Ambiguity Modeling in Latent Spaces

Carl Henrik Ek[1], Jon Rihan[1], Philip H.S. Torr[1], Grégory Rogez[2], and Neil D. Lawrence[3]

[1] Oxford Brookes University, UK
{cek,jon.rihan,philiptorr}@brookes.ac.uk
[2] University of Zaragoza, Spain
grogez@unizar.es
[3] University of Manchester, UK
Neil.Lawrence@manchester.ac.uk

Abstract. We are interested in the situation where we have two or more representations of an underlying phenomenon. In particular we are interested in the scenario where the representation are complementary. This implies that a single individual representation is not sufficient to fully discriminate a specific instance of the underlying phenomenon, it also means that each representation is an ambiguous representation of the other complementary spaces. In this paper we present a latent variable model capable of consolidating multiple complementary representations. Our method extends canonical correlation analysis by introducing additional latent spaces that are specific to the different representations, thereby explaining the full variance of the observations. These additional spaces, explaining representation specific variance, separately model the variance in a representation ambiguous to the other. We develop a spectral algorithm for fast computation of the embeddings and a probabilistic model (based on Gaussian processes) for validation and inference. The proposed model has several potential application areas, we demonstrate its use for multi-modal regression on a benchmark human pose estimation data set.

1 Introduction

A common situation in machine learning is the consolidation of two disparate, but related, data sets. Examples include: consolidation of lip movement with cepstral coefficients for improving the quality of robust speech recognition; consolidation of two different language renderings of the same document for cross language information retrieval; and consolidation of human pose data with image information for marker-less motion capture.

Formally, we will consider the situation where we are provided with two data sets, $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_N]^{\mathrm{T}} \in \Re^{N \times D_Y}$ and $\mathbf{Z} = [\mathbf{z}_1 \dots \mathbf{z}_N]^{\mathrm{T}} \in \Re^{N \times D_Z}$, for which there is some kind of correspondence between each point. For example, each measurement could have been taken at the same time or under the same experimental conditions. We are interested in answering questions about the relationship $\mathbf{z}_n$ and $\mathbf{y}_n$. For example: what is the most likely $\mathbf{z}_n$, given $\mathbf{y}_n$? This question can be answered by direct modeling of the conditional probability $p(\mathbf{z}_n|\mathbf{y}_n)$. However, this distribution can be very complex in practice. If we for example used a regression model, it would only be valid if the

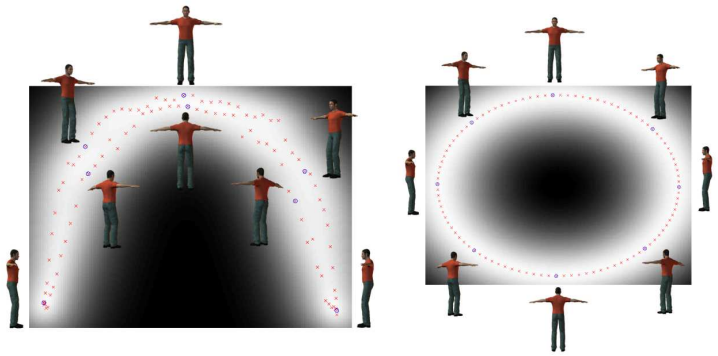

Fig. 1. *Latent representation associated with a static pose rotating* 360° *and its corresponding silhouette image features. The* x*-axis represents the dimension that is common to both spaces. The* y*-axis is* image feature specific *on the* **left** *and* pose feature specific *on the* **right**. *We have also used the GP-LVM model to associate each location in latent space with a likelihood. White represents high and black low regions of likelihood. Note the ambiguities in pose associated with the image feature space (ambiguous poses have similar* x *and* y *positions). In the pose space these ambiguities are resolved in the* y *axis.*

relationship between the observations was unimodal, this is often an invalid assumption. Multi-modalities that arise are a manifestation of a non-bijective relationship between $\mathbf{y}_n$ and $\mathbf{z}_n$, one that is difficult to express in a standard regression model. We could turn to a model for conditional probability estimation that allows for multi-modalities [13]. However, the nature of the multi-modal relationship is likely to be difficult to learn when the size of the data set is restricted. In this paper we propose an alternative approach, one that is based explicitly on assumptions about the relationship between $\mathbf{y}_n$ and $\mathbf{z}_n$. In particular we will assume that the data is generated by a lower dimensional latent variable. The approach is similar in character to that of canonical correlation analysis (CCA) with one key difference: the latent space associated with CCA describes only the characteristics of the data that are common to both the representations. We will construct a latent space that represents the *full data set*. We will subdivide the latent space into three *non-overlapping* partitions. One partition will be associated only with the $\mathbf{Y}$ data another partition is associated only with the $\mathbf{Z}$ data and the remaining partition is associated with the common or shared information between $\mathbf{Y}$ and $\mathbf{Z}$. The remaining non-shared or private latent subspaces model information not present in the corresponding observation space. This means when estimating $\mathbf{z}_n$ from $\mathbf{y}_n$ the private space represents the ambiguities of $\mathbf{z}_n$ when presented with $\mathbf{y}_n$

