# Associate Latent Encodings in Learning from Demonstrations

**Hang Yin**[1,2], **Francisco S. Melo**[1], **Aude Billard**[2] **and Ana Paiva**[1]
[1]GAIPS, INESC-ID and Instituto Superior Técnico, Universidade de Lisboa
[2]Learning Algorithms and Systems Laboratory, École Polytechnique Fédérale de Lausanne
{hang.yin, aude.billard}@epfl.ch {fmelo, ana.paiva}@inesc-id.pt

**Abstract**

We contribute a learning from demonstration approach for robots to acquire skills from multi-modal high-dimensional data. Both latent representations and associations of different modalities are proposed to be jointly learned through an adapted variational auto-encoder. The implementation and results are demonstrated in a robotic handwriting scenario, where the visual sensory input and the arm joint writing motion are learned and coupled. We show the latent representations successfully construct a task manifold for the observed sensor modalities. Moreover, the learned associations can be exploited to directly synthesize arm joint handwriting motion from an image input in an end-to-end manner. The advantages of learning associative latent encodings are further highlighted with the examples of inferring upon incomplete input images. A comparison with alternative methods demonstrates the superiority of the present approach in these challenging tasks.

Learning from demonstrations (LfD) is promising for an effective transfer of robotic skills from humans to robots. Traditional LfD approaches often learn with hand-crafted features (e.g., the poses of labeled objects) in a low dimensional space (e.g., robot operational space) (Khansari, Kronander, and Billard 2014)(Calinon 2015). This is limited for broader LfD applications in a more practical context, where high dimensional raw sensory data is pervasive. The desideratum of learning from high dimensional demonstrations solicits LfD to automate the extraction of task relevant features alongside learning the underlying task constraints. Such a representation learning has the potential to free human users, from the domain feature design as well as the restriction of sensor selection, thus substantially improving the applicability and the empirical value of the LfD framework.

Another often-overlooked aspect in LfD is learning from multi-modal demonstrations. Previous LfD works tend to focus on modeling data of a single sensor modality, even though there is nothing preventing the task demonstrations being observed through the lens of various robot sensors. On the other hand, human beings are quite proficient in learning and fusing the knowledge or experience gathered from different sensing systems. For instance, humans can perceive the shape of an object through both vision and tactile sensations. Such a redundant and associative description is beneficial when only partial information is presented to the agents in the task reproduction: humans can still effortlessly infer the shape of an object with a hand exploration in darkness. Therefore, by learning from multi-modal demonstrations, the robots are enabled to gain a more complete task description, a natural mechanism to estimate what is unknown from what is known, and as such, a capacity of robustly executing the task in face of uncertainty.



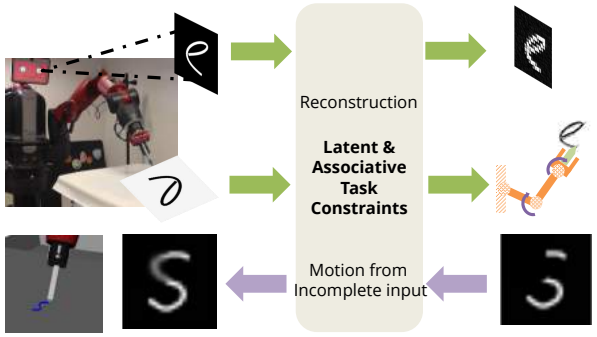



Figure 1: Learning representations for multiple sensory perceptions (vision and motion) and associating them in the latent space for the underlying task. The desired sensory/motor state, e.g., joint motion command, can be efficiently derived from incomplete or novel input e.g., symbol images.

In this paper, motivated by the above challenges, we present an approach that allows robots to learn demonstrated skills from high dimensional and multi-modal demonstration data. Among different LfD variants, we consider the problem of inverse optimal control (IOC) which extracts and represents task constraints as cost functions. We propose to learn the task cost function as well as its features by integrating a representation learning framework into LfD. Moreover, constrains in the latent space are imposed and jointly learned, by exploiting the fact that the multi-modal data is a redundant description for the underlying task to model. To this end, we can obtain succinct task manifolds and representations, which can be leveraged for efficient motion derivation from the raw sensory input (Figure 1). The main

contributions of this paper are:

- An approach which enables an agent to learn from high dimensional raw demonstration data, with an adaptation from the unsupervised representation learning.
- A KL-divergence-based metric that compactly associates the statistics of latent encodings of different demonstration modalities, resulting in an efficient stochastic gradient descent training algorithm.
- An end-to-end system that enables the robot to derive arm joint writing motion from observed symbol images, with robustness against image occlusion.

## Related Work

This section gives a brief discussion about related literature in topics of learning from demonstrations and representation learning.

