

## Spectral Latent Variable Models for Perceptual Inference

Atul Kanaujia  
Rutgers University

Cristian Sminchisescu  
TTI-C and University of Bonn

Dimitris Metaxas  
Rutgers University

### Abstract

We propose non-linear generative models referred to as **Sparse Spectral Latent Variable Models (SLVM)**, that combine the advantages of spectral embeddings with the ones of parametric latent variable models: (1) provide stable latent spaces that preserve global or local geometric properties of the modeled data; (2) offer low-dimensional generative models with probabilistic, bi-directional mappings between latent and ambient spaces, (3) are probabilistically consistent (i.e., reflect the data distribution, both jointly and marginally) and efficient to learn and use. We show that SLVMs compare favorably with competing methods based on PCA, GPLVM or GTM for the reconstruction of typical human motions like walking, running, pantomime or dancing in a benchmark dataset. Empirically, we observe that SLVMs are effective for the automatic 3d reconstruction of low-dimensional human motion in movies.

### 1. Perceptual Models

A variety of computer vision and machine learning tasks require the analysis of high-dimensional ambient signals, e.g. 2d images, 3d range scans or data obtained from human motion capture systems. The goal is to learn compact, perceptual (latent) models of the data generation process and use them to interpret new measurements. For example, the variability in an image sequence filming a rotating teapot is non-linearly produced by latent factors like rotation variables and the lighting direction. Our subjective, *perceived* dimensionality partly mirrors the latent factors, being significantly smaller than the one directly *measured* – the high-dimensional sequence of image pixel vectors. Similarly, filming a human running or walking requires megabytes of wildly varying images, yet in a representation that properly correlates the human joint angles, the intrinsic dimensionality is effectively 1 – the phase of the walking cycle. The argument can go on, but underlines the intuitive idea that the space of all images is much larger than the set of physically possible ones, which, in turn is larger than the one typically observed in most every day's scenes. If this is true, perceptual inference cannot proceed without an appropriate,

arguably probabilistic model of correlation, a natural way to link perceptual inferences and measurements. This implies a non-linear subset, or a manifold assumption, at least in the large-sample regime: the low-dimensional perceptual structure lives in the high-dimensional space of direct observations. To unfold it, we need faithful, topographic representations of the observed data – effectively forms of continuity and locality: nearby observations should map to nearby percepts and faraway observations should map faraway. Given this, we want to be able to consistently answer the following questions: *How to represent a percept or an image? What is the probability of an observed image? What is the probability of a given percept? What is the conditional probability of a percept given an image and vice-versa?*

One promising class of methods for constructing non-linear perceptual representations given observed data is spectral dimensionality reduction [16, 19, 2, 7, 23]. The methods are similar in their use of graph-based representations of the data, with nodes standing for observations and links for neighborhood relations. The connected graph can be viewed as a discrete approximation of the sub-manifold directly sampled from observed data. Different methods derive different matrices from the graph and their spectral decompositions (the top or bottom eigenvectors) reveal the low-dimensional, perceptual structure of the data, and in some cases, also its intrinsic dimensionality. Spectral methods have been demonstrated to be capable of unfolding highly non-linear structure, and some methods (e.g. ISOMAP, Hessian and Laplacian Eigenmaps) come with additional, strong asymptotic guarantees – if enough samples are available, they could, in principle recover the true manifold from which the data was generated. However, spectral methods are *not* probabilistic and lack a consistent way to relate perceptual and observed quantities, or evaluate their probability. This explains, perhaps, why their use in computer vision has been limited, despite their undeniable intuitive appeal. On the other hand, a variety of probabilistic, non-linear latent variable models are available (mixture of PCA, Factor Analyzers, etc.), but they lack a global perceptual coordinate system and are not guaranteed to preserve intuitive geometric properties of the data in the latent space, as spectral methods do.

In this paper we introduce probabilistic models with geometric properties in order to marry spectral embeddings and parametric latent variable models, and obtain: (1) stable latent spaces that preserve structural properties of the ambient data, *e.g.* its local or global geometry; (2) low-dimensional generative models with probabilistic, bi-directional mappings between latent and ambient spaces, and (3) probabilistically consistent and efficient estimates. We refer to these probabilistic constructs, *implicitly* defined on top of an irregular distribution (or unfolding) obtained from a spectral embedding as Sparse Spectral Latent Variable Models (SLVM). We show how SLVMs can be used successfully for complex visual inference tasks, in particular the *automatic discriminative 3d reconstruction* of low-dimensional human poses in non-instrumented monocular video.

### 1.1. Related Work

Our research relates to work in spectral manifold learning, latent variable models and visual tracking. Spectral methods can model intuitive local or global geometric constraints [16, 19, 2, 7, 23] and their, local-optima free, polynomial time calculations are amenable to efficient optimization – either sparse eigenvalue problems, or dense problems that can be solved with algebraic multigrid methods [3, 17].

