Hierarchical Latent Variable Models for Pose Inference

Catalin Ionescu and Cristian Sminchisescu
University of Bonn, Faculty of Mathematics and Natural Sciences, INS,
HUMOREV, Computer Vision and Machine Learning Group
{Catalin.Ionescu|Cristian.Sminchisescu}@ins.uni-bonn.de; http://sminchisescu.ins.uni-bonn.de

We present a hierarchical model that combines latent variable models of local distributions into a tree dependency structure. Two algorithms for training the model are proposed and a way to apply it to human pose prediction. It builds on the previous work on Spectral Latent Variable Model (SLVM) by hierarchically partitioning the space and building mappings between the different levels of the hierarchical latent representation. An overall likelihood of the joint model can be computed by integrating out the latent hierarchy. Direct optimization, a top-down and a bottom-up learning scheme have been investigated, leveraging the availability of mappings in both directions (latent↔ambient) in SLVM. For motivation, consider the case of a running person. The intrinsic dimensionality of a runner is, in the limit, the one of a 1d harmonic oscillator. But this minimalist model does not account for stylistic differences or for the lack of synchronization inherent of many subjects or scenes. What seems appropriate is a hierarchy, with the strongest (lowest-dimensional) model of correlation at the top (say), the weakest high-dimensional model of limbs moving unrestrictedly at the bottom, and various degrees of flexibility in-between (*e.g.* models with regularities among subsets of variables for each leg or arm, but without global constraints among all of them). The hierarchy can be used for the visual inference of 3D human body pose from images, either by estimating several representation levels simultaneously, or by automatically adapting the level of complexity to match the statistical regularity of the observation.

Spectral Latent Variable Models (SLVM): Assume vector-valued points in ambient space $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1\dots N}$ captured from a high-dimensional process, and corresponding latent space points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1\dots N}$, initially obtained using a spectral, non-linear embedding method like ISOMAP, LLE, Hessian or Laplacian Eigenmaps, *etc*. We model the joint distribution over latent and ambient variables as: $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$. The latent space prior $p(\mathbf{x})$ is modeled as a non-parametric kernel density estimate, with covariance $\boldsymbol{\theta}$: $p(\mathbf{x}) = \frac{1}{K}\sum_{i=1}^{K} K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}_i)$. In the model, we assume that ambient vectors are related to the latent ones using a nonlinear vector-valued function with parameters $\mathbf{W}$ and noise covariance $\boldsymbol{\sigma}$: $p(\mathbf{y}|\mathbf{x}, \mathbf{W}, \boldsymbol{\sigma}) \sim \mathcal{N}(\mathbf{y}|\mathbf{F}(\mathbf{x}, \mathbf{W}), \boldsymbol{\sigma})$, where $\mathcal{N}$ is a Gaussian distribution with mean $\mathbf{F}$ and covariance $\boldsymbol{\sigma}$. $\mathbf{F}$ is a generalized regression model: $\mathbf{F}(\mathbf{x}, \mathbf{W}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x})$ with $\boldsymbol{\phi}(\mathbf{x}) = [K_{\boldsymbol{\delta}}(\mathbf{x}, \mathbf{x}_1), \dots, K_{\boldsymbol{\delta}}(\mathbf{x}, \mathbf{x}_M)]^{\top}$, and kernels with covariance $\boldsymbol{\delta}$ placed at an M-size subset of $\mathbf{x}_i$. $\mathbf{W}$ is a weight matrix of size DxM.

The ambient marginal is obtained by integrating the latent space. The evidence, as well as derivatives w.r.t. model parameters, are computed using a simple Monte Carlo (MC) approximate using, say S, samples from the prior. This gives the MC estimate of the ambient marginal: $p(\mathbf{y}|\mathbf{W}, \boldsymbol{\sigma}) = \int p(\mathbf{y}|\mathbf{x}, \mathbf{W}, \boldsymbol{\sigma})p(\mathbf{x})\mathbf{dx} \approx \frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y}|\mathbf{x}^{(s)}, \mathbf{W}, \boldsymbol{\sigma})$. The latent space conditional is obtained using Bayes' rule: $p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{S}{K}\frac{p(\mathbf{y}|\mathbf{x})\sum_{i=1}^{K} K_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}_i)}{\sum_{s=1}^{S} p(\mathbf{y}|\mathbf{x}^{(s)},\mathbf{W},\boldsymbol{\sigma})}$. For pairs of ambient data points j and MC latent samples i, we abbreviate $p_{(i,j)} = p(\mathbf{x}_i|\mathbf{y}_j)$. The choice of prior $p(\mathbf{x})$ influences the membership probabilities. We can compute either the conditional mean or the mode (better for multimodal distributions) in latent space, using the same MC integration method: $\mathrm{E}\{\mathbf{x}|\mathbf{y}_n, \mathbf{W}, \boldsymbol{\sigma}\} = \int p(\mathbf{x}|\mathbf{y}_n, \mathbf{W}, \boldsymbol{\sigma})\mathbf{x}\mathbf{dx} = \sum_{i=1}^{K} p_{(i,n)}\mathbf{x}_i$, where $i_{max} = \arg\max_i p_{(i,n)}$. The model contains the ingredients for efficient computation in both latent and ambient space: a prior in latent space, an ambient marginal, the conditional distribution from latent to ambient space,

