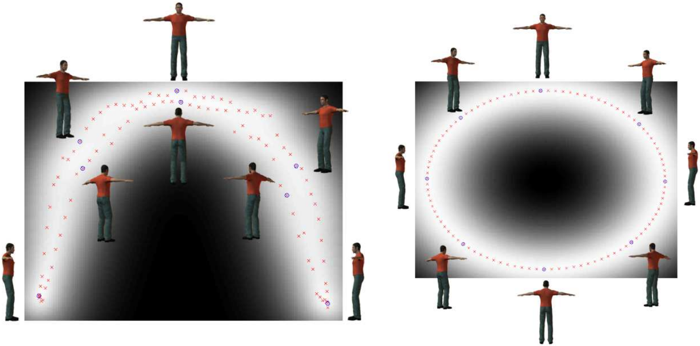
Ambiguity Modeling in Latent Spaces

Carl Henrik Ek, Jon Rihan, Philip H. S. Torr, Grégory Rogez
and **Neil D. Lawrence**

Machine Learning for Multimodal Interfaces, Utrecht, Netherlands

September 11, 2008

Motivation: Static pose rotating 360°

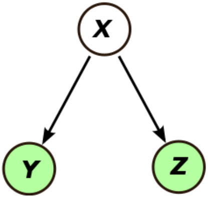


- Data consists of actual pose and features derived from silhoutte (data artificially generated in Poser)
- Visualization on the left from silhouette features. Visualization on the right from pose features.

Our Approach

- Reduce dimensionality of the data.
  - Non linear dimensionality reduction.
  - Underlying assumption that data is *really* low dimensional — *e.g.* a prototype with non-linear distortions.
- Fusion of different modalities.
  - Concatanate data observations
  - $\mathbf{Y} = [\mathbf{y}_1 \ldots \mathbf{y}_N]^{\mathrm{T}} \in \Re^{N \times D_Y}$ (silhouette)
  - $\mathbf{Z} = [\mathbf{z}_1 \ldots \mathbf{z}_N]^{\mathrm{T}} \in \Re^{N \times D_Z}$ (pose).

Fusion of the Data

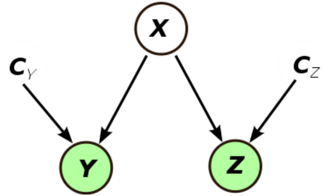


- Assume data sets have intrinsic low dimensionality, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathrm{T}}$ where $\mathbf{x}_n \in \Re^q$, $q \ll D_y$ and $q \ll D_z$.

$$y_{ni} = f_i^Y(\mathbf{x}_n) + \epsilon_{ni}^Y, \quad z_{ni} = f_i^Z(\mathbf{x}_n) + \epsilon_{ni}^Z.$$

- For Gaussian process priors over $f_i^Y(\cdot)$ and $f_i^Z(\cdot)$ this is a shared latent space variant of the GP-LVM (Shon et al., 2006; Ek et al., 2007; Navaratnam et al., 2007).

Probabilistic CCA



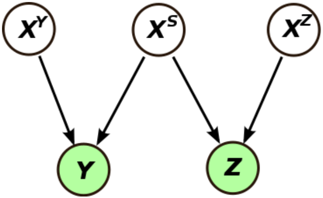
- If $f_i(\cdot)$ are taken to be linear and

$$\epsilon_n \sim N(\mathbf{0}, \mathbf{C})$$

this model is probabilistic canonical correlates analysis (Bach and Jordan, 2005).

- For non-linear $f_i^{\cdot}(\cdot)$ with Gaussian process priors we have GPLVM-CCA (Leen and Fyfe, 2006).