A variety of non-linear latent variable models exist *e.g.* mixtures of factor analyzers or PPCA [20]. These methods can model complicated non-linear structure but do not provide global latent coordinate systems or latent spaces which provably preserve local or global geometric properties of the data. Regular grid-based methods like the Generative Topographic Mapping GTM [5] do not scale beyond latent spaces higher than 2-3d, or for structured problems, where the latent space distribution is unlikely to be uniform (in fact GTM’s latent prior is a mixture of nodal delta functions – non-zero only on the grid). GTM is a useful non-linear method, yet it cannot unfold many convoluted manifolds (*e.g.* spirals, rolls) due to its data independent embedding grid, and local optima in training.<sup>1</sup> The Gaussian Process Latent Variable Model (GPLVM) [11] is a non-linear PCA technique based on a Gaussian Process mapping to ambient (data) space with zero mean unit Gaussian regularizer in latent space. Strictly, GPLVM defines a regularized conditional map to observed data, *not* a generative model. It is a competitive model primarily targeting data reconstruction error, but not designed to enforce the constraints we are after, *e.g.* the latent regularizer is data independent and geometric properties of ambient data are not explicitly preserved. GPLVM’s lack of latent space prior makes it some-

<sup>1</sup>Our model can be viewed a spectral generalization of the notable GTM precursor [5, 4], where we use a data-dependent, geometry preserving (rather than regular) embedding grid, sparsity constraints for the data map, a Gaussian mixture latent prior (as opposed to a delta sum) and MC sampling methods for training (as opposed to exact integration in GTM).

what agnostic in visual inference applications, where it is useful to penalize drifts from the manifold of typical configurations. A discussion and comparisons with both GPLVM and GTM appear in the experimental section. Memisevic [15] models the joint latent-ambient density using a separable product of non-parametric kernel density estimates and computes an embedding by optimizing a mutual information criterion over latent space coordinates – similar in spirit to GPLVM.

Recent work on visual tracking has identified the importance of low-dimensional models with intuitive geometric properties and latent-ambient mappings. Elgammal & Lee [8] fit an RBF to the corresponding points of an LLE-embedding, but their model devoid of a latent space prior is not fully probabilistic and there are no mappings to the latent space. In independent work, Sminchisescu & Jepson [17] augment spectral embeddings (*e.g.* Laplacian Eigenmaps) with both latent space priors and RBF mappings to ambient space. Their model is a latent variable one, but ambient to latent mappings are more difficult to compute, and the model is trained piece-wise. The model proposed here complements it. Urtasun *et al* [22] use GPLVM to track walking based on image tracks of the human joints obtained using the WSL tracker of Jepson’s *et al*. For more expressive kinematic representations, and in order to compensate for GPLVM’s lack of latent space prior, the authors [22] use an augmented, constrained (*latent, ambient*) state for tracking. This is feasible but, once again, renders the state estimation problem high-dimensional.

## 2. Spectral Latent Variable Models (SLVM)

We work with two sets of vectors,  $\mathcal{X}$  and  $\mathcal{Y}$ , of equal size  $N$  (initially in correspondence), in two spaces referred as latent (or perceptual) and ambient (or data). Both sets are un-ordered, hence there is no constraint to place either one on a regular grid (*i.e.* a matrix in 2d or the cells of cube in 3d). The latent space vectors are generically denoted  $\mathbf{x}$ , with  $\dim(\mathbf{x}) = d$ , the ambient vectors are  $\mathbf{y}$  with  $\dim(\mathbf{y}) = D$ , typically  $D \gg d$ .

**Spectral Embeddings:** Assume that vector-valued points in ambient space  $\mathcal{Y} = \{\mathbf{y}_i | i = 1 \dots N\}$  are captured from a high-dimensional process (images, sound, human motion capture systems), whereas corresponding latent space points  $\mathcal{X} = \{\mathbf{x}_i | i = 1 \dots N\}$  are initially obtained using any spectral, non-linear embedding method like ISOMAP, LLE, Hessian or Laplacian Eigenmaps, *etc.* The methods use graph-based representations of the observed data, with nodes that represent observations and links that stand for neighborhood relations. The connected graph can be viewed as a discrete approximation of the sub-manifold directly sampled from observed data. Different methods derive different matrices from the graph. Their spectral de-

compositions (the top or bottom eigenvectors) reveal the low-dimensional, latent structure of the data. We will use the distribution of latent points and a mapping to the ambient space in order to construct a joint probability distribution over latent and ambient variables.

**Latent Variable Models:** We model the joint distribution over latent and ambient variables using a constructive form:  $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$ . Without loss of generality we select the latent space prior  $p(\mathbf{x})$  to be a non-parametric kernel density estimate, with kernels  $K$  and covariance  $\theta$ , centered at embedded points  $\mathbf{x}_i$ , but more compact representations, e.g. Gaussian mixture models can be used instead:<sup>2</sup>

$$p(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^K K_{\theta}(\mathbf{x}, \mathbf{x}_i) \quad (1)$$

In the model, we assume that ambient vectors are related to the latent ones using a nonlinear vector-valued function  $\mathbf{F}(\mathbf{x}, \mathbf{W}, \alpha)$  with parameters  $\mathbf{W}$  and output noise covariance  $\sigma$ . Otherwise said, the joint distribution has a nonlinear constraint between two blocks of its variables:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{W}, \alpha, \sigma) \sim \mathcal{N}(\mathbf{y}|\mathbf{F}(\mathbf{x}, \mathbf{W}, \alpha), \sigma) \quad (2)$$

where  $\mathcal{N}$  is a Gaussian distribution with mean  $\mathbf{F}$  and covariance  $\sigma$ .  $\mathbf{F}$  is chosen to be a generalized (non-linear parametric) regression model:

$$\mathbf{F}(\mathbf{x}, \mathbf{W}, \alpha) = \mathbf{W}\phi(\mathbf{x}) \quad (3)$$

with  $\phi(\mathbf{x}) = [K_{\delta}(\mathbf{x}, \mathbf{x}_1), \dots, K_{\delta}(\mathbf{x}, \mathbf{x}_M)]^{\top}$ , having Gaussian kernels (other distributions can also be used) placed at an  $M$ -sized subset  $\mathbf{x}_i$  sampled from the prior  $p(\mathbf{x})$  (and selected automatically from a larger sample using a sparsity hyperprior, see below), with covariance  $\delta$ , and  $\mathbf{W}$  is a weight matrix of size  $D \times M$ .

