People Tracking with the Laplacian Eigenmaps Latent Variable Model

Zhengdong Lu Miguel Á. Carreira-Perpiñán
Dept. of Computer Science & Electrical Eng.,
OGI, Oregon Health & Science University
{zhengdon,miguel}@csee.ogi.edu

Cristian Sminchisescu
Toyota Technological Institute, Chicago
crismin@nagoya.uchicago.edu

Articulated pose tracking consists of extracting the 3D pose of a person from e.g. monocular video (see fig. 2). This problem is hard for several reasons. The space of poses is high-dimensional (between 30–60 joint angles or joint positions depending on the desired accuracy level) and the human poses that are actually feasible or typical lie in a complex low-dimensional manifold of it. This manifold is due to correlations of the joints during motion, physical constraints (limited range of variation of the angles, non-intersection of body parts) and kinematic constraints. Other difficulties include depth ambiguities (from the projection 3D to 2D) and a complex appearance of the body in the image, as well as general difficulties common to tracking in vision (such as data association and occlusion). Recent research in reconstructing articulated human motion has focused on methods that can exploit available prior knowledge on typical human poses or motions in an attempt to build more reliable algorithms. Here, the high-dimensional space of poses is represented as a low-dimensional latent space learned from motion-capture data or directly from images. This is then used to track.

We apply to this problem a recently introduced probabilistic dimensionality reduction method, the Laplacian Eigenmaps Latent Variable Model (LELVM) [2]. LELVM is based on a natural way of defining an out-of-sample mapping for Laplacian eigenmaps (LE) which, in addition, results in a density model. In LE, typically we first define a k-nearest-neighbour graph on the sample data $\{\mathbf{y}_n\}_{n=1}^N \subset \mathbb{R}^D$, weigh each edge $\mathbf{y}_n \sim \mathbf{y}_m$ by a Gaussian affinity function $K(\mathbf{y}_n, \mathbf{y}_m; \sigma)$ and build the graph Laplacian. Then the latent points $\{\mathbf{x}_n\}_{n=1}^N \subset \mathbb{R}^L$ result from minimising a quadratic objective. We now define an out-of-sample mapping $\mathbf{F}(\mathbf{y}) = \mathbf{x}$ for a new point $\mathbf{y}$ as a semi-supervised learning problem, by recomputing the embedding as before, but keeping the old embedding $\{\mathbf{x}_n\}_{n=1}^N$ fixed. This is the most natural way of adding a new point to the embedding without disturbing the previously embedded points. The solution can be obtained in closed form and yields an out-of-sample dimensionality reduction mapping $\mathbf{x} = \mathbf{F}(\mathbf{y})$ equal to a Nadaraya-Watson estimator (kernel regression) using as data $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ and the kernel K. We can take this a step further by defining a LVM that has as joint distribution a kernel density estimate (KDE) $p(\mathbf{x}, \mathbf{y}) = \frac{1}{N}\sum_{n=1}^N K_{\mathbf{y}}(\mathbf{y}, \mathbf{y}_n) K_{\mathbf{x}}(\mathbf{x}, \mathbf{x}_n)$. The marginals in observed and latent space are also KDEs and the dimensionality reduction and reconstruction mappings are given by kernel regression: the conditional means $\mathbf{f}(\mathbf{x}) = \mathrm{E}\{\mathbf{y}|\mathbf{x}\}$ and $\mathbf{F}(\mathbf{y}) = \mathrm{E}\{\mathbf{x}|\mathbf{y}\}$. This out-of-sample extension is different from that of [1], which does not yield a density or a reconstruction mapping.

