Preprint version; final version available at https://ieeexplore.ieee.org
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) (2022), pp: 9814-9820
Published by: IEEE

Disentangled Sequence Clustering for Human Intention Inference

Mark Zolotas and Yiannis Demiris

***Abstract*— Equipping robots with the ability to infer human intent is a vital precondition for effective collaboration. Most computational approaches towards this objective derive a probability distribution of "intent" conditioned on the robot's perceived state. However, these approaches typically assume task-specific labels of human intent are known *a priori*. To overcome this constraint, we propose the Disentangled Sequence Clustering Variational Autoencoder (DiSCVAE), a clustering framework capable of learning such a distribution of intent in an *unsupervised* manner. The proposed framework leverages recent advances in unsupervised learning to disentangle latent representations of sequence data, separating time-varying local features from time-invariant global attributes. As a novel extension, the DiSCVAE also infers a discrete variable to form a latent mixture model and thus enable clustering over these global sequence concepts, *e.g.* high-level intentions. We evaluate the DiSCVAE on a real-world human-robot interaction dataset collected using a robotic wheelchair. Our findings reveal that the inferred discrete variable coincides with human intent, holding promise for collaborative settings, such as shared control.**

I. Introduction

Humans are remarkably proficient at inferring the implicit intentions of others from their overt behaviour. Consequently, humans are adept at planning their actions when collaborating together. Intention inference may therefore prove equally imperative in creating fluid and effective human-robot collaborations. Robots endowed with this ability have been extensively explored [1]–[3], yet their integration into real-world settings remains an open research problem.

One major impediment to real-world instances of robots performing human intention inference is the assumption that a known representation of intent exists. For example, most methods in collaborative robotics assume a discrete set of task goals is known *a priori*. Under this assumption, the robot can infer a distribution of human intent by applying Bayesian reasoning over the entire goal space [3], [4]. Whilst such a distribution offers a versatile and practical representation of intent, the need for predefined labels is not always feasible unless restricted to a specific task scope.

Another fundamental challenge is that many diverse actions often fulfil the same intention. A popular class of probabilistic algorithms for overcoming this challenge are generative models, which derive a distribution of observations by introducing latent random variables to capture any hidden underlying structure. Within the confines of intention inference, the modelled latent space is then presumed to represent all possible causal relations between intentions and observed human behaviour [5]–[7]. The advent of deep generative models, such as Variational Autoencoders (VAEs) [8], [9], has also enabled efficient inference of this latent space from abundant sources of high-dimensional data.

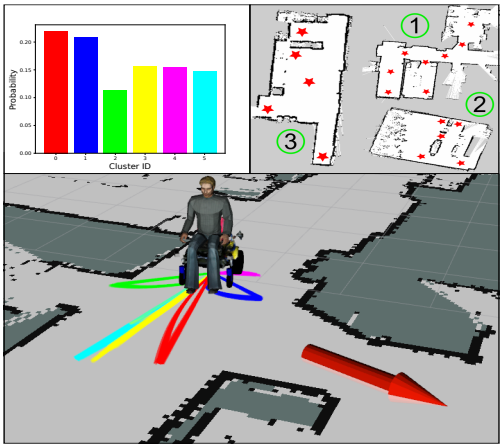


Fig. 1. Overview visualisation of the intention inference experiment on a robotic wheelchair. **Bottom:** Recorded output of an *actual* human subject navigating towards a goal (red arrow). **Top Right:** Maps of the three experiment settings, with red stars denoting target locations. **Top Left:** Probability histogram of the categorical variable modelling "intentions" at this particular snapshot of the data for $K = 6$ clusters. The bars are coloured to align with the wheelchair trajectories generated by sampling from the corresponding clusters. Multiple diverse trajectories can be sampled from the same cluster and each trajectory's length is dependent on the velocity commands drawn from the generative model. Figure best viewed in colour.

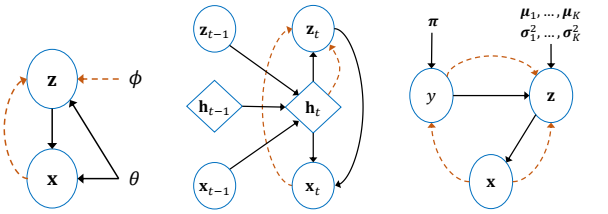
Inspired by the prospects of not only extracting hidden "intent" variables but also interpreting their meaning, we frame the intention inference problem as a process of *disentangling* the latent space. Disentanglement is a core research thrust in representation learning that refers to the recovery of abstract concepts from independent factors of variation assumed to be responsible for generating the observed data [10]–[12]. The interpretable structure of these independent factors is exceedingly desirable for human-in-the-loop scenarios [13], like robotic wheelchair assistance, however few applications have transferred over to the robotics domain [7].

We strive to bridge this gap by proposing an *unsupervised* disentanglement framework suitable for human intention inference. Capitalising on prior disentanglement techniques, we learn a latent representation of sequence observations that divides into a local (time-varying) and global (time-preserving) part [14], [15]. Our proposed variant simultaneously infers a categorical variable to construct a mixture model and thereby form clusters in the *global* latent space. In

M. Zolotas and Y. Demiris are with the Personal Robotics Lab, Dept. of Electrical and Electronic Engineering, Imperial College London, SW7 2BT, UK; Email: {mark.zolotas12, y.demiris}@imperial.ac.uk

This research was supported in part by an EPSRC Doctoral Training Award to MZ, and a Royal Academy of Engineering Chair in Emerging Technologies to YD.