A simple example of such an ambiguity is given in Figure 1 where the the proposed model has been applied to a toy data set of a rotated character. The x-axis direction in both plots is shared for both pose and silhouette. The y-axis in the left plot represents information specific to the silhouette, while in the in the right plot, information specific to the pose. When looking at the information in the x axis only, the pose is ambiguous.

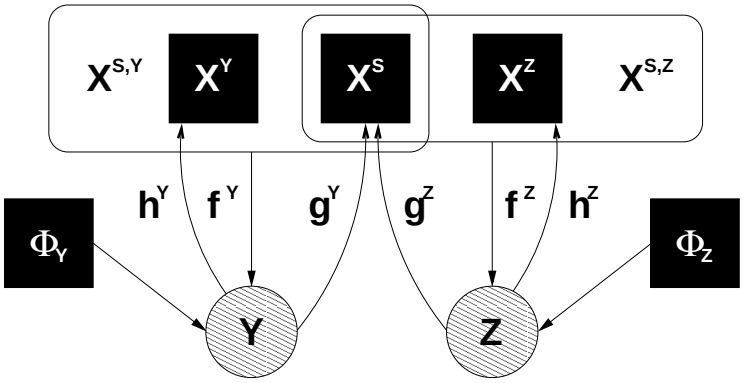

Fig. 2. *Graphical model of the NCCA Model. The two observations* $\mathbf{Y}$ *and* $\mathbf{Z}$ *are generated from low-dimensional embeddings* $\mathbf{X}^{S,Y}$ *and* $\mathbf{X}^{S,Z}$ *indicated by rounded rectangles. The embeddings share a common subspace* $\mathbf{X}^{S}$ *representing the shared variance in each observation space. This is variance in* $\mathbf{Y}$ *and* $\mathbf{Z}$ *that can be described as a function of* $\mathbf{Z}$ *and* $\mathbf{Y}$ *respectively. An additional subspace* $\mathbf{X}^{Y}$ *and* $\mathbf{X}^{Z}$ *completes the embedding, representing the non-shared variance between the observations.* Φ_Y *and* Φ_Z *collects the parameters associated with each mapping.*

However, in the right plot (from the motion capture) the pose is disambiguated on the y-axis, *i.e.* each pose is associated with a single location. The y-axis does not help in disambiguation in the left plot (which encodes silhouette information). Clearly, augmenting the latent space with a direction representing the 'private information' will be vital in disambiguating the pose from the silhouette.

Outline of the paper: In the next section we will present the NCCA model for data consolidation, we will then show results on both real and synthetic data in Section 3 followed by conclusions in Section 4.

2 The NCCA Model

Given two sets of corresponding observations $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^{\mathrm{T}}$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^{\mathrm{T}}$ where $\mathbf{y}_n \in \Re^{D_y}$ and $\mathbf{z}_n \in \Re^{D_z}$ we wish to characterize the relationship between the data sets through a latent variable model. We will assume that the two data sets can be generated by noise corrupted smooth functions that map from the latent space to the data-spaces in the following way,

$$y_{ni} = f_i^Y\left(\mathbf{x}_n^{\mathrm{s}}, \mathbf{x}_n^Y\right) + \epsilon_{ni}^Y, \quad z_{ni} = f_i^Z\left(\mathbf{x}_n^{\mathrm{s}}, \mathbf{x}_n^Z\right) + \epsilon_{ni}^Z, \tag{1}$$

where $\{y, z\}_{ni}$ represent dimension i of point n and ϵ_{ni}^Y, ϵ_{ni}^Z are sampled from a zero mean Gaussian distribution.