Thus, LELVM naturally extends a LE embedding (efficiently obtained as a sparse eigenvalue problem with a cost $\mathcal{O}(N^2)$) to global, continuous, differentiable mappings (Nadaraya-Watson estimators) and potentially multimodal densities having the form of a Gaussian KDE. This allows easy computation of posterior probabilities such as $p(\mathbf{x}|\mathbf{y})$ (unlike GPLVM). It can use a continuous latent space of arbitrary dimension L (unlike GTM) by simply choosing L eigenvectors in the LE embedding. It has no local optima since it is based on the LE embedding. LELVM is able to learn convoluted manifolds (e.g. the Swiss roll) and define mappings and densities for them. The only parameters to be set are the graph parameters (number of neighbours k and affinity width σ) and the smoothing bandwidths $\sigma_{\mathbf{x}}$ and $\sigma_{\mathbf{y}}$. Thus, it is a nonparametric LVM that neatly combines the advantages of latent variable models and spectral manifold learning methods.

Since LELVM defines a probabilistic model, it can be straightforwardly integrated into a sequential Bayesian estimation framework. Our state $\mathbf{s} = (\mathbf{x}, \mathbf{d})$ contains the latent pose $\mathbf{x}$ (2D sufficed in our case) and the rigid motion coordinates of the body $\mathbf{d}$ (centre-of-mass, orientation). The observed variables $\mathbf{z}$ consist of image features or the perspective projection of the markers on the camera plane. The mapping from state to observations combines the LELVM mapping $\mathbf{f}(\mathbf{x})$ (which provides the position $\mathbf{y}$ of the 3D markers) and the perspective projection (pinhole camera), plus Gaussian noise. Our dynamics model is $p(\mathbf{s}_t|\mathbf{s}_{t-1}) \propto p_{\mathbf{d}}(\mathbf{d}_t|\mathbf{d}_{t-1})\, p_{\mathbf{x}}(\mathbf{x}_t|\mathbf{x}_{t-1})\, p(\mathbf{x}_t)$, where both dynamics models for $\mathbf{d}$ and $\mathbf{x}$ are random walks, and $p(\mathbf{x}_t)$ is the LELVM prior. Thus the overall dynamics predicts states that are both near the previous state and yield feasible poses. As tracker we use the Gaussian mixture Sigma-point particle filter (GMSPPF) [3], but any probabilistic tracker for nonlinear, nongaussian models