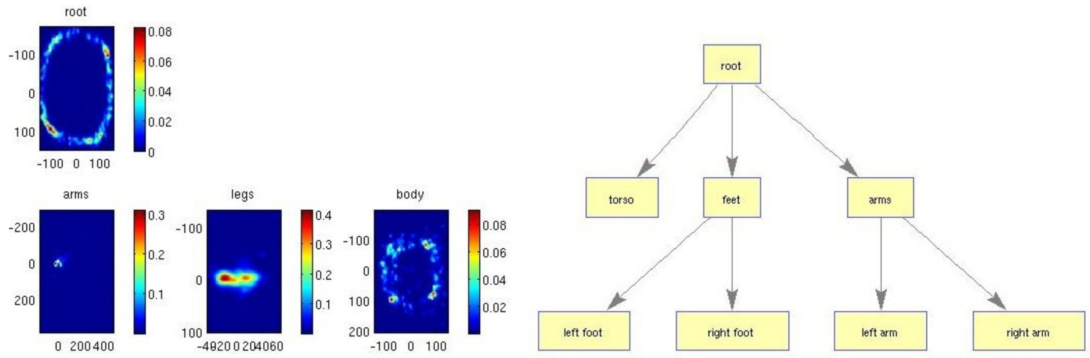

Figure 1: *Left* Latent spaces in a hierarchy trained with human poses. *Right* SLVM nodes.

and vice-versa. Latent conditionals given partially observed $\mathbf{y}$ vectors are easy computable – the conditional distribution of $\mathbf{y}$ is Gaussian and unobserved components can be integrated analytically (see [2] for details). The model can be trained by maximizing the log-likelihood of the data: $\mathcal{L} = \log \prod_{i=1}^{N} p(\mathbf{y}_n|\mathbf{W}, \boldsymbol{\sigma}) = \sum_{n=1}^{N} \log\{\frac{1}{S} \sum_{s=1}^{S} p(\mathbf{y}_n|\mathbf{x}^{(s)}, \mathbf{W}, \boldsymbol{\sigma})\}$. Maximizing the likelihood provides estimates for $\mathbf{W}, \boldsymbol{\sigma}$. The model is learned using EM: int the E-step we compute the membership probabilities of latent points generating datapoints, $p_{(i,j)}$. In the M-step we learn the sparse mapping and its noise model $(\mathbf{W}, \boldsymbol{\sigma})$, by solving a weighted regression problem [2].

Initialization: The initialization is identical for all learning procedures. We partition the data space and assign each part to a leaf. At each level of tree we compute non-linear embeddings using the datapoints representing subsets of variables in the level below, in order to obtain the data for the level above. We always combine (complementary) latent representations corresponding to the same original datapoint and don't crossover between different ones. Once the data at each node is available, KDE approximations to marginals (latent 'priors' in SLVMs) are constructed and mappings to the level below are learned (*i.e*. learn individual SLVMs for all parent-child node pairs).

Forward and posterior learning: The classical problem with latent variable models is they don't have data to easily train them. We use a form of Markov chain EM by keeping a KDE at the root fixed, sampling from it and propagating samples through the hierarchy. We then use the samples as data to train the SLVM nodes. Alternatively we can do a bottom-up learning taking advantage of the possibility of computing the expectation of the latent variables given the data. We can use these expected latent coordinates to populate the data in the higher levels of the hierarchy. With the nodes populated with data an EM procedure (SLVM) is used to refine the mappings between the nodes.

Pose prediction: This model can be used for pose estimation by training predictors (*e.g*. ridge regressor) between image features and latent variable coordinates corresponding to their poses at each level in the hierarchy. One possible choice for latent variable data used to train the predictors is the expectation of the training poses for that node. Another choice is the coordinates of the dimensionality reduction of the corresponding poses obtained in the initialization.

Topic: visual processing and pattern recognition. Preference: poster/oral

References

[1] Lawrence, N. D. et al. *Hierarchical Gaussian process latent variable models*. ICML, 2007

[2] Kanaujia, A.; Sminchisescu, C. and Metxas, D. *Spectral Latent Variable Models for Perceptual Inference*. ICCV, 2007