Distance preserving approaches to dimensionality reduction typically imply that there is a smooth mapping in the *reverse direction*. In particular, kernel-CCA [4] implicitly assumes that there is a smooth mapping from each of the data-spaces to a shared latent space,

$$x_{ni}^{\mathrm{s}} = g_i^Y\left(\mathbf{y}_n\right) = g_i^Z\left(\mathbf{z}_n\right). \tag{2}$$

Algorithm 1 NCCA Consolidation

Input:
$\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N],\ \mathbf{y}_i \in \Re^{D_\mathbf{Y}}$
$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N],\ \mathbf{z}_i \in \Re^{D_\mathbf{Z}}$
$\mathbf{K}_Y, \mathbf{K}_Z$

Stage 1, Learn latent embedding
Find kernel spaces from $\mathbf{K}_y$ and $\mathbf{K}_Z$ by kernel PCA:
1) Apply **CCA** to find shared embedded data $\mathbf{X}^{\mathrm{S}}$, Eq. (8)
2) Apply **NCCA** to find non-shared embedded data
$\mathbf{X}^{\mathrm{Y}}$ and $\mathbf{X}^{\mathrm{Z}}$, Eq. (9)

Stage 2, Learn mappings, *let J be either Y or Z,*
By GP-regression find:
1) Generative maps: $f^J : [\mathbf{X}^S; \mathbf{X}^J] \rightarrow \mathbf{J}$ Eq. (10)
2) Shared maps: $g^J : \mathbf{J} \rightarrow \mathbf{X}^S$ Eq. (11)
3) Non-shared maps: $h^J : \mathbf{J} \rightarrow \mathbf{X}^J$ Eq. (12)
See Figure 2

Return:
Pose Estimation: g^Y and f^Z
General Case: All maps learned above.

However, CCA does not characterize the nature of the other latent subspaces, $\mathbf{X}^Y$ and $\mathbf{X}^Z$. In Section 2.1 we will introduce an algorithm for extracting these spaces which we refer to as the *non-consolidating* subspaces. Underpinning the algorithm will be a further assumption about the non-consolidating subspaces,

$$x^Y_{ni} = h^Y_i(\mathbf{y}_n), \quad x^Z_{ni} = h^Z_i(\mathbf{z}_n), \tag{3}$$

where $h^Y_i(\cdot)$ and $h^Z_i(\cdot)$ are smooth functions. A graphical representation of the consolidation model is shown in Figure 2. Our approach will be as follows, we will construct a model by assuming the smooth mappings in (2) and (3) hold. We will then validate the model quality through assessing how well the resulting embeddings respect (1). We are inspired in our approach by the suggestion that spectral methods are used to initialize the GP-LVM in [5] and by the observation of [3] that the quality of an embedding is nicely indicated by the log likelihood of the GP-LVM.

To allow for non-linear relationships in the data we will first represent the observations in kernel induced feature spaces $\Psi_Y : Y \rightarrow \mathcal{F}^Y; \Psi_Z : Z \rightarrow \mathcal{F}^Z$, by introducing kernels $\mathbf{K}_Y$ and $\mathbf{K}_Z$. The first step in the model is to apply kernel canonical correlation analysis (CCA) [4] to find the directions of high correlation between the two feature spaces. We therefore briefly review the CCA algorithm. The objective in CCA is to find linear transformations $\mathbf{W}_Y$ and $\mathbf{W}_Z$ maximizing the correlation between $\mathbf{W}_Y\mathbf{Y}$ and $\mathbf{W}_Z\mathbf{Z}$. Applied in the kernel feature space of each observation,

$$\begin{aligned} \{\hat{\mathbf{W}}_Y, \hat{\mathbf{W}}_Z\} = \operatorname{argmax}_{\{\mathbf{W}_Y, \mathbf{W}_Z\}} & \operatorname{tr}\left(\mathbf{W}_Y^{\mathrm{T}}\mathbf{K}_Y^{\mathrm{T}}\mathbf{K}_Z\mathbf{W}_Z\right), \\ \text{s.t.} \quad & \operatorname{tr}\left(\mathbf{W}_Y^{\mathrm{T}}\mathbf{K}_Y^{\mathrm{T}}\mathbf{K}_Y\mathbf{W}_Y\right) = \mathbf{I} \\ & \operatorname{tr}\left(\mathbf{W}_Z^{\mathrm{T}}\mathbf{K}_Z^{\mathrm{T}}\mathbf{K}_Z\mathbf{W}_Z\right) = \mathbf{I}, \end{aligned} \tag{4}$$

the optima is found through an eigenvalue problem. In [4] it is suggested to apply CCA in the dominant principal subspace of each feature space instead of directly in the feature space, this constrains $\mathbf{W}_Y$ and $\mathbf{W}_Z$ to explain only the significant variance. We found this suggestion to be important in practice.

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. However, we wish to represent the full variance of each feature space. To achieve this further sets of bases representing the remaining variance are required. We derive a new algorithm, non consolidating component analysis, for finding these additional bases.