The model is made computationally efficient and more robust to overfitting by using hierarchical priors on the parameters  $\mathbf{W}$  of the mapping  $\mathbf{F}$ , in order to select a sparse subset for prediction [21, 14]:

$$p(\mathbf{W}|\alpha) \sim \prod_{j=1}^D \prod_{k=1}^N \mathcal{N}(w_{jk}|0, \frac{1}{\alpha_k}) \quad (4)$$

$$p(\alpha) = \prod_{i=1}^N \text{Gamma}(\alpha_i|a, b) \quad (5)$$

with  $\text{Gamma}(\alpha|a, b) = \Gamma(a)^{-1} b^a \alpha^{a-1} e^{-b\alpha}$ , and  $a = 10^{-2}$ ,  $b = 10^{-4}$  chosen to give broad hyperpriors. The ambient marginal is obtained by integrating the latent space:

$$p(\mathbf{y}|\mathbf{W}, \alpha, \sigma) = \int p(\mathbf{y}|\mathbf{x}, \mathbf{W}, \alpha, \sigma)p(\mathbf{x})d\mathbf{x} \quad (6)$$

<sup>2</sup>This is one possible construction for the joint distribution, arguably not the only possible. Instead, one can approximate as product of marginals in latent and ambient space  $p(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^N K_{\theta}(\mathbf{x}, \mathbf{x}_i)K_{\theta}(\mathbf{y}, \mathbf{y}_i)$ , rely on index correspondences in the embedding, and optimize the latent coordinates  $\mathbf{x}_i$  in order to account for correlations in the joint [15].

The evidence, as well as derivatives w.r.t. model parameters, can be computed using a simple Monte Carlo (MC) estimate using, say  $K$ , samples from the prior.<sup>3</sup> This gives the MC estimate of the ambient marginal:

$$p(\mathbf{y}|\mathbf{W}, \alpha, \sigma) = \frac{1}{K} \sum_{i=1}^K p(\mathbf{y}|\mathbf{x}_i, \mathbf{W}, \alpha, \sigma) \quad (7)$$

The latent space conditional is obtained using Bayes' rule:

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \quad (8)$$

$$= \frac{p(\mathbf{y}|\mathbf{x}) \sum_{i=1}^K K_{\theta}(\mathbf{x}, \mathbf{x}_i)}{\sum_{i=1}^K p(\mathbf{y}|\mathbf{x}_i, \mathbf{W}, \alpha, \sigma)} \quad (9)$$

For pairs of ambient data points  $j$  and MC latent samples  $i$ , we abbreviate  $p_{(i,j)} = p(\mathbf{x}_i|\mathbf{y}_j)$ . Notice how the choice of prior  $p(\mathbf{x})$  influences the membership probabilities in (8) and (10). We can compute either the conditional mean or the mode (better for multimodal distributions) in latent space, using the same MC integration method used for (7):

$$\mathbb{E}\{\mathbf{x}|\mathbf{y}_n, \mathbf{W}, \alpha, \sigma\} = \int p(\mathbf{x}|\mathbf{y}_n, \mathbf{W}, \alpha, \sigma)\mathbf{x}d\mathbf{x} \quad (10)$$

$$= \sum_{i=1}^K p_{(i,n)}\mathbf{x}_i \quad (11)$$

$$i_{max} = \arg \max_i p_{(i,n)} \quad (12)$$

The model contains all the ingredients for efficient computation in both latent and ambient space: eq. (1) gives the prior in latent space, (7) the ambient marginal, (2) provides the conditional distribution (or mapping) from latent to ambient space, and (10) and (12) give the mean or mode of the mapping from ambient to latent space (a more accurate but also more expensive mode-finding approximation than (12) can be obtained by direct gradient ascent on (8)). Latent conditionals given partially observed  $\mathbf{y}$  vectors are easy to compute, using (8). The  $\mathbf{y}$  distribution is Gaussian and unobserved components can be integrated analytically – this effectively removes them from the mean and the corresponding lines and columns of the covariance. The model can be trained by maximizing the log-likelihood of the data:

$$\mathcal{L} = \log \prod_{i=1}^N p(\mathbf{y}_i|\mathbf{W}, \alpha, \sigma) = \quad (13)$$

$$= \sum_{n=1}^N \log \left\{ \frac{1}{K} \sum_{i=1}^K p(\mathbf{y}_n|\mathbf{x}_i, \mathbf{W}, \alpha, \sigma) \right\} \quad (14)$$

<sup>3</sup>The number of samples and the number of prior components are chosen co-incidentally, as  $K$ , for notational convenience, and so was the sample index  $\mathbf{x}_i$  to match the datapoint  $\mathbf{y}_i$ . In general, the latent MC samples are distributed according to the latent density, and may be different from the latent coordinates of datapoints obtained from the spectral embedding. Sampling from the kernel density estimate is efficient and can be done once for all – the same set can be reused used for all MC calculations, both training and testing.