Preprint version; final version available at https://ieeexplore.ieee.org
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) (2022), pp: 9814-9820
Published by: IEEE



(a) VAE (b) VRNN (c) GMVAE

Fig. 2. Deep generative models for: (a) variational inference [8], [9]; (b) a sequential VAE that conditions on the deterministic hidden states of an RNN at each timestep (VRNN [16]); (c) a VAE with a Gaussian mixture prior (GMVAE). Dashed lines denote inference and bold lines indicate generation.

the scope of intention inference, we view the *continuous* local variable as representative of desirable low-level trajectories, whilst the *discrete* counterpart signifies high-level intentions. To summarise, this paper's contributions are:

- A framework for clustering disentangled representations of sequences, coined as the *Disentangled Sequence Clustering Variational Autoencoder (DiSCVAE)*;
- Findings from a robotic wheelchair experiment (see Fig. 1) that demonstrate how clusters learnt without explicit supervision can be interpreted as user-intended navigation behaviours, or strongly correlated with "labels" of such intent in a semi-supervised context.

II. Preliminaries

Before defining the DiSCVAE, we describe supporting material from representation learning, starting with the VAE displayed in Fig. 2a. The VAE is a deep generative model consisting of a generative and recognition network. These networks are jointly trained by applying the reparameterisation trick [8], [9] and maximising the evidence lower bound (ELBO) $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ on the marginal log-likelihood:

$$\begin{aligned}\log p_\theta(\mathbf{x}) &\geq \mathcal{L}_{\theta,\phi}(\mathbf{x}) \qquad (1)\\ &\equiv \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] - \mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p_\theta(\mathbf{z})\big),\end{aligned}$$

where the first term is the reconstruction error of reproducing observations $\mathbf{x}$, and the second KL divergence term is a regulariser that encourages the variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ to be close to the prior $p_\theta(\mathbf{z})$. For notational convenience, parameters ϕ and θ will be omitted hereafter.

Deep generative models can also be parameterised by Recurrent Neural Networks (RNNs) to represent temporal data under the VAE learning principle. A notable example is the Variational RNN (VRNN) [16] shown in Fig. 2b, which conditions on latent variables and observations from previous timesteps via its deterministic hidden state, $\mathbf{h}_t(\mathbf{x}_{t-1}, \mathbf{z}_{t-1}, \mathbf{h}_{t-1})$, leading to the joint distribution:

$$\begin{aligned}p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T}) &= \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{z}_{\leq t}, \mathbf{x}_{<t}) p(\mathbf{z}_t \mid \mathbf{x}_{<t}, \mathbf{z}_{<t}) \qquad (2)\\ &= \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{z}_t, \mathbf{h}_t) p(\mathbf{z}_t \mid \mathbf{h}_t),\end{aligned}$$

where the true posterior is conditioned on information pertaining to previous observations $\mathbf{x}_{<t}$ and latent states $\mathbf{z}_{<t}$, hence accounting for temporal dependencies. The VRNN state $\mathbf{h}_t$ is also shared with the inference procedure to yield the variational posterior distribution:

$$q(\mathbf{z}_{\leq T} \mid \mathbf{x}_{\leq T}) = \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t}) = \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_t). \quad (3)$$

The DiSCVAE developed in the following section elects an approach akin to the VRNN, where latent variables are injected into the forward autoregressive dynamics.

III. Disentangled Sequence Clustering Variational Autoencoder

In this section, we introduce the *Disentangled Sequence Clustering VAE (DiSCVAE)*[1], a framework suited for human intention inference. Clustering is initially presented as a Gaussian mixture adaptation of the VAE prior. The complete DiSCVAE is then specified by combining this adaptation with a sequential model that disentangles latent variables. Finally, we relate back to the intention inference domain.

A. Clustering with Variational Autoencoders

A crucial aspect of generative models is choosing a prior capable of fostering structure or clusters in the data. Previous research has tackled clustering with VAEs by segmenting the latent space into distinct classes using a Gaussian mixture prior, *i.e.* a GMVAE [17], [18].

Our approach is similar to earlier GMVAEs, except for two modifications. First, we leverage the categorical reparameterisation trick to obtain differentiable samples of discrete variables [19], [20]. Second, we alter the ELBO to mitigate the precarious issues of posterior collapse and cluster degeneracy [15]. Posterior collapse refers to latent variables being ignored or overpowered by highly expressive decoders during training, such that the posterior mimics the prior. Whilst cluster degeneracy is when multiple modes of the prior have collapsed into one [17].

The GMVAE outlined below is the foundation for how the DiSCVAE uncovers K clusters (see Fig. 2c). Assuming observations $\mathbf{x}$ are generated according to some stochastic process with discrete latent variable y and continuous latent variable $\mathbf{z}$, then the joint probability can be written as:

$$\begin{aligned}p(\mathbf{x}, \mathbf{z}, y) &= p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z} \mid y) p(y) \qquad (4)\\ y &\sim \mathrm{Cat}(\boldsymbol{\pi})\\ \mathbf{z} &\sim \mathcal{N}\big(\boldsymbol{\mu}_z(y), \mathrm{diag}(\boldsymbol{\sigma}_z^2(y))\big)\\ \mathbf{x} &\sim \mathcal{N}\big(\boldsymbol{\mu}_x(\mathbf{z}), \boldsymbol{I}\big) \text{ or } \mathcal{B}\big(\boldsymbol{\mu}_x(\mathbf{z})\big),\end{aligned}$$

where functions $\boldsymbol{\mu}_z$, $\boldsymbol{\sigma}_z^2$ and $\boldsymbol{\mu}_x$ are neural networks whose outputs parameterise the distributions of $\mathbf{z}$ and $\mathbf{x}$. The generative process involves three steps: (1) sampling y from a categorical distribution parameterised by probability vector $\boldsymbol{\pi}$ with π_k set to K^{-1}; (2) sampling $\mathbf{z}$ from the marginal prior $p(\mathbf{z} \mid y)$, resulting in a Gaussian mixture with a diagonal covariance matrix and uniform mixture weights; and (3) generating data $\mathbf{x}$ from a likelihood function $p(\mathbf{x} \mid \mathbf{z})$.