2.1 NCCA

Once a set of basis-vectors in each feature space have been found that describe the shared variance, we need to find directions in each feature space that individually represents the remaining variance of each data space. We therefore proceed by seeking the directions of maximum variance in the data that are *orthogonal* to the directions given by the canonical correlates. We call the following procedure *non-consolidating components analysis* (NCCA). The NCCA algorithm is applied in the same space as CCA, but now we seek the first direction $\mathbf{v}_1$ of maximum variance which is orthogonal to the canonical directions that were already extracted,

$$\mathbf{v}_1 = \text{argmax}_{\mathbf{v}_1} \mathbf{v}_1^{\mathrm{T}} \mathbf{K} \mathbf{v}_1 \tag{5}$$

subject to: $\mathbf{v}_1^{\mathrm{T}} \mathbf{v}_1 = 1$ and $\mathbf{v}_1^{\mathrm{T}} \mathbf{W} = \mathbf{0}$, (here we have temporarily dropped the partition subscript), $\mathbf{W}$ are the canonical directions and $\mathbf{K}$ is the covariance matrix in the dominant principal subspace of the feature space. The optimal $\mathbf{v}_1$ is found via an eigenvalue problem,

$$\left(\mathbf{C} - \mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{K}\right) \mathbf{v}_1 = \lambda_1 \mathbf{v}_1. \tag{6}$$

For successive directions further eigenvalue problems of the form

$$\left(\mathbf{K} - \left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sum_{i=1}^{k-1} \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}}\right) \mathbf{K}\right) \mathbf{v}_k = \lambda_k \mathbf{v}_k \tag{7}$$

need to be solved.

After applying CCA and NCCA we have recovered the following embeddings of the data

$$\mathbf{X}^S = \tfrac{1}{2} \left(\mathbf{W}_Y \mathbf{F}_Y + \mathbf{W}_Z \mathbf{F}_Z\right) \tag{8}$$

$$\mathbf{X}^Y = \mathbf{V}_Y \mathbf{F}_Y; \quad \mathbf{X}^Z = \mathbf{V}_Z \mathbf{F}_Z, \tag{9}$$

where $\mathbf{F}_Y$ and $\mathbf{F}_Z$ represent the kernel PCA representation of each observation space. The latent variables $\mathbf{X}^Y$, $\mathbf{X}^Z$ represent the non shared variance of each feature space and $\mathbf{X}^S$ represents the shared variance.

Our methodology results in a purely spectral algorithm: the optimization problems are convex and they lead to unique solutions. However, these spectral methods are perhaps less useful when it comes to inquisition of the resulting model. The pre-image problem means that handling missing data can be rather involved [9]. Probabilistic latent variable models lack the elegant convex solutions provided by spectral methods,

but they facilitate model inquisition. Harmeling [3] has performed a series of embedding experiments for which the ground truth is available. By comparing the embeddings from several different spectral algorithms with the ground truth, a good correpondence between the likelihood of the GP-LVM and the quality of the embedding is shown. Intuitively this is because: if the assumptions in (2) and (3) hold and the manifold has been correctly 'unraveled' (1) should also hold. If (1) holds then the likelihood of the GP-LVM will be high and inferences undertaken with the GP-LVM will be accurate. This allows us to proceed by combining our algorithm with the GP-LVM for model selection and inference.

The NCCA algorithm results in implicit mappings from the observation spaces to the embeddings or, if non function based kernels are used — such as those resulting from the MVU algorithm [14], a mapping can be learned explicitly. However, this leaves us with the pre-image problem [10]. For a given latent location, what is the correct observation? The next stage is, therefore, to build Gaussian process mappings from the latent to the data space. This will result in a combination of GP-LVM models that can be used for any inference tasks in the model. This means that as a post processing step, we learn mappings to regenerate the observations spaces $\mathbf{Y}$ and $\mathbf{Z}$ from the embeddings. We define the $\mathbf{Y}$ and $\mathbf{Z}$ specific latent space as $\mathbf{X}^{S,Y} = [\mathbf{X}^S; \mathbf{X}^Y]$, $\mathbf{X}^{S,Z} = [\mathbf{X}^S; \mathbf{X}^Z]$ respectively. The mappings,

$$f^{\{Y,Z\}} : \mathbf{y}_i = f^Y(\mathbf{x}_i^{S,Y}) + \epsilon_f^Y; \mathbf{z}_i = f^Z(\mathbf{x}_i^{S,Z}) + \epsilon_f^Z, \tag{10}$$

$$g^{\{Y,Z\}} : \mathbf{x}_i^S = g^Y(\mathbf{y}_i) + \epsilon_g^Y = g^Z(\mathbf{z}_i) + \epsilon_g^Z, \tag{11}$$

$$h^{\{Y,Z\}} : \mathbf{x}_i^Y = h^Y(\mathbf{y}_i) + \epsilon_h^Y; \mathbf{x}_i^Z = h^Z(\mathbf{z}_i) + \epsilon_h^Z, \tag{12}$$

where $\epsilon_{\{f,g,h\}}^{\{Y,Z\}}$ are samples from zero mean Gaussian distributions, are learned using GP-regression [8].