Maximizing the likelihood provides estimates for  $\mathbf{W}$ ,  $\alpha$ ,  $\sigma$  (consider  $\sigma$  is diagonal with values  $\sigma$ ):

$$\Sigma = (\sigma \Phi^\top \mathbf{G} \Phi + \mathbf{S})^{-1} \quad (15)$$

$$\mathbf{W}^\top = \sigma \Sigma \Phi^\top \mathbf{R} \mathbf{Y} \quad (16)$$

where  $\mathbf{S} = \text{diag}(\alpha_1, \dots, \alpha_M)$  with  $\alpha$  corresponding only to the active set,  $\mathbf{G} = \text{diag}(G_1, \dots, G_K)$  with  $G_i = \sum_{j=1}^N p(i,j)$ ,  $\mathbf{R}$  is a  $K \times N$  matrix with elements  $p(i,j)$ , and  $\mathbf{Y}$  is an  $N \times D$  matrix that stores the output vectors  $\mathbf{y}_i, i = 1 \dots N$  row-wise, and  $\Phi$  is a  $K \times M$  matrix with elements  $\mathcal{G}(\mathbf{x}_i | \mathbf{x}_j, \theta)$ . The inverse variance is estimated from prediction error:

$$\sigma = \frac{1}{ND} \sum_{n=1}^N \sum_{k=1}^K p(kn) \|\mathbf{W}^* \phi(\mathbf{x}) - \mathbf{y}_n\|^2 \quad (17)$$

where a ‘\*’ superscript identifies an updated variable estimate. The hyperparameters are re-estimated using the relevance determination equations [14]:

$$\alpha_i^* = \frac{\gamma_i}{\|\boldsymbol{\mu}_i\|^2}, \quad \gamma_i = 1 - \alpha_i \Sigma_{ii}^* \quad (18)$$

where  $\boldsymbol{\mu}_i$  is the  $i$ -th column of  $\mathbf{W}$ . The algorithm is summarized in fig. 1.

### 3. Feedforward 3D Pose Prediction

We estimate a distribution over solutions in latent space, given input descriptors derived from images obtained by detecting the person using a bounding-box. We then map from latent states to 3d human joint angles (using *e.g.* (2)) in order to recover body configurations for visualization or error reporting. To predict latent space distributions from image features, we use a probabilistic conditional mixture of expert predictors (see [18]), here sparse Bayesian linear regressors. Each one is paired with an observation dependent gate (a softmax function with sparse linear regressor exponent) that scores its competence given different images. As these change, different experts may be active and their rankings (relative probabilities) may change. The model is trained using a double-loop EM algorithm [9, 18].

## 4. Experiments

We illustrate the SLVM on simple S-sheet and Swiss-roll toy datasets and experiment also with a computer vision application: the reconstruction of low-dimensional representations of 3d human body and facial poses from monocular images of people photographed against non-stationary backgrounds.

### 4.1. The S-Sheet and Swiss Roll

This set of experiments are illustrated in fig. 2. The original S-sheet data set (top)a) consists of 1000 points

**Input:** Set of high-dimensional, ambient points  $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1 \dots N}$ .

**Output:** Sparse, Spectral Latent Variable Model (SLVM), with parameters  $(\mathbf{W}, \alpha, \sigma)$  and latent space distribution  $\mathcal{X}$  that preserves local or global geometric properties of  $\mathcal{Y}$ .

**Step 1.** Compute spectral (non-linear) embedding of  $\mathcal{Y}$  to obtain corresponding latent points  $\mathcal{X} = \{\mathbf{x}_i\}_{i=1 \dots N}$ , using any standard embedding method like ISOMAP, LLE, HE, LE.

**Step 2.** Construct latent space prior as non-parametric kernel density estimate or Gaussian mixture *c.f.* (1).

**Step 3.** EM Algorithm to learn Latent Variable Model

- **Initialize**  $(\mathbf{W}_0, \alpha_0, \sigma_0)$  given  $p(i,i) = 1$  (hard assignment based on spectral correspondences  $(\mathbf{x}_i, \mathbf{y}_i)_{i=1 \dots N}$ ).
- **E-step:** Compute posterior probabilities  $p(i,n)$  for the assignment of latent to ambient space points using (8), based on Monte-Carlo sample from the prior  $p(\mathbf{x})$  (for efficiency, the same sample can be reused for training and testing – this will *not coincide*, in general, with the spectral coordinates of the ambient datapoints  $\mathbf{y}_i$ ).
- **M-step:** Solve weighted non-linear Bayesian regression problem to update  $(\mathbf{W}, \alpha, \sigma)$  according to (15), (17), and (18). This uses Laplace approximation for the hyperparameters and analytical integration for the weights, and optimization with greedy weight subset selection [14, 21].