New Model

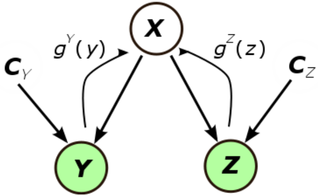


$$y_{ni} = f_i^Y\left(\mathbf{x}_n^{\mathrm{S}}, \mathbf{x}_n^Y\right) + \epsilon_{ni}^Y, \quad z_{ni} = f_i^Z\left(\mathbf{x}_n^{\mathrm{S}}, \mathbf{x}_n^Z\right) + \epsilon_{ni}^Z,$$

- The mappings are occurring from a latent space which is split into three parts, $\mathbf{X}^Y = \left\{\mathbf{x}_n^Y\right\}_{n=1}^N$, $\mathbf{X}^Z = \left\{\mathbf{x}_n^Z\right\}_{n=1}^N$ and $\mathbf{X}^S = \left\{\mathbf{x}_n^{\mathrm{S}}\right\}_{n=1}^N$.
- The[1] $\mathbf{X}^Y$ and $\mathbf{X}^Z$ take the role of $\mathbf{C}^Z$ and $\mathbf{C}^Y$.

[1]For linear mappings and $q^Y = D^Y - 1$ and $q^Z = D^Z - 1$ CCA is recovered.

Non Linear CCA



- Kernel-CCA (see *e.g.* Kuss and Graepel, 2003) implicitly assumes that there is a smooth mapping from each of the data-spaces to a shared latent space,

$$x^{\mathrm{s}}_{ni} = g^Y_i(\mathbf{y}_n) = g^Z_i(\mathbf{z}_n).$$

- We augment CCA to extract private spaces, $\mathbf{X}^Y$ and $\mathbf{X}^Z$.
- To do this we make further assumption about the non-consolidating subspaces,

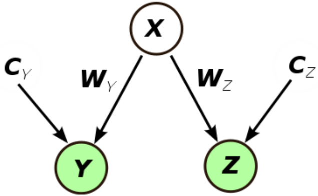
$$x^Y_{ni} = h^Y_i(\mathbf{y}_n), \quad x^Z_{ni} = h^Z_i(\mathbf{z}_n),$$

where $h^Y_i(\cdot)$ and $h^Z_i(\cdot)$ are smooth functions.

Initialize the GP-LVM

- Spectral methods used to initialize the GP-LVM (Lawrence, 2005).
- Harmeling (2007) observed that high quality embeddings are backed up by high GP-LVM log likelihoods.
- First step: apply kernel CCA to find shared sub-space.

Canonical Correlates Analysis



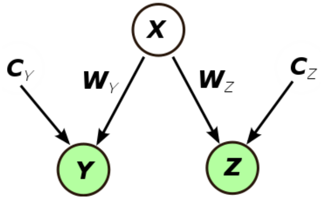
- Find linear transformations $\mathbf{W}_Y$ and $\mathbf{W}_Z$ maximizing the correlation between $\mathbf{W}_Y\mathbf{Y}$ and $\mathbf{W}_Z\mathbf{Z}$.

$$\{\hat{\mathbf{W}}_Y, \hat{\mathbf{W}}_Z\} = \mathrm{argmax}_{\{\mathbf{W}_Y,\mathbf{W}_Z\}} \mathrm{tr}\left(\mathbf{W}_Y^{\mathrm{T}} \Sigma_{YZ} \mathbf{W}_Z\right)$$

$$\mathrm{s.t.tr}\left(\mathbf{W}_Y^{\mathrm{T}} \Sigma_{YY} \mathbf{W}_Y\right) = \mathbf{I} \;\; \mathrm{tr}\left(\mathbf{W}_Z^{\mathrm{T}} \Sigma_{ZZ} \mathbf{W}_Z\right) = \mathbf{I}$$

the optima is found through an eigenvalue problem.

Non Linear Canonical Correlates Analysis



- We apply CCA in the dominant principal subspace of each feature space instead of directly in the feature space (Kuss and Graepel, 2003).
- Applying CCA recovers two sets of bases $\mathbf{W}_Y$ and $\mathbf{W}_Z$ explaining the correlated or shared variance between the two feature spaces.