*Learning from Demonstrations* - in the research of implicit learning from demonstrations, various features have been investigated to parameterize the target cost functions. A quadratic form was first researched as the inverse problem of linear quadratic optimal control in Kalman's pioneering work (Kalman 1964). In (Rückert et al. 2013) and (Yin, Paiva, and Billard 2014), structured quadratic forms were learned to fulfill prior constraints of the task representation. As a more universal cost encoding, radial basis function (RBF) was vastly adopted as a popular feature choice (Ziebart et al. 2008). This parameterization often results in an intractable demonstration likelihood evaluation, for which discretization (Dvijotham and Todorov 2010), Laplacian approximation (Levine and Koltun 2012) and trajectory sampling (Kalakrishnan et al. 2013) were resorted as approximations. Moreover, research effort has been made to avoid an explicit feature design. For example, Gaussian Process was used to obtain a nonparametric cost representation (Levine, Popovic, and Koltun 2011). Recently, this line of research was advanced by learning features of discrete state (Wulfmeier, Ondruska, and Posner 2015) and continuous robot joint state (Finn, Levine, and Abbeel 2016). Our work learns cost features on high dimensional vision data, as was envisioned in (Wulfmeier, Ondruska, and Posner 2015). Also, the presented variational formulation leads to an optimization of the lower bound of the original demonstration likelihood, which can be efficiently evaluated with sparse samples.

*Representation Learning* - in the task of learning from unlabeled data, representation learning has achieved remarkable progresses with the contributions of variational Bayes likelihood (Kingma and Welling 2014) and adversarial training objectives (Goodfellow et al. 2014). These advancements caught the attention of roboticists on embedding or building controllers on high dimensional sensory input such as vision data. Notable results have been obtained in different domains including robotic visuomotor skills (Levine et al. 2015), human gait modeling (Chen et al. 2015) and pixels-based reinforcement learning tasks (Watter et al. 2015). Our work differs from (Levine et al. 2015) in that we consider the task of learning robotic skills from humans. Also our work proposes to learn a generative model, which is arguably more flexible than searching a discriminant policy in dealing with incomplete sensory information. The work of (Chen et al. 2015) suggested enforcing dynamic movement primitives in the latent space extracted by a denoising auto-encoder, while (Watter et al. 2015) chose to impose locally linear constraints for the latent dynamics. These are similar to our work in terms of feature embedding and latent association. However, our work emphasizes the association among multiple general sensor readings from the perspective of imitation learning. With such an association enforced, the task can be learned by linking perceptual inputs and encoded motor commands in a direct manner.

## Background

This section presents necessary background to develop our contributed approach. It starts with the introduction of the target scenario and the relevant notations. Then this part is followed by a very brief overview about variational autoencoder (VAE) (Kingma and Welling 2014), which is adapted in the subsequent method development.

### Preliminaries

We consider an imitation learning problem where a robot needs to learn from expert demonstrations collected from multi-modal sensors. In particular, learning handwriting is taken as a running example throughout this paper. The demonstration with multiple modalities implies various types of sensory information (e.g., arm motion and visually observed letter images) are collected from a same underlying task procedure (writing the target letter). The demonstrations are denoted as $\{\boldsymbol{x}_i\}$. Without the loss of generality, the $i$-th joint observation $\boldsymbol{x}_i$ is comprised of two modalities as $\boldsymbol{x}_i = \{\boldsymbol{x}_v^i, \boldsymbol{x}_m^i\}$, with subscripts indicating vision (pixels of letter images) and motion (arm joint trajectory or its proper parametric representation) respectively. The goal of imitation is to rationalize the expert behaviors, by learning a model, e.g., a probabilistic one with maximum entropy in the exponential family (Ziebart et al. 2008):

$$p(\boldsymbol{x}) = \frac{\exp(-\mathcal{J}(\boldsymbol{x}, \boldsymbol{\theta}))}{\int \exp(-\mathcal{J}(\boldsymbol{x}', \boldsymbol{\theta}))d\boldsymbol{x}'} \tag{1}$$

In this equation, as a statistic momentum, $\mathcal{J}(\boldsymbol{x}, \boldsymbol{\theta})$ assigns a real value to ensure the observation of the interested modality $\boldsymbol{x} = \{\boldsymbol{x}_v^i\}$ or $\boldsymbol{x} = \{\boldsymbol{x}_m^i\}$ is more likely under the distribution with a proper model parameter $\boldsymbol{\theta}$. The learning of $\boldsymbol{\theta}$ can be exercised by maximizing the likelihood or its appropriate surrogate.

### Variational Autoencoder

Recently, the variational auto-encoder (VAE) (Kingma and Welling 2014) emerged as a popular unsupervised learning framework in representation learning. VAE assumes latent variables to model complicated correlations between high dimensional features, such as camera pixels and arm joint trajectories in our case. Concretely, taking the example of $\boldsymbol{x}_v$, the image pixels are assumed to be generated from a latent distribution $p_0(\boldsymbol{z}_v)$. By taking the formula of total prob-

ability, (1) is re-written as:

$$p(\boldsymbol{x}_v^i) = \int p_{\boldsymbol{\theta}_v}(\boldsymbol{x}_v^i|\boldsymbol{z}_v)p_0(\boldsymbol{z}_v)d\boldsymbol{z}_v \tag{2}$$

where $\boldsymbol{z}_v$ denotes the latent representation and its prior is $p_0$. $p_{\boldsymbol{\theta}_v}(\boldsymbol{x}_v|\boldsymbol{z}_v)$ is the process to obtain $\boldsymbol{x}_v^i$ by decoding a sample $\boldsymbol{z}_v$.