Figure 1. The SLVM Learning Algorithm

sampled regularly on the surface of S-sheet and (in most plots) color coded to show their relative spatial positions. In fig.b), we show reconstruction using GTM. The reconstruction is not accurate and scrambles the geometric ordering when traversing the S sheet. GTM uses a regular grid and data is embedded on this grid. Fig.c), top row, shows accurate reconstructions from the sparse SLVM – the 16 (out of 1000) latent space centers automatically selected by the model are shown in fig.d). Fig. e) shows the SLVM marginal in ambient space, computed using (7), for the basis set automatically computed by the model. Notice that unlike all the other plots shown in the figure, the color of the points shows probability, *not* geometric ordering. The prior distribution peaks higher on the principal curve of the S-shape, away from the borders of the manifold (where there

is less data, on average). The **bottom row** of fig. 2 illustrates the computation of conditional distributions and the prior in latent space for SLVM and GPLVM. In fig.a) we show the Swiss roll sheet dataset together with an out-of-sample point (at one extremity) for which the conditional latent space distribution is computed. Fig. b) shows the intuitive bi-modal distribution in latent space, computed *c.f.* (8) (GPLVM gives a unimodal approximation, which can be either one of the two shown if the embedding were unfolded correctly). In fig.c) we show the SLVM latent distribution, computed using (1). In fig.d) and (e) we show Swiss-roll embeddings and latent regularizer from GPLVM without and with backconstraints – neither one is able to correctly unfold the Swiss roll. Notice the difference in latent distributions: SLVM reflects the data density; GPLVM has no latent prior, only a zero mean unit variance Gaussian regularizer for learning the data map.

## 4.2. 3D Human Pose Reconstruction

In this section, we report quantitative comparisons and qualitative 3d reconstruction (joint angles) of human motion from video or photographs. We use a 41d rotational joint angle representation of the three-dimensional skeleton which is mapped to a 82d (*sin, cos*) encoding which varies continuously as angles rotate over  $360^\circ$ . This is given as input ( $y$ ) for training all latent-variable models.

**Image Descriptor:** A difficulty for reliable discriminative pose prediction is the design of image descriptors that are distinctive enough to differentiate among different poses, yet invariant to ‘within the same pose class’ deformations or spatial misalignments – people in similar stances, but differently proportioned, or photographed on different backgrounds. We use a Multilevel Spatial Blocks (**MSB**) encoding, in order to compute a description of the image at multiple levels of detail. The putative image bounding-box of a person is split using a regular grid of overlapping blocks, and increasingly large SIFT [13] descriptor cell sizes are extracted. We concatenate SIFTs within each layer and across layers, orderly, in order to obtain encodings of an entire image or sub-window. For our problem we use MSBs of size 1344, obtained by decomposing the image into blocks at 3 different levels, with 16, 4, 1 SIFT block, 4x4 cells per block, 12x12 pixel cell size – 8 image gradient orientations histogrammed within each cell.

**Database:** For qualitative experiments we use images from a movie (Run Lola Run) and the INRIA pedestrian database [6]. For quantitative experiments we use our own database consisting of  $3 \times 3247 = 9741$  quasi-real images, generated using a computer graphics human model that was animated using the Maya graphics package, and rendered on real image backgrounds. We have 3247 different 3d poses from the CMU motion capture database [1] and these are rendered to produce supervised (low-dimensional 3d human joint angle,

MSB image descriptor) pairs for different patterns of walks, either viewed frontally or from one side, dancing, conversation, bending and picking, running and pantomime (one of the 3 training sets of 3247 poses is placed on a clean background). We collect three test sets of 150 poses for each of the five motion classes. The test motions are executed by different subjects, not in the training set. We also render one test set on a clean background to use as baseline. The other two test sets are progressively more complicated: one has the model *randomly* placed at different locations, but on the same images as in the training set, the other has the model placed on unseen backgrounds. In all cases, a 320x240 bounding box of the model and the background is obtained, possibly using rescaling. There is significant variability and lack of centering in this dataset because certain poses are vertically (and horizontally) more symmetric than others (*e.g.* compare a person who picks an object with one who is standing, or pointing the arm in one direction).

We train multivalued predictors (conditional Bayesian mixture of 5 sparse linear regressors) on each activity in the dataset (latent variable models are learned separately for each activity). We use sparse linear regressors, in order to automatically turn-off noisy image descriptor components perturbed by background clutter *c.f.* §3. The error is computed w.r.t. the most probable expert, but we plan to also study the error in best  $k$  experts. This should reduce variance when comparing LVMS.

In fig. 3, we show quantitative comparisons (prediction error, per joint angle, in degrees) for 5 different motions (+ a cumulative plot) and 3 different imaging conditions: Clean backgrounds, Clutter1 backgrounds and Clutter2 backgrounds, not seen at all in the training set. We compare several methods, including our SLVM-(ISOMAP, HE, LE, LLE), GPLVM with and without back-constraints [11], GTM [5] and PCA, all with 2 latent space dimensions embedded from  $41 \times 2 = 82$ d (*sin, cos*) encoding of ambient human joint angles (recall that in each case, both a separate LVM and an image-based latent state predictor are trained). For visualization and error reporting we use the conditional estimate of the ambient state (function  $\mathbf{F}$ ) to map latent point estimates  $x$  to joint angles  $y$ .