NCCA I

- Need to describe private subspaces ($\mathbf{X}^Z$,$\mathbf{X}^Y$).
- Look for directions of maximum data variance that are *orthogonal* to the canonical correlates.
- Call the procedure *non-consolidating components analysis* (NCCA).
- Seek the first direction $\mathbf{v}_1$ of maximum variance orthogonal to $\mathbf{W}$.

$$\mathbf{v}_1 = \mathrm{argmax}_{\mathbf{v}_1} \mathbf{v}_1^{\mathrm{T}} \mathbf{K} \mathbf{v}_1$$

subject to: $\mathbf{v}_1^{\mathrm{T}} \mathbf{v}_1 = 1$ and $\mathbf{v}_1^{\mathrm{T}} \mathbf{W} = \mathbf{0}$.

- The optimal $\mathbf{v}_1$ is found via an eigenvalue problem,

$$\left(\mathbf{C} - \mathbf{W}\mathbf{W}^{\mathrm{T}} \mathbf{K} \right) \mathbf{v}_1 = \lambda_1 \mathbf{v}_1.$$

NCCA II

- For successive directions further eigenvalue problems of the form

$$\left(\mathbf{K} - \left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sum_{i=1}^{k-1} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}}\right) \mathbf{K}\right) \mathbf{v}_k = \lambda_k \mathbf{v}_k$$

  need to be solved.
- Embeddings then take form:

$$\mathbf{X}^S = \frac{1}{2}\left(\mathbf{W}_Y \mathbf{F}_Y + \mathbf{W}_Z \mathbf{F}_Z\right) \quad (1)$$

$$\mathbf{X}^Y = \mathbf{V}_Y \mathbf{F}_Y; \qquad \mathbf{X}^Z = \mathbf{V}_Z \mathbf{F}_Z, \quad (2)$$

  where $\mathbf{F}_Y$ and $\mathbf{F}_Z$ represent the kernel PCA representation of each observation space.

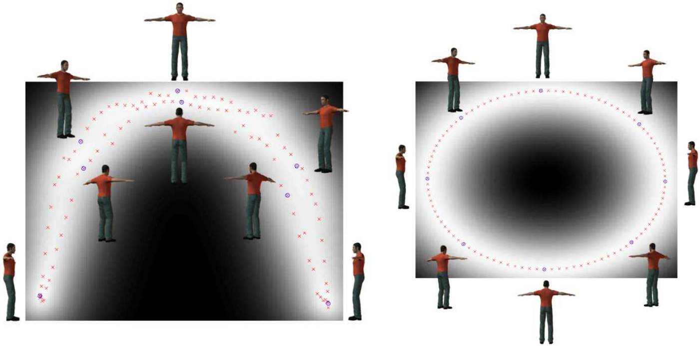
Initialization of a GP-LVM

- Purely spectral algorithm: the optimization problems are convex and they lead to unique solutions.
- Spectral methods are less useful in "inquisition" of the model.
- The pre-image problem means that handling missing data can be rather involved (Sanguinetti and Lawrence, 2006).
- Build Gaussian process mappings from the latent to the data space.
- This results in a GP-LVM model.

Inference I

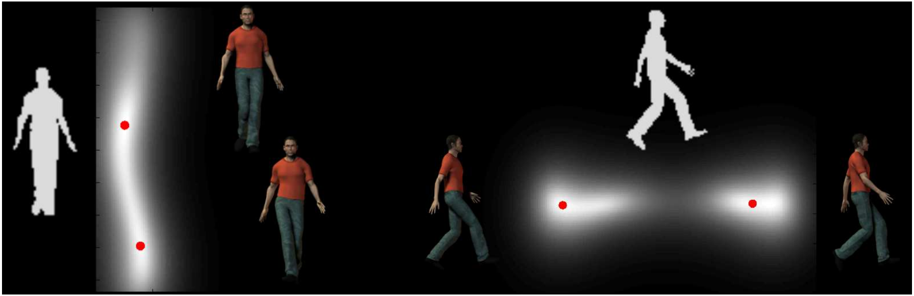
- Given a silhouette ($\mathbf{y}_*$), we can find the corresponding $\mathbf{x}_*^S$ position.
- The likelihood of different poses ($\mathbf{z}_*$) can then be visualized in the private space for the poses, $\mathbf{x}_*^Z$.
- Disambiguation (not dealt with here) can then be achieved through *e.g.* temporal information.