Note that we have, in effect, created a set of back-constrained GP-LVMs from our data [6]. We could have used the GP-LVM algorithm directly for learning this model, in practice though, the spectral approach we have described is much quicker and has fewer problems with local minima.

2.2 Inference

The proposed model represents two data sets using a low dimensional latent variable. Once the latent representations have been learned we are interested in inferring the location $\mathbf{z}_*$, corresponding to a previously unseen input $\mathbf{y}_*$. The input and the sought output locations latent representation coincide on the shared latent subspace $\mathbf{X}^S$, which can be determined from the input through the mapping g^Y. Therefore, to determine the full location of the corresponding output, it remains to determine the location over the private space associated with the output. However, the private subspace is orthogonal to the input specific latent subspace. This implies $\mathbf{y}_*$ can provide no further information to disambiguate over this space, *i.e.* each location over the private space corresponds to outputs that are ambiguous to the input location. We therefore proceed by finding the most probable $\mathbf{z}_*$'s generated by f^Z for different locations over $\mathbf{X}^Z$. From our model's perspective, this is equivalent to minimizing the predictive variance of f^Z [8]

with respect to $\mathbf{x}_*^Z$ under the constraint that $\mathbf{x}_*^S$ is given,

$$\hat{\mathbf{x}}_*^Z = \text{argmax}_{\mathbf{x}_*^Z} \left[k(\mathbf{x}_*^{S,Z}, \mathbf{x}_*^{S,Z}) - k(\mathbf{x}_*^{S,Z}, \mathbf{X}^{S,Z})^{\mathrm{T}} (\mathbf{K} + \beta^{-1}\mathbf{I}) k(\mathbf{x}_*^{S,Z}, \mathbf{X}^{S,Z}) \right]. \tag{13}$$

The optimal $\hat{\mathbf{x}}_*^Z$ is found by optimizing Eq. (13) using gradient based methods. We are looking to find all the locations $\mathbf{z}_*$ that are consistent with a specific $\mathbf{y}_*$. The separation of $\mathbf{Z}$ into shared and non-shared means that the ambiguities are very close in the shared subspace. Therefore, we can explore the different modes by looking for nearest neighbors in the shared subspace and initializing the GP-LVM optimizations from those neighbors.

3 Human Pose Estimation

We now consider the application of the model to human pose estimation. We will first briefly review relevant previous work in this area, much of which has provided the inspiration of our approach. Human pose estimation is the task of estimating the full pose configuration of a human from an image. Due to the high dimensionality of the image representation it is common practice, as a preprocessing stage, to represent each image by a lower dimensional image feature vector. In the simplest case, where there is no ambiguity between the image features and the pose, the relationship can be modeled with regression as was demonstrated by [1]. However, regression models are not sufficient to accurately describe the multi-modalities that we expect to arise as a result of ambiguities associated with common image features. An alternative approach to dealing with the multi-modalities is to use a conditional model over the image feature space given the poses [13]. However, due to the high dimensionality and relative data sparsity care must be taken in choosing the class of conditional models. One solution is to incorporate a low dimensional manifold within the conditional density model, thereby avoiding the curse of dimensionality. This approach is followed by [2, 7] who exploit the shared GP-LVM [11] to jointly learn a low dimensional representation of both the image features and the pose space. An advantage of basing the model on the GP-LVM [5] is that it provides a principled probabilistic framework for the resulting inference of pose, easily allowing, for example, the incorporation of dynamical models [2].

A key problem with the application of the shared GP-LVM in this context is that a single latent space is used to explain *all* the variance in the data. Since we know that only a portion of the variance is shared, with the remainder being specific to each data partition, it seems to make much more sense to encode this explicitly. The proposed NCCA model does this by decomposing the latent space into sub-spaces which encode the shared variance and subspaces which encode the variance that is private to each data set. These constraints on the latent spaces lead to much cleaner representation of the ambiguities in practice (as we shall see in Section 3.1). When combining the image features with the motion capture the shared latent space represents the variance in the pose space that can be discriminated from the image feature location. The ambiguities, if they exist, therefore necessarily lie in the portion of the latent space that is specific to the motion capture data. As we shall see this makes them much easier to visualize and interpret.

Further, it is likely that a significant amount of the variance in a descriptor does not help in disambiguating the pose. In the shared GP-LVM this information is still encoded in the model: the shared GP-LVM attempts to model *all* the variance in the data. The NCCA model encodes this information separately, which means it does not influence the inference procedure. This is a key advantage of our model compared to other conditional models, where inference is polluted by estimating this task irrelevant variance.