In our experiments, SLVM based on ISOMAP was the best performer, followed closely by Hessian Eigenmaps. GPLVM (with and without back-constraints) performed less well, but better than PCA. Local geometry preserving latent variable models based on LLE and LE didn’t perform as well as the other models. GTM in turn, gives significantly higher prediction error and has difficulty unfolding the high-dimensional human joint angle trajectories on its regular 2d grid. Dancing appears to be the hardest sequence for all models – primarily a training / testing issue: the motions are performed by different subjects and their intrinsic semantic variability is significantly higher – hence the

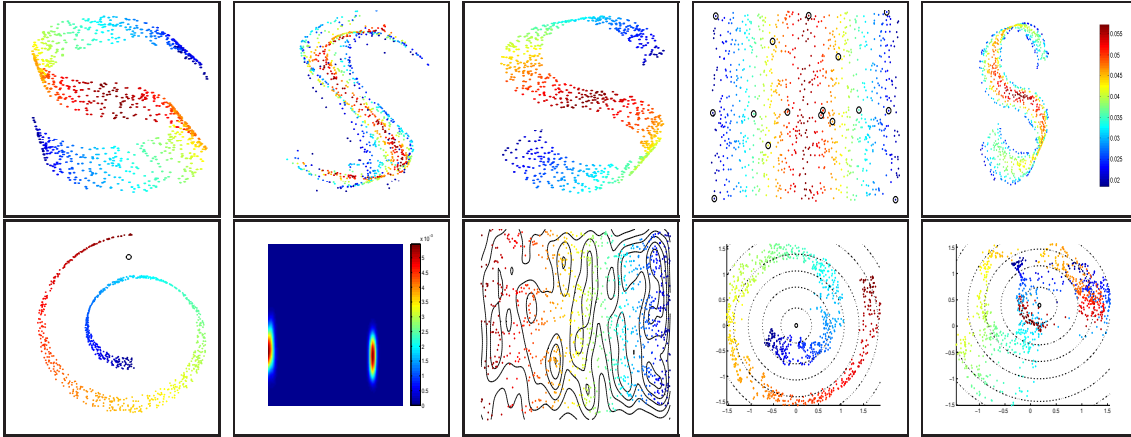


Figure 2. Analysis of SLVM, GTM and GPLVM on the S-Sheet and the Swiss roll. **Top Row:** (a) Original dataset of 1000 points, color coded to emphasize geometric ordering; (b) Data reconstructed (sampled) from GTM, with associations between datapoints color-coded – GTM scrambles the points and does not preserve their ambient geometric structure. The reconstruction is not accurate. (c) Reconstruction from our sparse SLVM model with associations color-coded; (d) Active (sparse) basis set with 1.6%=16 of the datapoints shown in latent space, as automatically selected by SLVM. (e) Ambient marginal of SLVM computed using (7) for the automatically selected basis in (d). Important: Notice that *unlike all the other plots shown in the figure the color of the points represents probability, not geometric ordering*. Notice higher probability on the principal curve. **Bottom row:** Computations of conditional distributions and prior in latent space by SLVM and GPLVM for the Swiss roll. (a) Dataset with ambient point for which SLVM conditional latent space distribution is computed. (b) The multimodal conditional distribution in latent space, computed *c.f.* (8). (c) Latent space structure and prior for SLVM. (d-e) Embeddings and latent priors computed by GPLVM without and with backconstraints. In this case, GPLVM cannot correctly unfold the Swiss roll. Notice differences in latent distributions: SLVM reflects the data density; GPLVM has no latent prior, only a data-independent, zero mean unit variance Gaussian regularizer (isocontours shown) for learning the ambient map.

motion trajectories are very different from the ones seen in training. Computationally, GPLVM is the most expensive model and PCA the cheapest to train, whereas in testing all models are about the same (earlier implementations of GPLVM we tested were a factor of 4 slower, but this has improved in the most recent version). SLVMs have competitive training times. A comparative table is shown in fig. 5. Another set of results we show in fig. 6 is based on

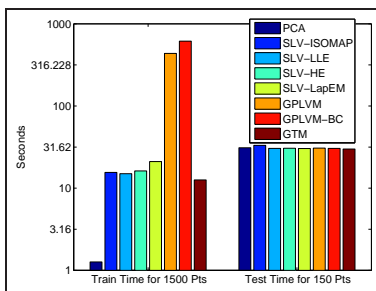


Figure 5. Training and test times for different LVMs.

real images from the INRIA pedestrian dataset [6] and the movie ‘Run Lola Run’. These are automatic 3d reconstructions obtained with our SLVM-Isomap. The humans are fast moving and filmed in non-instrumented environments. We use a model trained with 2000 walking and running poses *only* (quasireal data of our model placed on real backgrounds, rendered from 8 different viewpoints). As typical with many discriminative methods, the solutions are not al-

ways entirely accurate in a classical alignment sense (this is largely due to lack of typical training data) – these are nevertheless fully automatic reconstructions of a fast moving person (Lola), filmed with significant viewpoint and scale variability. Notice that the phase of the run and the synchronicity between arms and legs varies significantly across frames – naturally, we had no mean to train on Lola’s movement. Overall, we appreciate that the 3d reconstructions have reasonable perceptual accuracy.

**Face tracking and 3d head pose reconstruction:** The final set of results we present concern a different human sensing application – face tracking in monocular video fig. 7. We use a 2d face tracker based on a landmark representation – a 80d vector encoding 40 (x 2d) points. Different human face examples, both frontal and profiles, are used to learn a 2d SLVM-Isomap representing the variability of the high-dimensional set of landmarks. Each landmark is paired with an appearance descriptor (matched against nearby image regions using SSD), encoding its intensity distribution along the normal to the face contour, inside the face. The SLVM is used for face initialization and 2d tracking based on gradient descent in latent space. Fig. 7 shows the optimizer trajectory when fitting the image of a face profile – this is initialized at an average frontal face. As the face is tracked, the latent coordinate is given as input to a conditional mixture of experts to predict the global 3d face rotation.