Motivation



x-axes are the shared space for the two models and the y-axes are the private space for the silhouettes (left) and the pose (right). Shading is from the GP-LVM likelihood.

Toy Problem Result



- Pose inference from silhouette using two different silhouettes from the training data.
- *Left* image: continuous leg ambiguity.
- *Right* image: discrete leg ambiguity.

Experiments

- A walking sequence from the HumanEva database (Sigal and Black, 2006).
  - Four cycles in a circular walk.
  - Use two for training and two for testing for the same subject.
  - Each image is represented using a 100 dimensional integral HOG descriptor (Zhu et al., 2006).
  - Represent the pose space as the sum of a MVU kernel (Weinberger et al., 2004) applied to the full pose space and a linear kernel applied on the local motion.
  - Represent the HOG features with an MVU kernel.
- On HumanEva: one dimensional shared space explaining data variance: 9% image space. 18% pose space.
- To retain 95% of the total variance in each observation two dimensions are needed for private spaces.

Pose Specific Latent Space

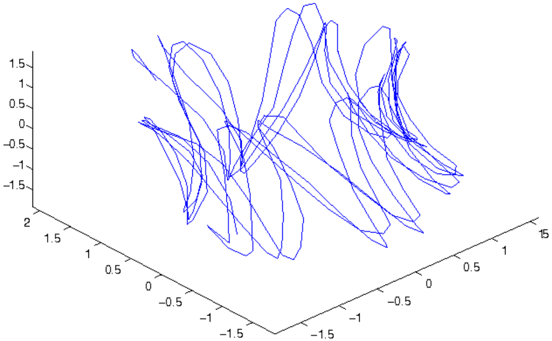
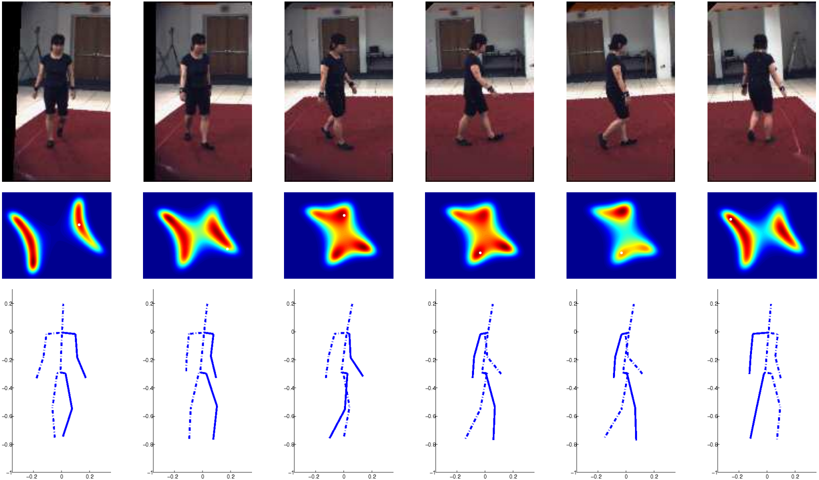


Figure: The latent space for the pose.

Experiments

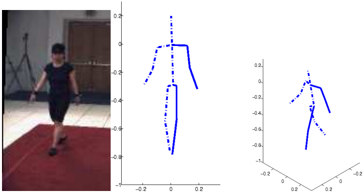
- Computation time about 10 minutes on a Intel Core Duo with 1GB of RAM.
- Inference procedure using 20 nearest neighbor initializations per image took a few seconds to compute.
- Comparison with shared GPLVM.

HumanEva Sequence Results

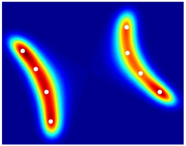


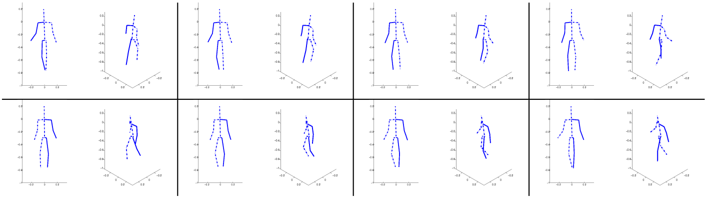
- *Top row*: original test set image. *Second row*: visualisation of ambiguities. *Bottom row*: pose from mode closest to ground truth.