[1]Code available at: `https://github.com/mazrk7/discvae`

Preprint version; final version available at https://ieeexplore.ieee.org
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) (2022), pp: 9814-9820
Published by: IEEE

A variational distribution $q(\mathbf{z}, y \,|\, \mathbf{x})$ for the true posterior can then be introduced in its factorised form as:

$$q(\mathbf{z}, y \,|\, \mathbf{x}) = q(\mathbf{z} \,|\, \mathbf{x}, y) q(y \,|\, \mathbf{x}), \tag{5}$$

where both the multivariate Gaussian $q(\mathbf{z} \,|\, \mathbf{x}, y)$ and categorical $q(y \,|\, \mathbf{x})$ are also parameterised by neural networks, with respective parameters, ϕ_z and ϕ_y, omitted from notation.

Provided with these inference $q(.)$ and generative $p(.)$ networks, the ELBO for this clustering model becomes:

$$\begin{aligned} \mathcal{L}(\mathbf{x}) &= \mathbb{E}_{q(\mathbf{z}, y \,|\, \mathbf{x})} \left[\log \frac{p(\mathbf{x}, \mathbf{z}, y)}{q(\mathbf{z}, y \,|\, \mathbf{x})} \right] \\ &= \mathbb{E}_{q(\mathbf{z}, y \,|\, \mathbf{x})} \left[\log p(\mathbf{x} \,|\, \mathbf{z}) \right] \\ &\quad - \mathbb{E}_{q(y \,|\, \mathbf{x})} \left[\mathrm{KL}\big(q(\mathbf{z} \,|\, \mathbf{x}, y) \,||\, p(\mathbf{z} \,|\, y)\big) \right] \\ &\quad - \mathrm{KL}\big(q(y \,|\, \mathbf{x}) \,||\, p(y)\big), \end{aligned} \tag{6}$$

where the first term is reconstruction loss of data $\mathbf{x}$, and the latter two terms push the variational posteriors close to their corresponding priors. As the standard reparameterisation trick is intractable for non-differentiable discrete samples, we employ a continuous relaxation of $q(y \,|\, \mathbf{x})$ [19], [20] that removes the need to marginalise over all K class values.

Optimising GMVAEs with powerful decoders is prone to cluster degeneracy due to the over-regularisation effect of the KL term on y opting for a uniform posterior [17]. As KL divergence is a known upper bound on mutual information between a latent variable and data during training [10], [11], we instead penalise mutual information in Eq. 6 by replacing $\mathrm{KL}\big(q(y \,|\, \mathbf{x}) \,||\, p(y)\big)$ with entropy $\mathcal{H}\big(q(y \,|\, \mathbf{x})\big)$ given uniform $p(y)$. We found this modification to be empirically effective at preventing mode collapse and it may even improve the other key trait of the DiSCVAE: disentanglement [11].

B. Model Specification

Having established how to categorise the VAE latent space learnt over static data, we now derive the DiSCVAE (see Fig. 3) as a sequential extension that automatically clusters and *disentangles* representations. Disentanglement amongst sequential VAEs commonly partitions latent representations into *time-invariant* and *time-dependent* subsets [14], [15]. Similarly, we express our disentangled representation of some input sequence $\mathbf{x}_{\leq T}$ at timestep t as $\mathbf{z}_t = [\mathbf{z}_G, \mathbf{z}_{t,L}]$, where $\mathbf{z}_G$ and $\mathbf{z}_{t,L}$ encode *global* and *local* features.

The novelty of our approach lies in how we solely cluster the global variable $\mathbf{z}_G$ extracted from sequences. Related temporal clustering models have either mapped the entire sequence $\mathbf{x}_{\leq T}$ to a discrete latent manifold [13] or inferred a categorical factor of variation to cluster over an *entangled* continuous latent representation [15]. Whereas the DiSCVAE clusters high-level attributes $\mathbf{z}_G$ in isolation from lower-level dynamics $\mathbf{z}_{t,L}$. Furthermore, this proposed formulation plays an important role in our interpretation of intention inference, as is made apparent in Section III-D.

Using the clustering scheme described in Section III-A, we define the generative model $p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T,L}, \mathbf{z}_G, y)$ as:

$$p(\mathbf{z}_G \,|\, y) p(y) \prod_{t=1}^{T} p(\mathbf{x}_t \,|\, \mathbf{z}_{t,L}, \mathbf{z}_G, \mathbf{h}_t^{\mathbf{z}_L}) p(\mathbf{z}_{t,L} \,|\, \mathbf{h}_t^{\mathbf{z}_L}). \tag{7}$$

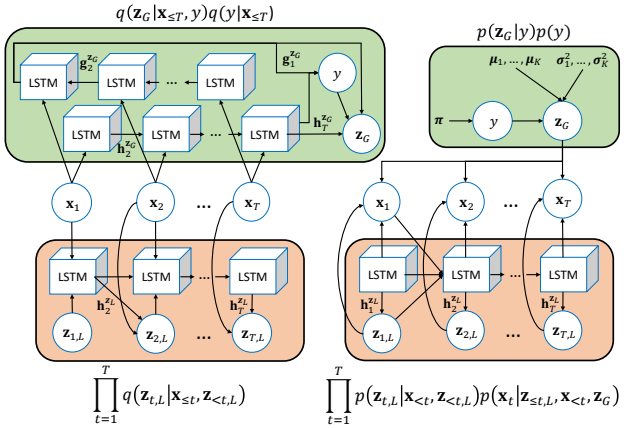


Fig. 3. Computation graph of the inference $q(.)$ and generative $p(.)$ networks. Green blocks contain global variables y and $\mathbf{z}_G$, with a bidirectional LSTM conditioning over input sequence $\mathbf{x}_{\leq T}$. Forward $\mathbf{h}_t^{\mathbf{z}_G}$ and backward $\mathbf{g}_t^{\mathbf{z}_G}$ states then compute the $q(.)$ distribution parameters. Orange blocks encompass the local sequence variable $\mathbf{z}_{t,L}$, where an LSTM's states $\mathbf{h}_t^{\mathbf{z}_L}$ are combined at each timestep with current inputs $\mathbf{x}_t$ to infer $\mathbf{z}_{t,L}$. Generating $\mathbf{x}_t$ requires *both* $\mathbf{z}_G$ and $\mathbf{z}_{t,L}$. Figure best viewed in colour.