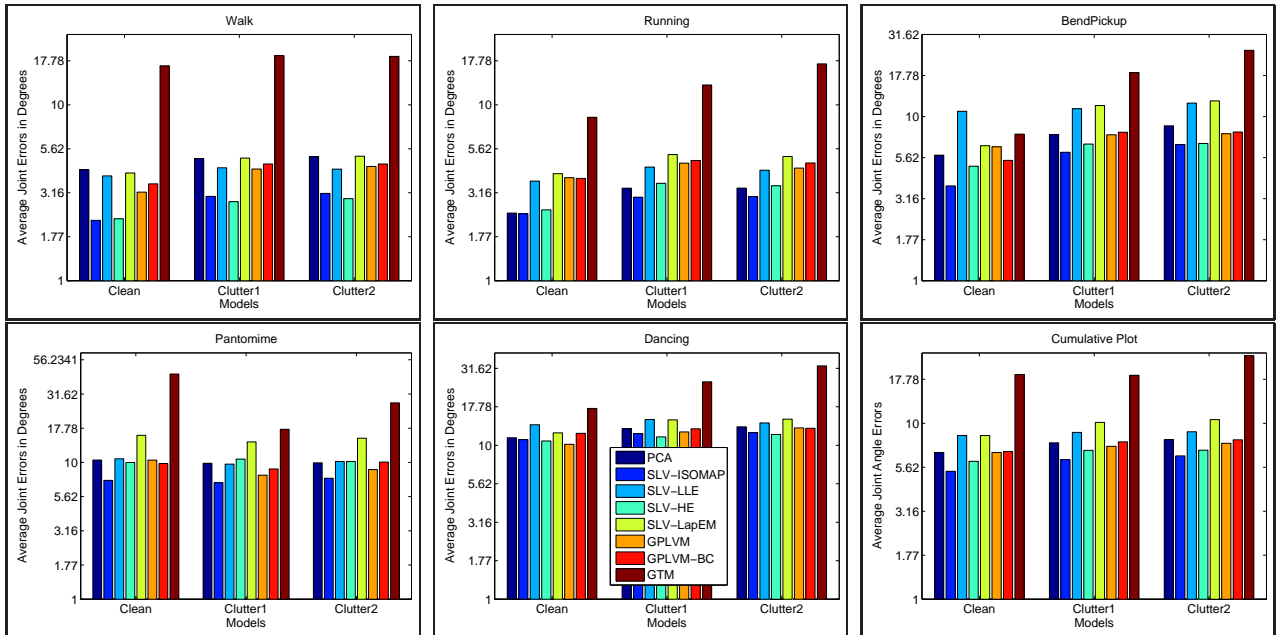


Figure 3. *Quantitative results* (prediction error, per joint angle, in degrees) for 5 different motions (+ a cumulative plot) and 3 different imaging conditions, as follows: people on Clean backgrounds, people on Clutter1 backgrounds, used in the training set (but the test image has the person in a different position w.r.t. the background, in a relative position *not* seen in the training set) and people on Clutter2 backgrounds, not seen at all in the training set. We compare several methods, including our SLVM with different spectral embeddings (ISOMAP, LLE, HE, LE), GPLVM (with and without back-constraints), GTM and PCA, all with 2 latent space dimensions. We use discriminative image-based predictors (multivalued mappings from images to latent space), based on conditional Bayesian mixtures of 5 experts, sparse linear regressors, for robustness to image descriptor components perturbed by variable background clutter *c.f.* §3. Both LVM models and corresponding predictors have been trained separately for each motion type. Error is computed w.r.t. the most probable expert.

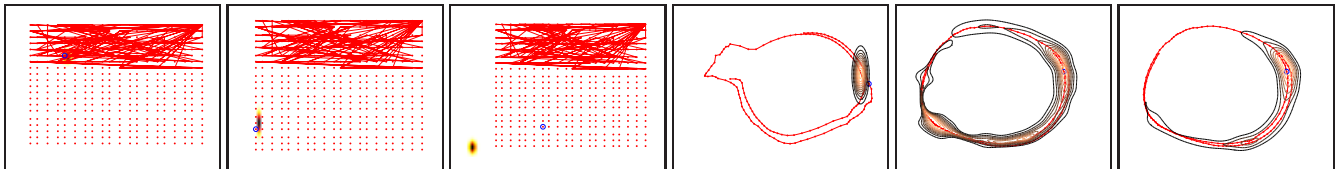


Figure 4. Posterior plots showing the predicted distribution  $p(\mathbf{x}|\mathbf{r})$  in latent space, given an image observable  $\mathbf{r}$  for GTM (a-c) defined on a regular 2d grid (with a spiky mixture of delta functions, each placed at a node) and a SLVM model (d). The topmost points in (a-c) are the training points and for illustration we link points corresponding to adjacent frames in the motion sequence. The ground truth is shown with a circle, the predicted posterior is color-coded. Half of the points on the grid have RBFs placed on top, regularly sub-sampled (*not* only top). Notice the loss of track, and the assignments of ambient points to multiple grid points (top). Figs (d) show image-based prediction from SLVM. Figs (e) and (f) show the conditional latent space distribution  $p(\mathbf{x}|\mathbf{y})$  for a SLVM walking model given only the left arm (shoulder and elbow, 5 out of 41 variables), the latent point corresponding to the complete vector of ‘ground truth’ joint angles is shown with a circle. Notice the 3 modes that arise due to missing data. Fig (f) shows  $p(\mathbf{x}|\mathbf{y})$  for the right leg (5 variables out of 41 given). This is more informative than the arm – the conditional is unimodal.