The above objective (2) is, however, not easy to evaluate. In fact, a large amount of samples might be required to obtain a sufficiently accurate estimation, because most $\boldsymbol{z}_v$ will not generate the observed $\boldsymbol{x}_v^i$ and the contribution of $p_{\boldsymbol{\theta}_v}(\boldsymbol{x}_v^i|\boldsymbol{z}_v)$ would be extremely small. VAE proposes to use a $\boldsymbol{\phi}$ parametrized proposal distribution $q_{\boldsymbol{\phi}_v}(\boldsymbol{z}_v|\boldsymbol{x}_v^i)$ to approximate $p(\boldsymbol{z}_v|\boldsymbol{x}_v^i)$. The approximation is regulated through a Kullback-Leibler (KL) divergence:

$$\mathrm{KL}[q_{\boldsymbol{\phi}_v}\|p(\boldsymbol{z}_v|\boldsymbol{x}_v)] = \mathbb{E}_{q_{\phi_v}}[\log q_{\boldsymbol{\phi}_v} - \log p(\boldsymbol{z}_v|\boldsymbol{x}_v^i)] \tag{3}$$

Applying Bayes rule and noticing that total probability $p(\boldsymbol{x}_v^i)$ is independent of $\boldsymbol{z}_v$, the above equation can be expanded and rearranged as:

$$\begin{aligned}\mathcal{L}_v(\boldsymbol{\theta}_v, \boldsymbol{\phi}_v, \boldsymbol{x}_v^i) =& \mathrm{KL}[q_{\boldsymbol{\phi}_v}\|p(\boldsymbol{z}_v|\boldsymbol{x}_v^i)] - \log p(\boldsymbol{x}_v^i) \\ =& \mathrm{KL}[q_{\boldsymbol{\phi}_v}(\boldsymbol{z}_v|\boldsymbol{x}_v^i)\|p_0(\boldsymbol{z}_v)] \\ &- \mathbb{E}_{q_{\boldsymbol{\phi}_v}(\boldsymbol{z}_v|\boldsymbol{x}_v^i)}[\log p_{\boldsymbol{\theta}_v}(\boldsymbol{x}_v^i|\boldsymbol{z}_v)]\end{aligned} \tag{4}$$

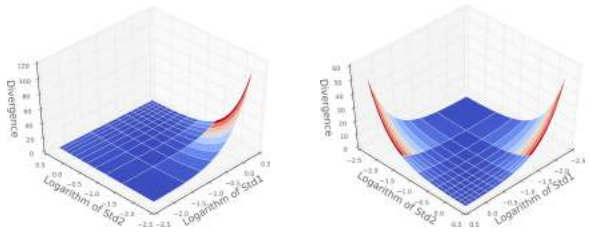
Because of the non-negativity of KL-divergence, the right hand side can be viewed as an upper bound of the negative logarithm of (2). Hence $\mathcal{L}_v$ can be used as a valid surrogate to optimize the original objective when (3) is small. In practice, $q_{\boldsymbol{\phi}_v}(\boldsymbol{z}_v|\boldsymbol{x}_v^i)$ is often parameterized as a Gaussian $\boldsymbol{z}_v \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}_v^i, \boldsymbol{\theta}_v), \boldsymbol{\Sigma}(\boldsymbol{x}_v^i, \boldsymbol{\theta}_v))$, which is sufficiently rich to provide a close approximation to $p(\boldsymbol{z}_v|\boldsymbol{x}_v^i)$. The mean and covariance are modeled as deep neural networks parameterized by $\boldsymbol{\phi}_v$. The generative process $p_{\boldsymbol{\theta}}(\boldsymbol{x}_v^i|\boldsymbol{z}_v)$ also uses a deep neural network to model the statistics of an exponential distribution, which could be a binomial distribution for pixel values or a multivariate Gaussian distribution if a real-valued output is expected. As per the prior of the latent variable, an isotropic Gaussian $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is often used. This simplifies the evaluation of the regularization term $\mathrm{KL}[q_{\boldsymbol{\phi}_v}(\boldsymbol{z}|\boldsymbol{x}^i)\|p_0(\boldsymbol{z})]$ to a closed-form. Finally, the integral term can be estimated by taking only a few samples from $q_{\boldsymbol{\phi}_v}$, and as is shown in (Kingma and Welling 2014), even a single sample is sufficient for stable iterations.