The mixture prior $p(\mathbf{z}_G \,|\, y)$ encourages mixture components (indexed by y) to emerge in the latent space of variable $\mathbf{z}_G$. Akin to a VRNN [16], the posterior of $\mathbf{z}_{t,L}$ is parameterised by deterministic state $\mathbf{h}_t^{\mathbf{z}_L}$. We also highlight the dependency on both $\mathbf{z}_{t,L}$ and $\mathbf{z}_G$ upon generating $\mathbf{x}_t$.

To perform posterior approximation, we adopt the variational distribution $q(\mathbf{z}_{\leq T,L}, \mathbf{z}_G, y \,|\, \mathbf{x}_{\leq T})$ and factorise it as:

$$q(\mathbf{z}_G \,|\, \mathbf{x}_{\leq T}, y) q(y \,|\, \mathbf{x}_{\leq T}) \prod_{t=1}^{T} q(\mathbf{z}_{t,L} \,|\, \mathbf{x}_t, \mathbf{h}_t^{\mathbf{z}_L}), \tag{8}$$

with a differentiable relaxation of categorical y injected into the process when training [19], [20].

Under the VAE paradigm, the DiSCVAE is trained by maximising the time-wise objective:

$$\begin{aligned} \mathcal{L}(\mathbf{x}_{\leq T}) &= \mathbb{E}_{q(\cdot)} \left[\log \frac{p(\mathbf{x}_{\leq T}, \mathbf{z}_{\leq T,L}, \mathbf{z}_G, y)}{q(\mathbf{z}_{\leq T,L}, \mathbf{z}_G, y \,|\, \mathbf{x}_{\leq T})} \right] \\ &= \mathbb{E}_{q(\cdot)} \Bigg[\sum_{t=1}^{T} \Big(\log p(\mathbf{x}_t \,|\, \mathbf{z}_{t,L}, \mathbf{z}_G, \mathbf{h}_t^{\mathbf{z}_L}) \\ &\quad - \mathrm{KL}\big(q(\mathbf{z}_{t,L} \,|\, \mathbf{x}_t, \mathbf{h}_t^{\mathbf{z}_L}) \,||\, p(\mathbf{z}_{t,L} \,|\, \mathbf{h}_t^{\mathbf{z}_L})\big) \Big) \\ &\quad - \mathrm{KL}\big(q(\mathbf{z}_G \,|\, \mathbf{x}_{\leq T}, y) \,||\, p(\mathbf{z}_G \,|\, y)\big) \\ &\quad + \mathcal{H}\big(q(y \,|\, \mathbf{x}_{\leq T})\big) \Bigg]. \end{aligned} \tag{9}$$

This summation of lower bounds across timesteps is decomposed into: (1) the expected log-likelihood of input sequences; (2) KL divergences for variables $\mathbf{z}_{t,L}$ and $\mathbf{z}_G$; and (3) entropy regularisation to alleviate mode collapse.

C. Network Architecture

The DiSCVAE is graphically illustrated in Fig. 3. An RNN parameterises the posteriors over $\mathbf{z}_{t,L}$, with the hidden state $\mathbf{h}_t^{\mathbf{z}_L}$ allowing $\mathbf{x}_{<t}$ and $\mathbf{z}_{<t,L}$ to be indirectly conditioned on in Eqs. 7 and 8. For time-invariant variables y and $\mathbf{z}_G$, a bidirectional RNN extracts feature representations from the

Preprint version; final version available at https://ieeexplore.ieee.org
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) (2022), pp: 9814-9820
Published by: IEEE

Algorithm 1: Sampling to produce diverse predictions of goal states from the inferred cluster c

Input: Observation sequence $\mathbf{x}_{\leq t}$; sample length n;
Initialise: $\mathbf{h}_t \leftarrow \mathbf{0}$; $\mathbf{z}_{t,L} \leftarrow \mathbf{0}$;
Output: Predicted states $\tilde{\mathbf{x}}_{t+1}, \ldots, \tilde{\mathbf{x}}_{t+n}$
  Feed prefix $\mathbf{x}_{\leq t}$ into inference model (Eq. 8)
  Assign to cluster c (Eq. 10)
  Draw fixed global sample from $p(\mathbf{z}_G \mid y = c)$
  for $i \in \{t+1, \ldots, t+n\}$ **do**
    Update $\mathbf{h}_i \leftarrow \text{RNN}(\mathbf{z}_{i-1,L}, \mathbf{x}_{i-1}, \mathbf{h}_{i-1})$
    Sample local dynamics from $p(\mathbf{z}_{i,L} \mid \mathbf{h}_i)$
    Predict $\tilde{\mathbf{x}}_i \sim p(\mathbf{x}_i \mid \mathbf{z}_{i,L}, \mathbf{h}_i, \mathbf{z}_G)$
  end for

entire sequence $\mathbf{x}_{\leq T}$, analogous to prior architectures [14]. Bidirectional forward $\mathbf{h}_t$ and reverse $\mathbf{g}_t$ states are computed by iterating through $\mathbf{x}_{\leq T}$ in both directions, before being merged by summation. RNNs have LSTM cells and multilayer perceptrons (MLPs) are dispersed throughout to output the mean and variance of Gaussian distributions.