## 5. Conclusions

We have presented spectral latent variable models (SLVM) and showed their potential for visual inference applications. We have argued in support of low-dimensional models that: (1) preserve intuitive geometric properties of the ambient distribution, *e.g.* locality, as required for visual tracking applications; (2) provide mappings, or more generally multimodal conditional distributions between latent and ambient spaces, and (3) are probabilistically consistent, efficient to learn and estimate and applicable with any spec-

tral non-linear embedding method like ISOMAP, LLE or LE. To make (1)-(3) possible, we propose models that combine the geometric and computational properties of spectral embeddings with the probabilistic formulation and the mappings offered by latent variable models. We demonstrate quantitatively that SLVMs compare favorably with existing linear and non-linear techniques and show empirically that (in conjunction with discriminative pose prediction methods and multilevel image encodings), SLVMs are effective for the *automatic 3d reconstruction* of low-dimensional human poses from non-instrumented monocular images.



Figure 6. Qualitative 3d reconstruction results obtained on images from the movie ‘Run Lola Run’ (block of leftmost 3 images) and the INRIA pedestrian dataset (rightmost 2 images) [6]. (a) Top row shows the original images, (b) Bottom row shows automatic 3d reconstructions.

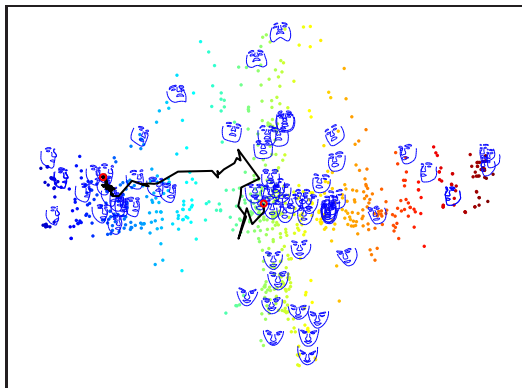


Figure 7. The trajectory of a gradient descent optimizer in the latent space of our SLVM-Isomap. The model is initialized at a frontal face and used to fit the image of a face profile.

**Future work:** We currently investigate alternative approximations to the latent space prior distribution as well as alternative GP mappings and constraints between latent and ambient points. We also plan to study the behavior of our algorithms when unfolding more complex structures, e.g. motion combinations and higher dimensional latent spaces, where non-linear models are expected to perform best.

**Acknowledgements:** This work has been supported in part by the NSF and the EC, under awards IIS-0535140 and MCEXT-025481.

## References

[1] CMU Human Motion Capture DataBase. Available online at <http://mocap.cs.cmu.edu/search.html>, 2003. 5

[2] M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *NIPS*, 2002. 1, 2

[3] J. Bengio, J. Paiement, and P. Vincent. Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps and Spectral Clustering. In *NIPS*, 2003. 2

[4] C. Bishop, M. Svensen, and C. K. I. Williams. Developments of the generative topographic mapping. *Neurocomputing*, (21):203–224, 1998. 2

[5] C. Bishop, M. Svensen, and C. K. I. Williams. Gtm: The generative topographic mapping. *Neural Computation*, (1):215–234, 1998. 2, 5

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 5, 6, 8

[7] D. Donoho and C. Grimes. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-dimensional Data. *Proc. Nat. Acad. Arts and Sciences*, 2003. 1, 2

[8] A. Elgammal and C. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR*, 2004. 2

[9] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, (6):181–214, 1994. 4

[10] A. Kanaujia, C. Sminchisescu, and D. Metaxas. Sparse Spectral Latent Variable Models for Perceptual Inference. Technical Report DCS-TR-610, Rutgers University, February 2007.

[11] N. Lawrence. Probabilistic non-linear component analysis with gaussian process latent variable models. *JMLR*, (6):1783–1816, 2005. 2, 5

[12] R. Li, M. Yang, S. Sclaroff, and T. Tian. Monocular Tracking of 3D Human Motion with a Coordinated Mixture of Factor Analyzers. In *ECCV*, 2006.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004. 5

[14] D. Mackay. Comparison of Approximate Methods for Handling Hyperparameters. *Neural Computation*, 11(5), 1998. 3, 4

[15] R. Memisevic. Kernel Information Embeddings. In *ICML*, 2006. 2, 3

[16] S. Roweis and L. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 2000. 1, 2

[17] C. Sminchisescu and A. Jepson. Generative Modeling for Continuous Non-Linearly Embedded Visual Inference. In *ICML*, pages 759–766, Banff, 2004. 2

[18] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3D Human Motion Estimation. In *CVPR*, volume 1, pages 390–397, 2005. 4

[19] J. Tenenbaum, V. Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 2000. 1, 2

[20] M. Tipping. Mixtures of probabilistic principal component analysers. *Neural Computation*, 1998. 2

[21] M. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *JMLR*, 2001. 3, 4

[22] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking in small training sets. In *ICCV*, 2005. 2

[23] K. Weinberger and L. Saul. Unsupervised Learning of Image Manifolds by Semidefinite Programming. In *CVPR*, 2004. 1, 2