Once again we direct the reader to Figure 1 to see this effect. The y-axis in the left plot is encoding the spurious information from the image features. It does not help with encoding the true pose. It also is prevented from corrupting the information that arises from the motion capture data (right plot).

3.1 Experiments

We considered a walking sequence from the HumanEva database [12]. There are four cycles in a circular walk, we use two for training and two for testing for the same subject. In the original data the subject is walking in a counter-clockwise direction, to introduce further ambiguities into the data we transform each image and pose to also include the clockwise motion. Each image is represented using a 100 dimensional integral HOG descriptor [15] with 4 orientation bins and the pose space by the $3D$ locations of 19 major body joints. There are two types of motion in the data, the global motion of the subject moving around in $3D$ space and the local body relative motion, *i.e.* each stride. We assume that each local movement in the training data is possible at all global locations. To decorrelate the two motions we represent the pose space as the sum of a MVU kernel [14] applied to the full pose space and a linear kernel applied on the local motion. The NCCA algorithm with this kernel over the pose space and a MVU kernel over the image features results in a one dimensional shared space explaining 9% and 18% of the variance in the image feature and pose space respectively. To retain 95% of the variance in each observation two dimensions are needed to represent the non-shared variance for both the pose and the image feature space. The pose specific latent space takes the shape of a torus, the larger circle is associated with the heading direction and the smaller circles associated with the stride at that position Figure 4. The total computation time for learning the embedding and the required mappings was about 10 minutes on a Intel Core Duo with 1GB of RAM. In Figure 6 the $2nd$ and $3rd$ row show inference of two different image features from the test data is shown. The inference procedure using 20 nearest neighbor initializations per image took a few seconds to compute.

Shared GP-LVM: The inference procedure in the NCCA model consists of a discriminative mapping followed by the optimization over a sub-set of the pose specific latent space. In comparison to the shared GP-LVM [2, 7] the optimization is done over the full latent representation of both image feature and pose. This means that the objective is influenced by how well the latent locations represents variance in the image features that are irrelevant for discriminating the pose. In contrast, the optimization in the NCCA model is done over latent dimensions representing only pose relevant variance. We applied the shared GP-LVM model suggested in [7] to the above data set. To compare models with similar inference complexity we learn a two dimensional shared latent representation of image feature and pose. The optimization on the latent space is

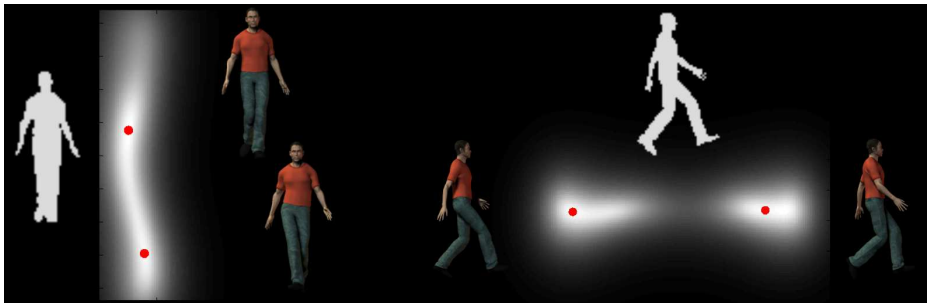

Fig. 3. *Pose inference from silhouette using two different silhouettes from the training data. From the silhouette in the left image it is not possible to determine the positioning of the legs, this results in an elongated region of high probability in the pose private subspace that describs a full stride. The right image shows a silhouette from which it is not possible to differentiate between the right and the left leg. This results in two clear modes over the non-shared dimensions representing the two possible leg labellings in the silhouette.*

initialized by the nearest neighbors in the training data. Note that this is a search in the 100 dimensional image feature space compared to the algorithm we present were the nearest neighbor search takes place in a *one* dimensional space. In Figure 6 the bottom two rows shows the results of applying the Shared GP-LVM to inference the pose of the same images as for the NCCA model.

4 Conclusion

We have presented a practical approach to consolidating two data sets with known correspondences via a latent variable model. We constructed a generative latent variable model for inference and model validation and a spectral algorithm for fast learning of the embeddings, both these interpretations of our model built upon canonical correlation analysis. The resulting model was successful in visualizing the ambiguities on a benchmark human motion data set. Moreover, not only is the presented model fast to train, but also it is efficient in the test phase. Inference is realized by a fast discriminative model that constrains the related generative model. This results in a much simpler estimation compared to previous generative approaches.

References

1. A. Agarwal and B. Triggs. Recovering 3 d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58, 2006.
2. Carl Henrik Ek, Philip H.S. Torr, and Neil 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.