## Learning and Associating Latent Encodings

Variational auto-encoder jointly learns the encoding and decoding features to obtain a compact representation of the original demonstrations. This section presents our main contribution, an associative variational auto-encoder, which adapts the original framework to couple different demonstration modalities in the latent space. We show that the extracted representations are flexible to conduct efficient inference in the context of motion synthesis.

### Associating Latent Representations

An associative variational auto-encoder consists of multiple variational auto-encoders, each of which models one modality of the demonstration. The factored probabilistic model




(a) Standard KL-divergence (b) Symmetrical KL-divergence

Figure 2: Standard and symmetrical KL-divergences between $\mathcal{N}(0, \sigma_1^2)$ and $\mathcal{N}(0, \sigma_2^2)$. The standard KL-divergence fails to capture the discrepancy for certain cases, e.g., $\sigma_1 = e^{-2}$ and $\sigma_2 = 1$, while the symmetrical one is invariant w.r.t. the commutation.

is correlated if we consider that each modality corresponds to a different perspective on the underlying task. This implies their latent encodings can be correlated by a metric, in the general form $h(\boldsymbol{z}_v, \boldsymbol{z}_m) = 0$. While there exist numerous assumptions to capture this relation, it is reasonable to adopt an identity constraint. Indeed, the model flexibility is not much compromised by such an assumption, thanks to the expressiveness of the encoding and decoding features. Concretely, the association can be expressed by matching the distributions of the probabilistic latent encodings, namely $q_{\boldsymbol{\phi}_v(\boldsymbol{z}|\boldsymbol{x}_v^i)} = q_{\boldsymbol{\phi}_m(\boldsymbol{z}|\boldsymbol{x}_m^i)}, \forall \boldsymbol{z}$. We propose to quantify this relation with a symmetrical composition of KL-divergences:

$$\begin{aligned}\mathcal{L}_{assoc} =& \mathrm{KL}(q_{\boldsymbol{\phi}_v}(\boldsymbol{z}_v|\boldsymbol{x}_v^i)\|q_{\boldsymbol{\phi}_m}(\boldsymbol{z}_m|\boldsymbol{x}_m^i)) \\ &+\mathrm{KL}(q_{\boldsymbol{\phi}_m}(\boldsymbol{z}_m|\boldsymbol{x}_m^i)\|q_{\boldsymbol{\phi}_v}(\boldsymbol{z}_v|\boldsymbol{x}_v^i)) \\ =& \frac{1}{2}[\log\frac{|\Sigma_m(\boldsymbol{x}_m^i)|}{|\Sigma_v(\boldsymbol{x}_v^i)|} + \log\frac{|\Sigma_v(\boldsymbol{x}_v^i)|}{|\Sigma_m(\boldsymbol{x}_m^i)|} \\ &+(\mu_m(\boldsymbol{x}_m^i) - \mu_v(\boldsymbol{x}_v^i))\Sigma_m^{-1}(\boldsymbol{x}_m^i)(\mu_m(\boldsymbol{x}_m^i) - \mu_v(\boldsymbol{x}_v^i)) \\ &+(\mu_v(\boldsymbol{x}_v^i) - \mu_m(\boldsymbol{x}_m^i))\Sigma_v^{-1}(\boldsymbol{x}_v^i)(\mu_v(\boldsymbol{x}_v^i) - \mu_m(\boldsymbol{x}_m^i)) \\ &+\mathrm{tr}(\Sigma_m^{-1}(\boldsymbol{x}_m^i)\Sigma_v(\boldsymbol{x}_v^i)) + \mathrm{tr}(\Sigma_v^{-1}(\boldsymbol{x}_v^i)\Sigma_m(\boldsymbol{x}_m^i))]\end{aligned} \tag{5}$$

which is still of a closed-form and differentiable with respect to the model parameters $\boldsymbol{\phi}_v$ and $\boldsymbol{\phi}_m$. Under such a constraint penalty, the sequence of modalities is exchangeable, as is shown in Figure 2(a) and 2(b). In our experience, this symmetrical formation generally yields better results comparing with a standard KL-divergence.

Applying (4) to $\boldsymbol{x}_v$ and $\boldsymbol{x}_m$, we obtain a joint objective as $\mathcal{L}(\boldsymbol{\theta}_v, \boldsymbol{\theta}_m, \boldsymbol{\phi}_v, \boldsymbol{\phi}_m, \boldsymbol{x}_v^i, \boldsymbol{x}_m^i) = \mathcal{L}_v + \mathcal{L}_m + \lambda\mathcal{L}_{assoc}$, with $\lambda$ denoting the weight of the imposed constraint. To this end, the introduced loss term of association adds no extra computational complexity to the regular variational auto-encoder training, for which stochastic gradient descent still applies. The overall model architecture is wrapped in Figure 3. Learning with such a model can be understood as extracting low dimensional task manifolds that are, in an ideal condition, fully overlapped . The projections of different observation modalities are co-located on the manifolds. Exploit-