D. Intention Inference

Let us now recall the problem of intention inference. We first posit that the latent class attribute y could model a K-dimensional repertoire of action plans when considering human interaction data for a specific task. From this perspective, intention inference is a matter of assigning clusters (or action plans) to observations $\mathbf{x}_{\leq T}$ of human behaviour and the environment (*e.g.* joystick commands and sensor data). Human intent is thus computed as the most probable element of the component posterior:

$$c = \arg\max_k q(y_k \mid \mathbf{x}_{\leq T}), \tag{10}$$

where c is the assigned cluster identity, *i.e.* the inferred intention label. The goal associated with this cluster is then modelled by $\mathbf{z}_G$, and local variable $\mathbf{z}_{t,L}$ captures the various behaviours capable of accomplishing the inferred plan.

In the robotic wheelchair scenario, most related works on intention estimation represent user intent [21], [22] as a target wheelchair state $\tilde{\mathbf{x}}_T$. Bayesian reasoning over the entire observation sequence $\mathbf{x}_{\leq T}$ using an entangled latent variable can yield such a state [3], [5], [6]. In contrast, the DiSCVAE employs a disentangled representation $\mathbf{z}_t = [\mathbf{z}_G, \mathbf{z}_{t,L}]$, where the goal state variable is explicitly separated from the user action and environment dynamics. The major benefit of this separation is controlled generation, where repeatedly sampling $\mathbf{z}_{t,L}$ can enable diversity in how trajectories $\tilde{\mathbf{x}}_t$ pan out according to the global plan. The procedure for inferring intention label c amongst a collection of action plans and generating diverse trajectories is summarised in Algorithm 1.

IV. Intention Inference on Robotic Wheelchairs

To validate the DiSCVAE utility at intention inference, we consider a dataset of real users navigating a wheelchair. The objective here is to infer user-intended action plans from observations of their joystick commands and surroundings.

A. Dataset

Eight healthy subjects (aged 25-33) with experience using a robotic wheelchair were recruited to navigate three mapped environments (top right of Fig. 1). Each subject was requested to manually control the wheelchair using its joystick and follow a random route designated by goal arrows appearing on a graphical interface, as in Fig. 1.

Experiment data collected during trials were recorded at a rate of 10 Hz, with sequences of length $T = 20$. This sequence length T is inspired by related work on estimating the short-term intentions of robotic wheelchair operators [22]. Every sequence was composed of user joystick commands $\mathbf{a}_t \in \mathbb{R}^2$ (linear and angular velocities), as well as rangefinder readings $\mathbf{l}_t \in \mathbb{R}^{360}$ (1° angular resolution), with both synchronised to the elected system frequency. The resulting dataset amounted to a total of 8883 sequences.

To assess the generalisability of our intention inference framework, we segregate the dataset based on the experiment environment. As a result, trials that took place in Map 3 are excluded from the training and validation sets, leaving splits of 5881/1580/1422 for training/testing/validation. Dividing the dataset in this way allows us to investigate performance under variations in task context, verifying whether the DiSCVAE can elucidate human intent irrespective of such change.

B. Labelling Routine

Even without access to predefined labels for manoeuvres made by subjects while pursuing task goals, we can appoint approximations of user "intent" to shed light on the analysis. As such, an automated labelling routine is devised below.

Each sequence is initially categorised as either "narrow" or "wide" depending on a measure of threat applied in obstacle avoidance for indoor navigation [23]:

$$s_t = \frac{1}{N} \sum_{i=1}^{N} \text{sat}_{[0,1]} \left(\frac{D_s + R - l_t^i}{D_s} \right), \tag{11}$$

where the aggregate threat score s_t at timestep t for $N = 360$ laser readings l_t^i is a saturated function of these ranges, the robot's radius R (0.5 m for the wheelchair), and a safe distance parameter D_s (set to 0.8 m). In essence, this score reflects the danger of imminent obstacles and qualifies narrow sequences whenever it exceeds a certain threshold.

Next, we discern the intended navigation manoeuvres of participants from the wheelchair's odometry. After empirically testing various thresholds for translational and angular velocities, we determined six manoeuvres: in-place rotations (left/right), forward and reverse motion, as well as forward turns (left/right). This results in 12 classes that account for the influence of both the environment and user actions. Referring to Fig. 1, the majority class across Maps 1 & 2 is the wide in-place rotation (left *and* right), whilst for Map 3 it is the narrow reverse. This switch in label frequency highlights the task diversity caused by different maps.

C. Implementation

The robotic wheelchair has an on-board computer and three laser sensors, two at the front and one at the back

Preprint version; final version available at https://ieeexplore.ieee.org
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) (2022), pp: 9814-9820
Published by: IEEE

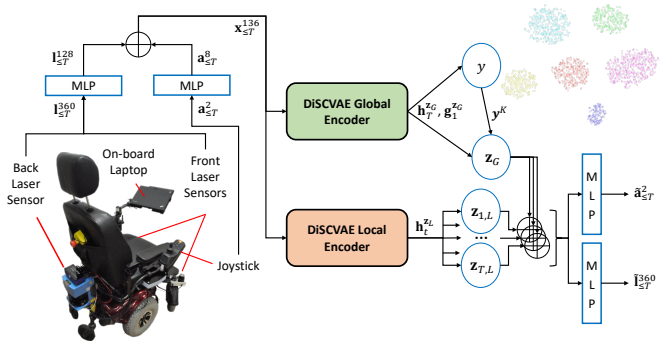


Fig. 4. Architecture for the robotic wheelchair experiment. Joystick and laser data are fed into separate MLPs to produce a concatenated sequence, $\mathbf{x}_t \in \mathbb{R}^{136}$, which feeds into the DiSCVAE encoder (Fig. 3). Forward and backward states, $\mathbf{h}_T^{\mathbf{z}_G}$ and $\mathbf{g}_1^{\mathbf{z}_G}$, allow inference of $\mathbf{z}_G$, whilst $\mathbf{z}_{t,L}$ instead conditions on hidden state $\mathbf{h}_t^{\mathbf{z}_L}$. These latent variables are then passed onto MLPs that decode the joystick commands $\tilde{\mathbf{a}}_t$ and range values $\tilde{\mathbf{l}}_t$.

for a full 360° field of view. For readers interested in the robotic platform, please refer to our earlier work [24].