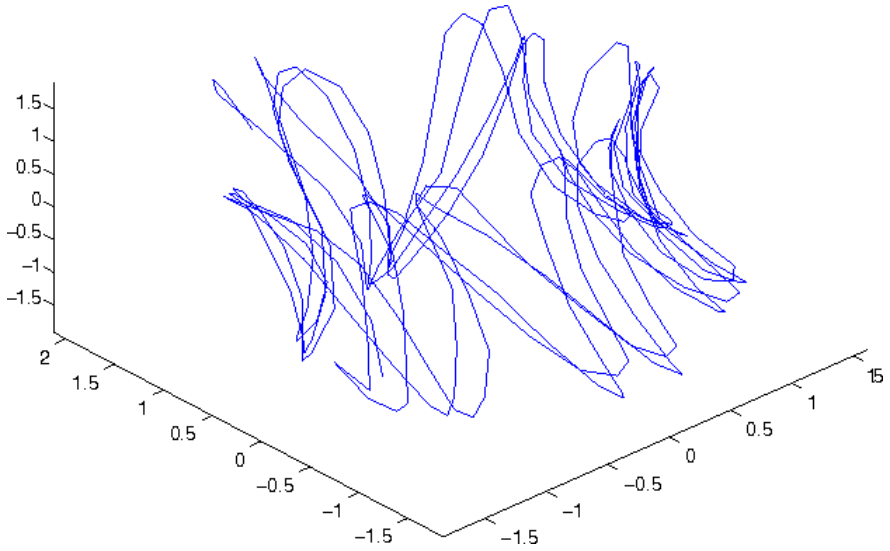

Fig. 4. *The pose specific latent representation associated with the HumanEva data. Applying the NCCA algorithm results in a one dimensional shared subspace and a two dimensional pose private space. The larger circle in the embedding is associated with the heading direction while the smaller circles encodes the configuration of arms and legs.*

3. Stefan Harmeling. Exploring model selection techniques for nonlinear dimensionality reduction. Technical Report EDI-INF-RR-0960, University of Edinburgh, 2007.
4. Malte Kuss and Thore Graepel. The geometry of kernel canonical correlation analysis. Technical Report TR-108, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2003.
5. Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian Process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, 2005.
6. Neil D. Lawrence and Joaquin Qui nonero-Candela. Local distance preservation in the gp-lvm through back constraints. In Russell Greiner and Dale Schuurmans, editors, *ICML '06: Proceedings of the 23rd international conference on Machine learning*, volume 21, pages 513–520, New York, NY, USA, 2006. ACM.
7. Ram Navaratnam, Andrew Fitzgibbon, and Roberto Cipolla. The joint manifold model. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.
8. Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
9. Guido Sanguinetti and Neil D. Lawrence. Missing data in kernel pca. In *ECML*, Berlin, 2006.
10. B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press Cambridge, MA, USA, 2001.
11. A.P. Shon, K. Grochow, A. Hertzmann, and R.P.N. Rao. Learning shared latent structure for image synthesis and robotic imitation. *Proc. NIPS*, pages 1233–1240, 2006.
12. L. Sigal and M.J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 2006.

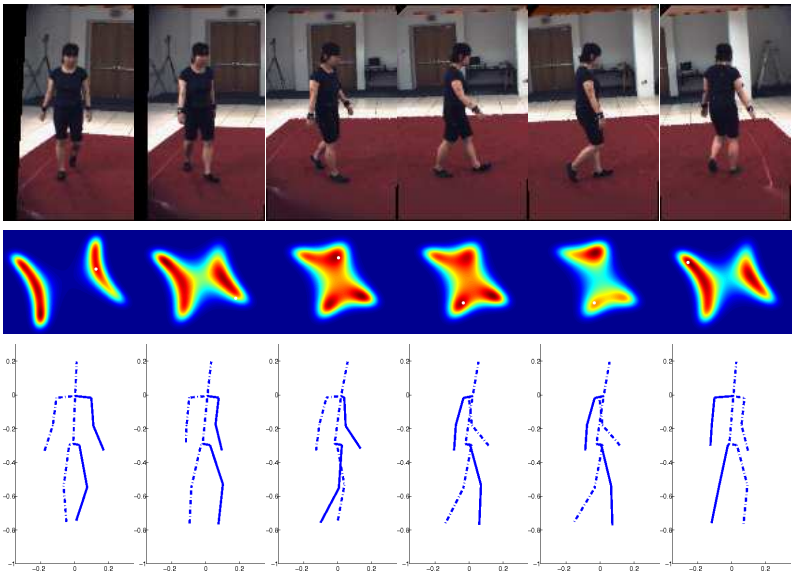

Fig. 5. *Pose inference on a sequence of images from the HumanEva data set. Top row: original test set image. Second row: visualisation of the modes in the non-shared portion of the pose specific latent space. Note how the modes evolve as the subject moves. When the subject is heading in a direction perpendicular to the view-plane, it is not possible to disambiguate the heading direction image (*1*,* 2 *and* 6*) this is indicated by two elongated modes. In image (*$3-5$*) it is not possible to disambiguate the configuration of the arms and legs this gives rise to a set of discrete modes over the latent space each associated with a different configuration. Bottom row: the pose coming from the mode closest to the ground truth is shown. The different types of mode are explored further in Figure 6.*