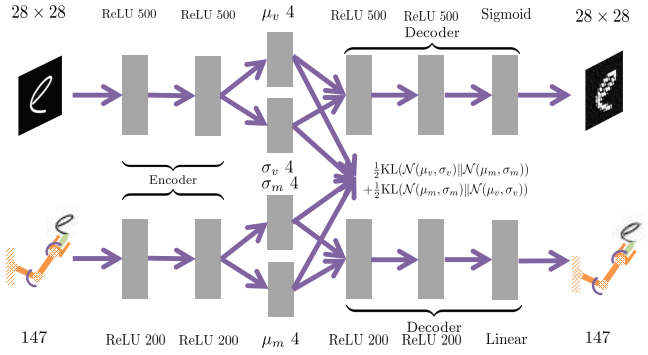



Figure 3: Model architecture of learning latent representations and association on different modalities of demonstrations. Latent layers of representation is annotated with feature type (Rectified Linear Unit) and size. The association is captured by a symmetrical KL-divergence.

ing this fact, we can infer one modality given the other one, such as deriving arm joint motion from a target letter image:

$$p(\boldsymbol{x}_m|\boldsymbol{x}_v) = \int p(\boldsymbol{x_m}|\boldsymbol{z})q_{\boldsymbol{\phi}_v}(\boldsymbol{z}|\boldsymbol{x}_v)d\boldsymbol{z} \tag{6}$$

Such an evaluation can be performed in an efficient way by sampling from the low-dimensional manifold $\boldsymbol{z}$.

Furthermore, the learning of each modality can be understood from the perspective of inverse optimal control, which learns the cost function $\mathcal{J}$ in (1). To see this, we take the negative logarithm of (1) to retrieve $\mathcal{J}$ with a constant term. From (4), when encoder $q_{\boldsymbol{\phi}_v}$ and decoder $p_{\boldsymbol{\theta}_v}$ model multivariate Gaussians with fixed covariances such as $\mathcal{N}(\boldsymbol{\mu}^e(\boldsymbol{x}_v), \sigma_1^2\boldsymbol{I})$ and $\mathcal{N}(\boldsymbol{\mu}^d(\boldsymbol{z}_v), \sigma_2^2\boldsymbol{I})$, we have:

$$\begin{aligned}\mathcal{J}(\boldsymbol{x}_v) &= -\log p(\boldsymbol{x}_v) + C \\ &= \frac{1}{2\sigma_1^2}\|\boldsymbol{\mu^e(x_v)}\|_2 + \frac{1}{2\sigma_2^2}\mathbb{E}_{q_{\boldsymbol{\phi}_v}}[\|\boldsymbol{x}_v - \boldsymbol{\mu}^d(\boldsymbol{z}_v)\|_2] \\ &+ C' - \mathrm{KL}[q_{\boldsymbol{\phi}_v}\|p(\boldsymbol{z}_v|\boldsymbol{x}_v)]\end{aligned} \tag{7}$$

Thus, up to a constant term, the MaxEnt cost $\mathcal{J}$ is upper bounded by the first two terms which are a featured quadratic cost and a reconstruction loss measuring how well the original data $\boldsymbol{x}_v$ is compressed by the feature. Ideally we expect the learned encoder and decoder to yield small KL divergence and reconstruction loss so only the first quadratic term dominates. To this end, we obtain an estimation of the cost $\mathcal{J}$ with a simple quadratic metric in a nonlinear feature space, which is shared by the two associative modalities.

### Efficient Inference on Incomplete Input

The learned full probability provides additional inference options alongside correlating modalities. The low dimensional latent encodings can be leveraged to evaluate the marginal probability thus making the inference within the space of each modality tractable. This can be applied to an even more challenging scenario: while the input features are incomplete or corrupted in comparison with the training demonstrations, the robot is still expected to derive the desired motion in a robust manner.

We address the above challenge by first recovering the complete input feature and then inferring upon the target modality. Concretely, the incomplete input feature, e.g., a letter image $\tilde{\boldsymbol{x}}_v$ with some parts occluded, is projected into the feature space to obtain its latent encoding. With the latent encoding as an initial guess, the manifold can be explored to search a most likely latent point whose reconstructed feature matches the observable part of $\tilde{\boldsymbol{x}}_v$. Quantitatively, it is proposed to solve:

$$\boldsymbol{z}_v^* = \underset{\boldsymbol{z}_v}{\operatorname{argmin}} - \log p_0(\boldsymbol{z}_v) + \eta\|\boldsymbol{x}_v^{(obs)}(\boldsymbol{z}_v) - \tilde{\boldsymbol{x}}_v^{(obs)}\| \tag{8}$$

where $\eta$ weights the difference between the observable parts of the reconstructed and the target images. This objective literally seeks an image that, on one hand matches the observable part of the target one, and on the other hand, is more probable w.r.t the learned letter image knowledge. The inference can be performed with the cross entropy method (de Boer et al. 2005) . The cross entropy method optimizes the target objective by alternating between taking samples from a proposal distribution, e.g., a multivariate Gaussian distribution, and estimating it with the samples weighted under the target objective. Since the samples are taken in a low dimensional space, this method can secure an efficient exploration on the manifold.