Topic: visual processing and pattern recognition
Preference: oral

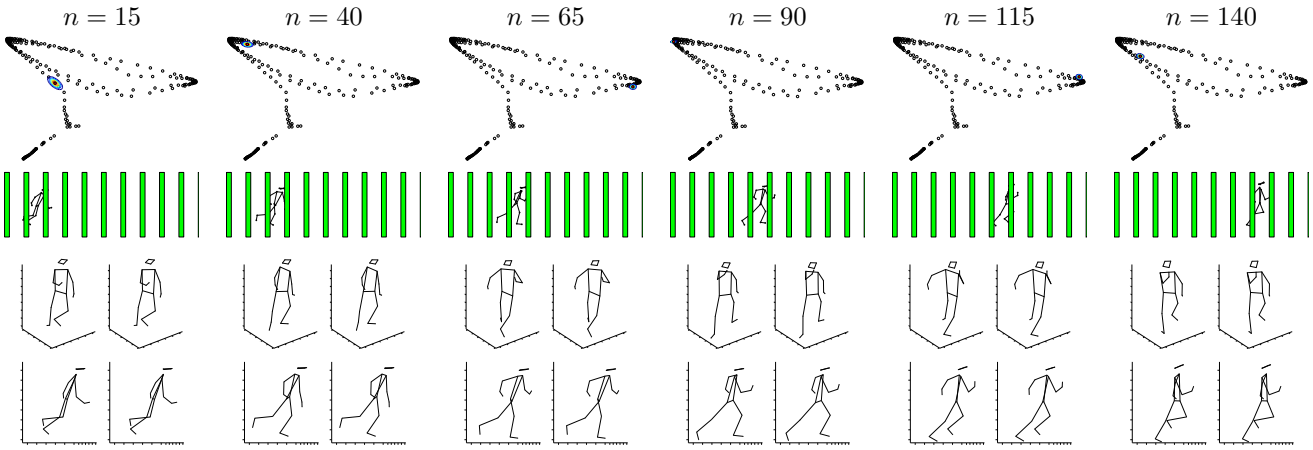

Figure 1: OSU running man motion capture data. We use 217 datapoints for training LELVM and for tracking. *Row 1*: tracking in the 2D latent space. The contours (very tight in this sequence) are the posterior probability. *Row 2*: perspective-projection-based observations with occlusions. *Row 3*: true pose of the running man (left subplot) and reconstructed pose (right subplot). *Row 4*: as in the third row but from a side view.

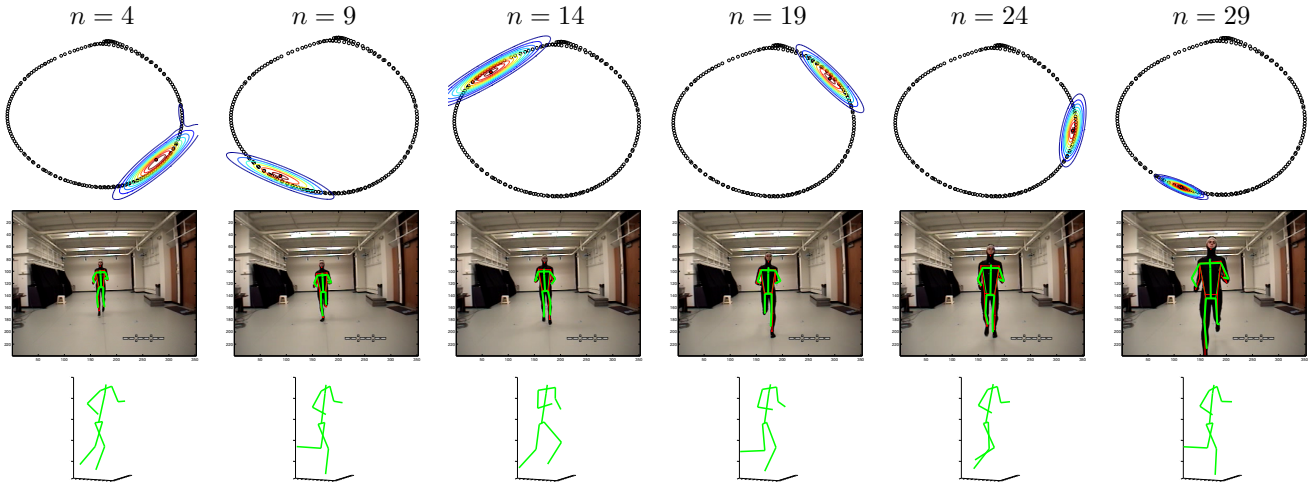

Figure 2: Tracking of a person running straight towards the camera. Notice the scale changes and possible forward-backward ambiguities in the 3D estimates. We train the LELVM using 180 datapoints (2.5 running cycles) from a different person; 2D tracks were obtained by manually marking the video. *Row 1*: tracking in the 2D latent space. *Row 2*: tracking based on markers. The red dots are the 2D tracks and the green stick man is the 3D reconstruction obtained using our model. *Row 3*: our 3D reconstruction from a different viewpoint.

would work as well. Figs. 1–2 show a subset of our experiments on image sequences (synthetic and real) of people running. The model can cope well with persistent partial occlusion, severely subsampled training data and missing markers; and works well when trained on motion data from one person but applied to track a different person (with different body proportions and motion style).

References

[1] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219, Oct. 2004.

[2] M. Á. Carreira-Perpiñán and Z. Lu. The Laplacian Eigenmaps Latent Variable Model. *AISTATS*, 2007.

[3] R. van der Merwe and E. A. Wan. Gaussian mixture sigma-point particle filters for sequential probabilistic inference in dynamic state-space models. *ICASSP*, volume 6, pages 701–704, 2003.