Fig. 4 portrays the network architecture for this experiment. Before entering the network, input sequences are normalised per modality using the mean and standard deviation of the training set. To process the two input modalities, laser readings $\mathbf{l}_{\leq T}$ and control commands $\mathbf{a}_{\leq T}$ are first passed through separate MLPs. The derived code vectors are then concatenated $\mathbf{x}_{\leq T}$ and fed into the DiSCVAE encoder to infer latent variables $\mathbf{z}_G$ and $\mathbf{z}_{\leq T,L}$. Two individual decoders are conditioned on these variables to reconstruct the original input sequences. Both sensory modalities are modelled as Gaussian variables with fixed variance. No odometry information was supplied at any point to this network.

D. Evaluation Protocol & Model Selection

The evaluation protocol for this experiment is as follows. Although labelled data are unavailable in most practical settings, including ours, we are still interested in digesting the prospects of the DiSCVAE for downstream tasks, such as semi-supervised classification. Accordingly, we train a k-nearest neighbour (KNN) classifier over the learnt latent representation, $\mathbf{z}_G$, and judge intention estimation performance using two pervasive classification metrics: accuracy and the F1-score. Another typical measure in the field is mean squared error (MSE) [6], hence we compare trajectory predictions of user actions $\tilde{\mathbf{a}}_t$ and laser readings $\tilde{\mathbf{l}}_t$ for 10 forward sampled states against "ground truth" future states.

Using this protocol, model selection was conducted on the holdout validation set. A grid search over the network hyperparameters found 512 hidden units to be suitable for the single-layer MLPs (ReLU activations) and bidirectional LSTM states. More layers and hidden units garnered no improvements in accuracy and overall MSE. However, 128 units was chosen for the shared $\mathbf{h}_t^{\mathbf{z}_L}$ state, as higher values had the trade-off of enhancing MSE but worsening accuracy, and so we opted for better classification. Table I also reports on the dimensionality effects of global $\mathbf{z}_G$ and local $\mathbf{z}_{t,L}$ for a fixed model setting. The most noteworthy pattern observed is the steep fall in accuracy when $\dim(\mathbf{z}_{t,L}) > 16$. Given that

TABLE I
HYPERPARAMETER SELECTION ON VALIDATION SET

$\dim(\mathbf{z}_G)$	16			32			64		
$\dim(\mathbf{z}_{t,L})$	16	32	64	16	32	64	16	32	64
Acc (%) ↑	**77.9**	54.3	22.9	72.9	41.4	15	74.9	28.8	14.5
MSE ↓	4.52	4.56	**4.43**	4.69	4.69	4.45	4.47	4.55	4.5

TABLE II
NORMALISED MUTUAL INFORMATION TO DETERMINE K

No. Clusters K	4	6	10	13	16	25	36
NMI ↑	0.141	0.133	0.206	**0.264**	0.244	0.24	0.26

smaller latent spaces raised MSE, a balanced dimensionality of 16 was configured for local and global features.

Another core design choice of the DiSCVAE is to select the number of clusters K. Without access to ground truth labels, we rely on an unsupervised metric, known as Normalised Mutual Information (NMI), to assess clustering quality. The NMI score occupies the range $[0, 1]$ and is thus unaffected by different K clusterings. This metric has also been used amongst similar VAEs for discrete representation learning [13]. Table II provides NMI scores as K varies, where $K = 13$ was settled on due to its marginal superiority and resemblance to the class count from Section IV-B.

E. Experimental Results

Six methods are considered in this experiment, each imitating the same network structure as in Fig. 4:

- **HMM:** A ubiquitous baseline in the literature [3], [6];
- **SeqSVM:** A sequential SVM baseline [5];
- **BiLSTM:** A bidirectional LSTM classifier, akin to [25];
- **VRNN:** An autoregressive VAE model [16];
- **DSeqVAE:** A disentangled sequential autoencoder [14];
- **DiSCVAE:** The proposed model of Section III-B.

The top three *supervised* models learn mappings between the inputs and labels identified in Section IV-B, where baselines utilised the trained BiLSTM encoder for feature extraction. Meanwhile, the bottom three VAE-based methods optimise their respective ELBOs, with a KNN trained on learnt latent variables for a *semi-supervised* approach. Hyperparameters are consistent across methods, *e.g.* equal dimensions for the static and global latent variables of the DSeqVAE and DiSCVAE, respectively. The Adam optimiser [26] was used to train models with a batch size of 32 and initial learning rate of 10^{-3} that exponentially decayed by 0.5 every 10k steps. From the range 3×10^{-3} to 10^{-4}, this learning rate had the most stable and effective ELBO optimisation performance. All models were optimised for 10 runs at different random seeds with early stopping (∼75 epochs for the DiSCVAE).

For qualitative analysis, a key asset of the DiSCVAE is that sampling states from different clusters can exhibit visually diverse characteristics. Fig. 1 portrays sampled trajectories from each mixture component during a subject's recorded interaction. There is clear variability in the trajectory outcomes predicted at this wheelchair configuration ($K = 6$ to ease trajectory visualisation). The histogram over categorical

Preprint version; final version available at https://ieeexplore.ieee.org
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) (2022), pp: 9814-9820
Published by: IEEE

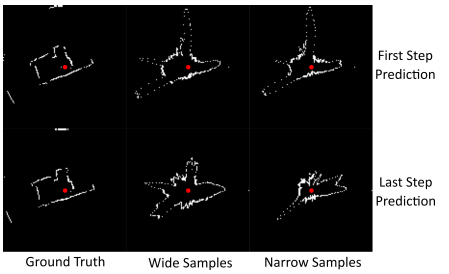


Fig. 5. 2D grids of predicted laser scans on the test set when sampling from "wide" and "narrow" type clusters. Wide samples create spacious proximity around the wheelchair (red dot), whilst narrow samples enclose space.