## Implementation and Experiment

This section presents the implementation and application of the proposed method in the task of associating handwriting arm motion and the letter image. Details about the experiment setup are given and the presented approach is also compared with other alternatives.

### Data Augmentation

The dataset used in the experiment is UJI Char Pen 2 dataset, from which, for simplicity, only one-stroke-formed alphabetical letters and digits are involved. The images are generated from the 2D online handwriting motion to obtain $28 \times 28$ grayscale thumbnails, resulting a $\boldsymbol{x}_v$ of a length of 784. Iterative LQR (Todorov and Li 2005) is used to derive the optimal joint motion of a 7-DOFs Baxter robot arm. The derived arm joint motion is expected to fit the letter trajectory in the Cartesian space with joint torque efforts minimized. Then joint trajectories are further parameterized with a function approximator for each DOF, yielding a 147-dimension vector $\boldsymbol{x}_m$.

Unfortunately, the original dataset is sparse and unbalanced thus the model tends to fail in learning rare samples. We propose to address this by augmenting the dataset. Specifically, this is done by learning each character with a probabilistic model and re-sampling with perturbations constrained in a kinematics feature space, see (Yin et al. 2016) for details of the method. Eventually, above 70000 pairs of

images and arm motion are obtained, with about 1000 samples per each character.

## Model Implementation

Similar to the standard variational auto-encoder, neural network (NN) models are used as the data encoder $q(\boldsymbol{z}|\boldsymbol{x})$ and decoder $p(\boldsymbol{x}|\boldsymbol{z})$. Each of the NNs is comprised of two layers of rectified linear units (ReLU) as the nonlinear hidden features. Sigmoid functions are adopted as the output features of the vision modality, in order to obtain valid gray-scale values. The entire model is trained through the stochastic gradient descent with an adaptive moment estimation (ADAM) (Kingma and Ba 2015), a learning rate of $1e^{-4}$ and a batch size of 64. The other hyper parameters, including the length of the latent variable and the weight of association term, are selected according to the cross-validation of the reconstruction performance. The code implementation and trained models are publicly accessible[1].

To illustrate the strength of feature learning, Gaussian Mixture Models (GMM) on raw data are also trained as baselines. Training these models with full covariance matrices suffers from severe overfitting issues due to the high dimensionality of the data. To alleviate it, we also use other variants. These encompass a GMM model with diagonal covariance matrices, a GMM model with a PCA dimension reduction and the combination of these two. For the PCA preprocessing, the number of eigenvectors is selected to explain $99\%$ data variance, yielding a reduced dimension of 240 for the image modality and 37 for the motion modality. The number of mixture components is determined based on the BIC criterion. In our experiment, GMMs with 350 components and diagonal covariance matrices give the best BIC score. Since a diagonal matrix cannot capture the correlation across feature dimensions, the best full covariance models with 10 components are also included in subsequent comparisons.

## Deriving Motion from Image

A natural application of the learned encodings and association is to infer one data modality from the other one. In our handwriting context, this implies the model can be used to immediately derive the handwriting motion when a symbol image is presented.

Figure 4 depicts concrete samples of deriving writing motion from symbol images. It is worth noting that the images here are not from the dataset itself but generated in real-time. By real-time generation, we mean the symbols are drawn by hand on a canvas or a user interface. The images are then retrieved and fed to the model to obtain the writing motion in real-time. For the convenience of visualization, all of the joint motion is transformed into the Cartesian space and rendered as 3D plots.

As is clear from the figure, the proposed approach generates the most plausible arm joint motion for the drawn image samples. Because of the rich mode patterns of data, a model learned in the original feature space requires a large amount of local models to fully cover the data modes. Henceforth, among the alternative methods, GMM with diagonal covariance matrices, which allows for a larger number of components, appears to have a comparatively better performance. However, due to the high dimensionality, such a shallow model still fails at times. Additionally, the PCA, aiming to reduce the data dimension, is not helpful in this task. In fact, the methods with PCA preprocessing perform worse than the GMMs learned in the original feature space. This can be partially explained by the fact that the PCA inherently learns linear correlations as the features, which are not expressive in general cases. In our experiment, we observe that sometimes the generated movement forms an incomplete loop, like the cases of "g" and "8" in Figure 4. A possible cause is that, in the data augmentation, the samples are perturbed without an explicit constraint of maintaining the closeness of a loop thus the samples with a loop cut dominate the training data. We expect to obtain improved performance when the data of a better quality is used.