HumanEva — Mode Exploration I

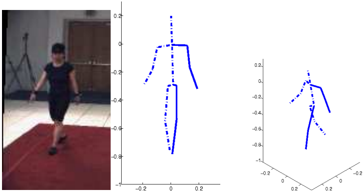


NCCA

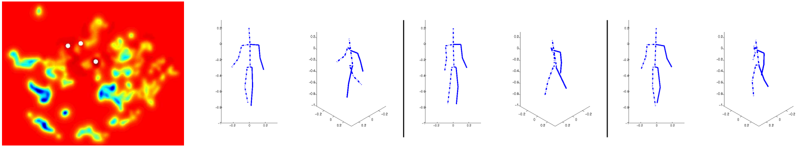




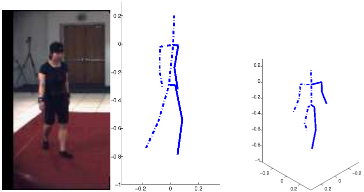
HumanEva — Mode Exploration I



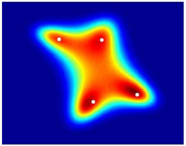
Shared GP-LVM

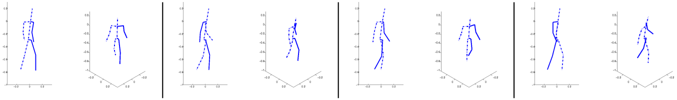


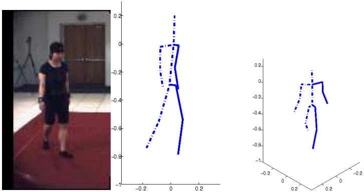
HumanEva — Mode Exploration II



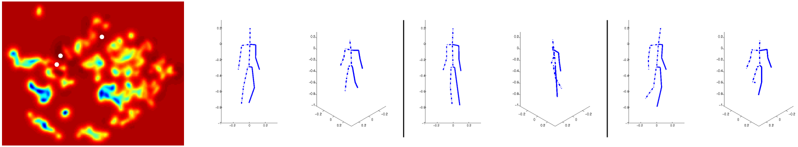
NCCA







Shared GP-LVM



Discussion

- Careful fusion of multimodal data at training stage allows for elegant disambiguation when only part of the data is available at test time.
- Further work:
  - Refinement with GPLVM algorithm.
  - Disambiguation with temporal information.

References I

F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, [PDF].

C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In *4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2007)*, volume LNCS 4892, pages 132–143, Brno, Czech Republic, Jun. 2007. Springer-Verlag.

S. Harmeling. Exploring model selection techniques for nonlinear dimensionality reduction. Technical Report EDI-INF-RR-0960, University of Edinburgh, [PDF].

M. Kuss and T. Graepel. The geometry of kernel canonical correlation analysis. Technical Report TR-108, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, [PDF].

N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian Process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, 2005. ISSN 1533-7928. [URL].

References II

G. Leen and C. Fyfe. A Gaussian process latent variable model formulation of canonical correlation analysis. Bruges (Belgium), 26-28 April 2006 2006. [PDF].

R. Navaratnam, A. Fitzgibbon, and R. Cipolla. The joint manifold model. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.

G. Sanguinetti and N. D. Lawrence. Missing data in kernel pca. In *ECML*, Lecture Notes in Computer Science, Berlin, 2006. Springer-Verlag.

A. Shon, K. Grochow, A. Hertzmann, and R. Rao. Learning shared latent structure for image synthesis and robotic imitation. *Proc. NIPS*, pages 1233–1240, 2006.

L. Sigal and M. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 2006.

K. Weinberger, F. Sha, and L. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. *ACM International Conference Proceeding Series*, 2004.

Q. Zhu, S. Avidan, M. Yeh, and K. Cheng. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. *CVPR*, 1(2):4, 2006.