TABLE III
PERFORMANCE ON TEST SET (10 RANDOM SEEDS)

Model	Acc (%) ↑	F1 ↑	$\tilde{\mathbf{a}}_{MSE}$ ↓	$\tilde{\mathbf{l}}_{MSE}$ ↓
HMM	12.3 ± 2.9	0.09 ± 0.02	-	-
SeqSVM	48.3 ± 0.9	0.41 ± 0.01	-	-
BiLSTM	56.3 ± 1.9	0.43 ± 0.02	-	-
VRNN	65.1 ± 2.8	0.58 ± 0.04	0.15 ± 0.02	2.8 ± 0.04
DSeqVAE	73.2 ± 2.0	0.65 ± 0.02	0.26 ± 0.02	**2.14 ± 0.06**
DiSCVAE	**82.3 ± 1.8**	**0.78 ± 0.03**	**0.14 ± 0.01**	2.7 ± 0.05

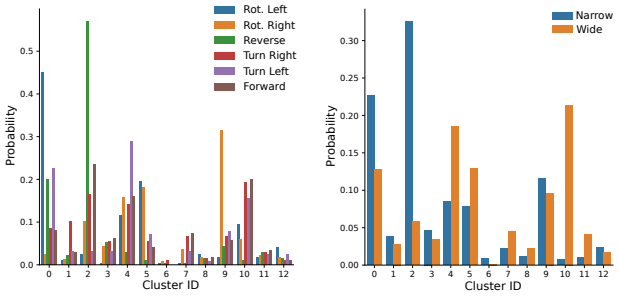
y (top left of Fig. 1) also indicates that the most probable trajectory aligns with the wheelchair user's current goal (red arrow), *i.e.* the correct "intention". As for generating future environment states, Fig. 5 displays how samples from clusters manifest when categorised as either "wide" or "narrow".

Table III contains quantitative results for this experiment. As anticipated, the highly variable nature of wheelchair control in an unconstrained navigation task makes classifying intent challenging. The baselines perform poorly and even the supervised BiLSTM obtains a classification accuracy of merely 56.3% on the unseen test environment. Nevertheless, learning representations of user interaction data can reap benefits in intention inference, as performance is drastically improved by a KNN classifier trained over the latent spaces of the VAE-based methods. The DiSCVAE acquires the best accuracy, F1-scores and MSE on joystick commands. The DSeqVAE instead attains the best error rates on forecasted laser readings at the expense of under-representing the relevant low-dimensional joystick signal. Cluster specialisation in the DiSCVAE may explain the better $\tilde{\mathbf{a}}_{MSE}$.

F. Illuminating the Clusters

Straying away from the purely discriminative task of classifying intent, we now use our framework to decipher the navigation behaviours, or "global" factors of variation, intended by users. In particular, we plot assignment distributions of y on the test set examples to understand the underlying meaning of our clustered latent space. "Local" factors of variation in this application capture temporal dynamics in state, *e.g.* wheelchair velocities.

Fig. 6a provides further clarity on how certain clusters have learnt independent wheelchair manoeuvres. For instance, cluster 2 is distinctly linked with the wheelchair's reverse motion. Likewise, clusters 0 and 9 pair with left and right in-place rotations. The spatial state assignments shown in Fig. 6b also delineate how these clusters are most often categorised as "narrow", which is to be expected of evasive actions taking place in cluttered spaces. On the contrary, predominantly forward-oriented manoeuvres fall into "wide" clusters (*e.g.* 4 and 10). These findings suggest that wheelchair action plans have been aptly inferred.



(a) Wheelchair Manoeuvres (b) Spatial States

Fig. 6. Assignment distribution of y for $K = 13$ with post-processed labels for **(a)** wheelchair manoeuvres and **(b)** perceived spatial context. The plot illuminates how various clusters are associated with user intent under different environmental conditions. For example, most backward motion and "narrow" state samples reside in cluster 2. Similar patterns are noticeable for in-place rotations (0 and 9) and "wide" forward motion (4 and 10).

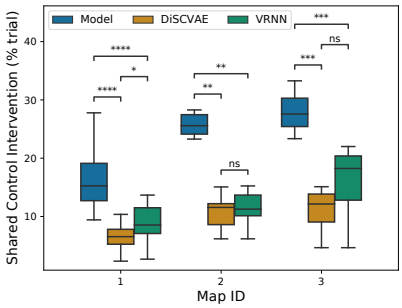


Fig. 7. Percentage of trials per map where shared control would have wrongly intervened. The Model approach is significantly more likely to trigger incorrect assistance. Less variable VRNN and DiSCVAE performance across maps also hints at better robustness to changes in task conditions.

G. Prospects for Shared Control

Lastly, we examine a shared control use-case, where intention inference plays a vital role [1]. Shared control concerns the interaction between robots and humans when *both* exert control over a system to accomplish a common goal [2]. Despite shared control being inactive for this experiment, we simulate its operation in post-processing to gauge success.

More precisely, we address the known issue in shared control of administering *wrong* assistance whenever there is a misalignment between the robot's and user's internal models. To quantify this mismatch, we monitor the percentage of each navigation trial where a shared control methodology [24] would have intervened had it been operational. Given how the subjects are experienced, healthy individuals that incurred no wheelchair collisions, it is safe to assume they never required assistance. We compare wheelchair trajectories produced by the VRNN, DiSCVAE, and a constant velocity "Model" using differential drive kinematics.

Fig. 7 offers results on shared control intervention rates. Performing the two-sided Mann-Whitney U test finds significantly better rates for the VRNN and DiSCVAE over

Preprint version; final version available at https://ieeexplore.ieee.org
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) (2022), pp: 9814-9820
Published by: IEEE

the Model across all maps ($p \leq 0.01$). Excluding Map 1 ($p \leq 0.05$), the positive trend in the DiSCVAE surpassing the VRNN is not significant. Though the DiSCVAE has the advantage of capturing uncertainty around its estimated intent via the categorical y, *e.g.* when a strict left-turn is hard to distinguish from a forward left-turn (blue and red in Fig. 1). This holds potential for shared control seeking to realign mismatched internal models by explaining to a user why the robot chose not to assist under uncertainty [24].

