rule and \( p(o_t|x_t) \) is the observation model
⇒ We use the Covariance Scaled Sampling, a mixture density propagation algorithm adapted for high-dimensional problems
  • covariance-based random sampling + mode finding using local continuous optimization (with heavy tail observation likelihood distributions)

---

Experiment Set 1

- 2500 sample motion capture joint angle walking data
- 9d model (6d rigid motion + learned 3d latent coordinates), forward map with 500 kernels
- 2s monocular tracking of walking, 9d learned model

---

Experiment Set 2

- 9000 sample training set contains running, walking and conversation activities
- 12d model (6d rigid motion + learned 6d latent coordinates)
- Forward map based on 900 kernels
- 5s monocular video, fast motion in clutter
- Good foreground / background segmentation is not easy (see e.g. shadows on the wall)
- Occlusion of limbs makes tracking without learning difficult

Some of the component failure modes

- No push-down physical priors
- Only trained with running data