13. C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. *Proc. Conf. Computer Vision and Pattern Recognition*, pages 217–323, 2005.
14. K.Q. Weinberger, F. Sha, and L.K. Saul. Learning a kernel matrix for nonlinear dimensionality reduction. *ACM International Conference Proceeding Series*, 2004.
15. Q. Zhu, S. Avidan, M.C. Yeh, and K.T. Cheng. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. *CVPR*, 1(2):4, 2006.

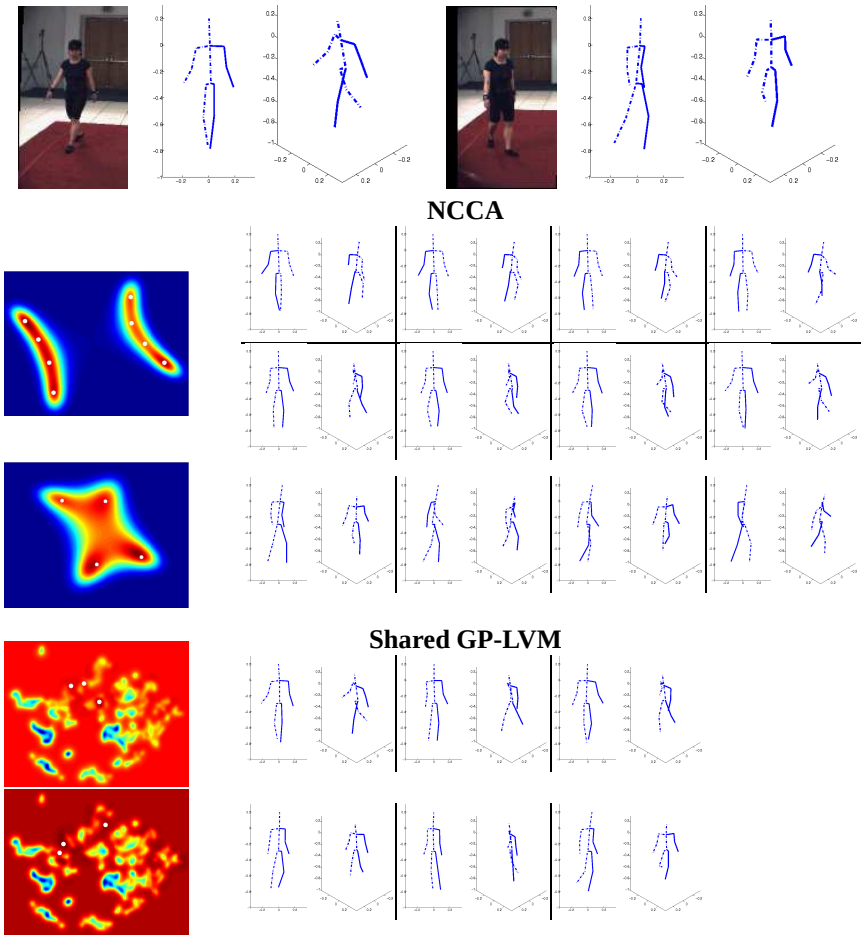

Fig. 6. *The top row shows two images from the training data. The* $2nd$ *and* $3rd$ *row shows results from infering the pose using the NCCA consolidation, the first column shows the likelihood sampled over the pose specific latent space constrained by the image features, the remaining columns shows the modes associated with the locations of the white dots over the pose specific latent space.* **NCCA:** *In the* $2nd$ *row the position of the leg and the heading angle cannot be determined in a robust way from the image features. This is reflected by two elongated modes over the latent space representing the two possible headings. The poses along each mode represents different leg configurations. The top row of the* $2nd$ *column shows the poses generated by sampling along the right mode and the bottom row along the left mode. In the* $3rd$ *row the position of the leg and the heading angle is still ambiguous to the feature, however here the ambiguity is between a discrete set of poses indicated by four clear modes in the likelihood over the pose specific latent space.* **SGP-LVM:** *The* $4th$ *and* $5th$ *row show the results of doing inference using the SGP-LVM model. Even though the most likely modes found are in good correspondece to the ambiguities in the images the latent space is cluttered by local minima that the optimization can get stuck in.*