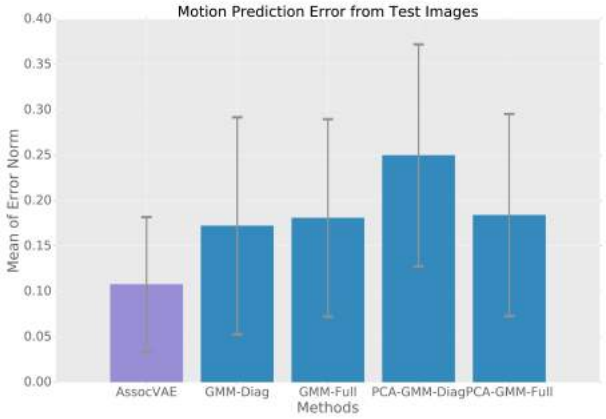



Figure 5: Error comparison of different models on predicting the arm joint motion from a symbol image of the test dataset.

The qualitative visual results are also in accordance with the numerical result. In this experiment, motion trajectories are predicted for the test dataset and the Euclidean distance between the prediction and ground-truth is measured in the function approximator basis space. As is clear from Figure 5, the presented associative VAE outperforms the competing methods by a significant margin. These results demonstrate the advantage of the proposed nonlinear feature learning in such a challenging task that involves high dimensional raw sensory input.

## Handling Incomplete Images

In this experiment, the letters are again written by a person whose handwriting is not included in the dataset. However, the model only receives a corrupted symbol image, with a random quartile covered. We use the cross-entropy method to optimize the objective proposed in (8). In order to guarantee the real-time performance, the number of iterations and samples are both limited to 20. Figure 6 presents some examples of the experiment and clearly illustrates how the pro-

[1]https://github.com/navigator8972/vae_assoc

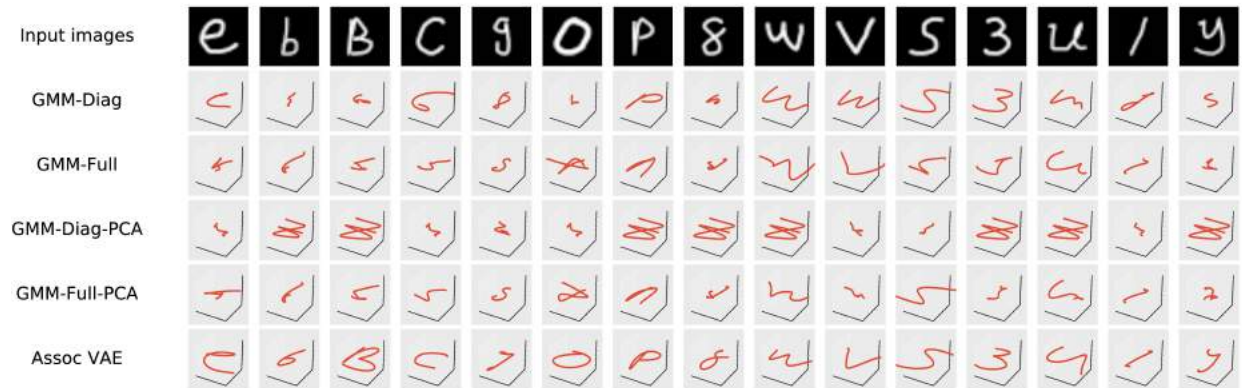



Figure 4: Deriving handwriting motion with different models and symbol images outside the test dataset: the resulted trajectories are transformed to the Cartesian space and shown in 3D plots.

posed inference proceeds. Initially, the algorithm attempts to make up the missed pixels with a plausible component. Then the recovered part is progressively refined and sharpened as the iteration continues. At last, the resultant latent encoding appears to be a good representation of the full underlying image, leading to correct writing motion (the last column). In practice, we observe that 20 iterations are often more than enough to reconstruct the image, thanks to the efficiency from the learned latent representation. With a projection from the observed pixels, the obtained initial guess is expected to be close to the ideal reconstruction on the manifold. In addition, the learned low dimension parameter space only desires a limited number of samples to secure a stable exploration.

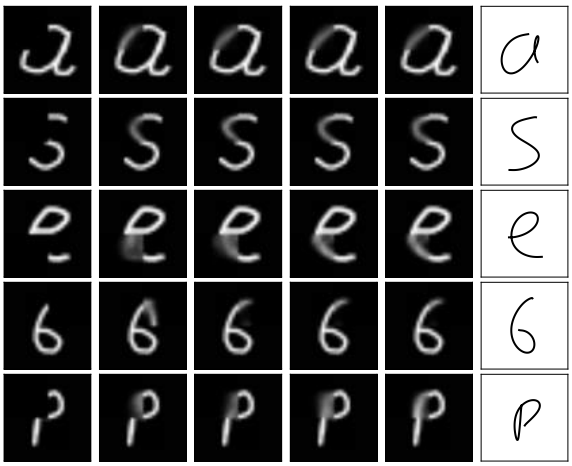


Figure 6: Inferring arm joint motion given occluded letter images: the latent encodings are explored to search complete images to match the observed parts before deriving the associated handwriting motion. The first column: input images; the second to the fifth columns: evolution of the recovered full images in iteration steps of 3, 8, 13, 18; the last column: Cartesian letter trajectories resulted from the inferred arm joint motion.