V. Discussion

There are a few notable limitations to this work. One is that learning disentangled representations is sensitive to hyperparameter tuning, as shown in Section IV-D. To aid with model selection and prevent posterior collapse, further investigation into different architectures and other information theoretic advances is thus necessary [10], [11]. Moreover, disentanglement and interpretability are difficult to define, often demanding access to labels for validation [10], [12]. Therefore, a study into whether users believe the DiSCVAE representations of intent are "interpretable" or helpful for the wheelchair task is integral in claiming disentanglement.

In human-robot interaction tasks, intention recognition is typically addressed by equipping a robot with a probabilistic model that infers intent from human actions [3], [4]. Whilst the growing interest in scalable learning techniques for modelling agent intent has spurred on applications in robotics [7], [25], disentanglement learning remains sparse in the literature. The only known comparable work to ours is a conditional VAE that disentangled latent variables in a multi-agent driving setting [7]. Albeit similar in principle, we believe our approach is the first to infer a discrete "intent" variable from human behaviour by clustering action plans.

VI. Conclusions

In this paper, we embraced an unsupervised outlook on human intention inference through a framework that disentangles and clusters latent representations of input sequences. A robotic wheelchair experiment on intention inference gleaned insights into how our proposed DiSCVAE could discern primitive action plans, *e.g.* rotating in-place or reversing, without supervision. The elevated classification performance in semi-supervised learning also posits that disentanglement is a worthwhile avenue to explore in intention inference.

There are numerous promising research directions for an unsupervised means of inferring intent in human-robot interaction. The task-agnostic prior and inferred global latent variable could be exploited in long-term downstream tasks, such as user modelling, to augment the wider adoption of collaborative robotics in unconstrained environments. A truly interpretable latent structure could also prove fruitful in assistive robots that warrant explanation by visually relaying inferred intentions back to end-users [24].

References

[1] Y. Demiris, "Prediction of intent in robotics and multi-agent systems," *Cogn. Process.*, vol. 8, no. 3, pp. 151–158, 2007.

[2] D. P. Losey, C. G. McDonald, E. Battaglia, and M. K. O'Malley, "A Review of Intent Detection, Arbitration, and Communication Aspects of Shared Control for Physical Human-Robot Interaction," *Appl. Mech. Rev.*, vol. 70, no. 1, 2018.

[3] S. Jain and B. Argall, "Probabilistic Human Intent Recognition for Shared Autonomy in Assistive Robotics," *J. Hum. Robot Interact.*, vol. 9, no. 1, 2019.

[4] S. Javdani, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization," *Robot. Sci. Syst.: online proceedings*, 2015.

[5] Z. Wang, K. Mülling, M. P. Deisenroth, H. B. Amor, D. Vogt, B. Schölkopf, and J. Peters, "Probabilistic movement modeling for intention inference in human-robot interaction," *Int. J. Rob. Res.*, vol. 32, no. 7, pp. 841–858, 2013.

[6] A. K. Tanwani and S. Calinon, "A generative model for intention recognition and manipulation assistance in teleoperation," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 43–50.

[7] Y. Hu, W. Zhan, L. Sun, and M. Tomizuka, "Multi-modal Probabilistic Prediction of Interactive Behavior via an Interpretable Model," in *IEEE Intell. Veh. Symp. Proc.*, 2019, pp. 557–563.

[8] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114*, 2013.

[9] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," in *Int. Conf. Mach. Learn.*, 2014, pp. 1278–1286.

[10] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[11] E. Dupont, "Learning Disentangled Joint Continuous and Discrete Representations," in *Adv. Neural Inf. Process. Syst.*, 2018, pp. 710–720.

[12] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," in *Int. Conf. Mach. Learn.*, 2019, pp. 4114–4124.

[13] V. Fortuin, M. Hüser, F. Locatello, H. Strathmann, and G. Rätsch, "Som-vae: Interpretable discrete representation learning on time series," *arXiv:1806.02199*, 2018.

[14] L. Yingzhen and S. Mandt, "Disentangled Sequential Autoencoder," in *Int. Conf. Mach. Learn.*, 2018, pp. 5670–5679.

[15] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, *et al.*, "Hierarchical Generative Modeling for Controllable Speech Synthesis," *arXiv:1810.07217*, 2018.

[16] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A Recurrent Latent Variable Model for Sequential Data," in *Adv. Neural Inf. Process. Syst.*, 2015, pp. 2980–2988.

[17] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders," *arXiv:1611.02648*, 2016.

[18] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering," in *Int. Jt. Conf. Artif. Intell.*, 2017, pp. 1965–1972.

[19] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv:1611.00712*, 2016.

[20] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv:1611.01144*, 2016.

[21] T. Carlson and Y. Demiris, "Collaborative Control for a Robotic Wheelchair: Evaluation of Performance, Attention, and Workload," *IEEE Trans. Syst. Man Cybern.*, vol. 42, no. 3, pp. 876–888, 2012.

[22] J. Poon, Y. Cui, J. V. Miro, T. Matsubara, and K. Sugimoto, "Local driving assistance from demonstration for mobility aids," in *IEEE Int. Conf. Robot. Autom.*, 2017, pp. 5935–5941.

[23] J. W. Durham and F. Bullo, "Smooth Nearness-Diagram Navigation," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2008, pp. 690–695.

[24] M. Zolotas and Y. Demiris, "Towards Explainable Shared Control using Augmented Reality," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 3020–3026.

[25] D. Nicolis, A. M. Zanchettin, and P. Rocco, "Human intention estimation based on neural networks for enhanced collaboration with robots," in *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1326–1333.

[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980*, 2014.