The GMM-based models are not compared here as it could be notoriously expensive to apply the cross-entropy method to sample pixels of hundreds of dimensions in the original space. We also emphasize that this experiment showcases an unique benefit of learning a generative model of demonstrations. Indeed, it provides a principled way to handle sensor uncertainties in the task execution. The robot systems can benefit from this in terms of skill generalization and robustness. Approaches in which sensory states are mapped directly to actions are unable to achieve this.

## Conclusion

In this paper, we propose an approach that enables robot agents to extract complex task representations and constraints from high dimensional demonstrations. To realize this, a variational autoencoder is adapted with an imposition of a symmetrical KL-divergence metric in the latent space. The resulting objective remains differentiable and therefore the demonstration features and association can be jointly learned with existing standard techniques. The effectiveness of feature extraction and associating control to perception is demonstrated in a series of handwriting related tasks, which include inferring writing motion from symbol images that are high dimensional or even incomplete.

Possible extensions include the introduction of dynamical latent representations. This requires to learn the temporal association in the latent space besides the relations among sensory modalities. The work (Watter et al. 2015) presents a way to realize this in a control task with a single modal sensory feedback. Also, it will be interesting to explore special representation architectures. Along this line of research, we can expect to reuse fruitful results from relevant domains to cope with specific sensor modalities, e.g., using convolutional filters to construct features for locally correlated demonstration data such as videos and natural language.

## Acknowledgments

This work is partially funded by Swiss National Center of Robotics Research and national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/

CEC/50021/2013 and doctoral grant (SFRH/BD/51933/2012) under IST-EPFL Joint Doctoral Initiative.

## References

Calinon, S. 2015. Robot learning with task-parameterized generative models. In *Proceedings of the International Symposium of Robotics Research (ISRR)*.

Chen, N.; Bayer, J.; Urban, S.; and van der Smagt, P. 2015. Efficient movement representation by embedding dynamic movement primitives in deep autoencoders. In *Proceedings of IEEE International Conference on Humanoid Robots (Humanoids)*, 434–440.

de Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research* 134(1):19–67.

Dvijotham, K., and Todorov, E. 2010. Inverse optimal control with linearly-solvable mdps. In *Proceedings of the International Conference on Machine Learning (ICML)*, 335–342.

Finn, C.; Levine, S.; and Abbeel, P. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. *Proceedings of the International Conference on Machine Learning (ICML)* abs/1603.00448.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Proceedings of Neural Information Processing Systems (NIPS)*. Curran Associates, Inc. 2672–2680.

Kalakrishnan, M.; Pastor, P.; Righetti, L.; and Schaal, S. 2013. Learning objective functions for manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 1331–1336.

Kalman, R. E. 1964. When is a linear control system optimal. *Journal of Basic Engineering* 51–60.

Khansari, M.; Kronander, K.; and Billard, A. 2014. Modeling robot discrete movements with state-varying stiffness and damping: A framework for integrated motion generation and impedance control. In *Proceedings of Robotics: Science and Systems (RSS)*.

Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, volume abs/1412.6980.

Kingma, D. P., and Welling, M. 2014. Stochastic gradient vb and the variational auto-encoder. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Levine, S., and Koltun, V. 2012. Continuous inverse optimal control with locally optimal examples. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Levine, S.; Finn, C.; Darrell, T.; and Abbeel, P. 2015. End-to-end training of deep visuomotor policies. *CoRR* abs/1504.00702.

Levine, S.; Popovic, Z.; and Koltun, V. 2011. Nonlinear inverse reinforcement learning with gaussian processes. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F.; and Weinberger, K. Q., eds., *Proceedings of Neural Information Processing Systems (NIPS)*. Curran Associates, Inc. 19–27.

Rückert, E. A.; Neumann, G.; Toussaint, M.; and Maass, W. 2013. Learned graphical models for probabilistic planning provide a new class of movement primitives. *Frontiers in Computational Neuroscience* 6(97).

Todorov, E., and Li, W. 2005. A generalized iterative lqg method for locally-optimal feedback control of constrained nonlinear stochastic systems. In *Proceedings of the 2005, American Control Conference, 2005.*, 300–306 vol. 1.

Watter, M.; Springenberg, J.; Boedecker, J.; and Riedmiller, M. 2015. Embed to control: A locally linear latent dynamics model for control from raw images. In *Proceedings of Neural Information Processing Systems (NIPS)*. 2728–2736.

Wulfmeier, M.; Ondruska, P.; and Posner, I. 2015. Maximum entropy deep inverse reinforcement learning. *CoRR* abs/1507.04888.

Yin, H.; Alves-Olivera, P.; Melo, F. S.; Billard, A.; and Paiva, A. 2016. Synthesizing robotic handwriting motion by learning from human demonstrations. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*.

Yin, H.; Paiva, A.; and Billard, A. 2014. Learning cost function and trajectory for robotic writing motion. In *Proceedings of IEEE International Conference on Humanoid Robots (Humanoids)*.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 1433–1438.