Predicting the What and How - A Probabilistic Semi-Supervised Approach to Multi-Task Human Activity Modeling

Judith Bütepage Hedvig Kjellström Danica Kragic
Robotics, Perception and Learning
KTH Royal Institute of Technology
butepage@kth.se, hedvig@kth.se, dani@kth.se

Abstract

Video-based prediction of human activity is usually performed on one of two levels: either a model is trained to anticipate high-level action labels or it is trained to predict future trajectories either in skeletal joint space or in image pixel space. This separation of classification and regression tasks implies that models cannot make use of the mutual information between continuous and semantic observations. However, if a model knew that an observed human wants to drink from a nearby glass, the space of possible trajectories would be highly constrained to reaching movements. Likewise, if a model had predicted a reaching trajectory, the inference of future semantic labels would rank "lifting" more likely than "walking". In this work, we propose a semi-supervised generative latent variable model that addresses both of these levels by modeling continuous observations as well as semantic labels. This fusion of signals allows the model to solve several tasks, such as action detection and anticipation as well as motion prediction and synthesis, simultaneously. We demonstrate this ability on the UTKinect-Action3D dataset, which consists of noisy, partially labeled multi-action sequences. The aim of this work is to encourage research within the field of human activity modeling based on mixed categorical and continuous data.

1. Introduction

Prediction of future events plays a key role in human and animal cognition. On the one hand, the brain constantly predicts the sensory consequences produced by ones own actions. On the other hand, it also attempts to infer the intentions and to anticipate the actions of other agents in the environment. Computer vision applications such as human-robot interaction or autonomous vehicles can benefit from the implementation of similar predictive processes as they allow for adaptive action execution and task planning.

This work was supported by the EU through the project socSMCs (H2020-FETPROACT-2014) and the Swedish Foundation for Strategic Research.

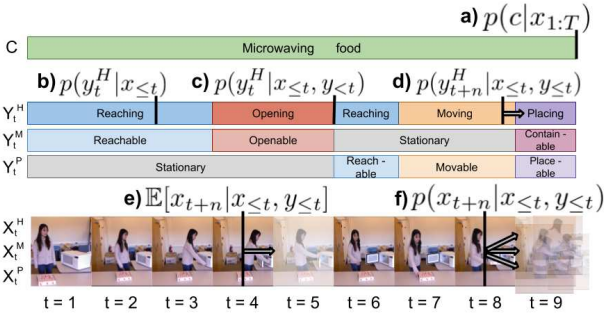


Figure 1: Among others, human activity modeling is concerned with **a)** action classification, **b)** action prediction, **c)** action detection, **d)** action anticipation, **e)** motion prediction and **f)** motion synthesis. The black bars indicate when the respective decision, e.g. classification, is made. Images belong to the CAD - 120 dataset [7].

Most approaches towards human activity modeling focus on either discrete, semantic labels or on continuous trajectory prediction. Label classification, prediction and detection (Figure 1a), b) and c)) are supposed to classify observed trajectories either at the end of a sequence (classification), as soon as possible (prediction) or at action onset (detection). Only action anticipation (Figure 1d)) is concerned with inferring labels of future actions. Human motion prediction and synthesis (Figure 1e) and 1f)) on the other hand aim at modeling the future continuous motion trajectories given past observations. Compared to prediction, motion synthesis should anticipate different possible trajectories instead of only the most likely one.

From a modeling perspective, these different types of tasks and mixed categorical and continuous data should influence each other. A model that is able to anticipate a future label should be better at detecting the actual onset of the action. Likewise a model that carries information about the ongoing action should be able to synthesize different appropriate motion trajectories.

In this work, we present a generative, temporal latent variable model that can capture the complex dependencies of continuous features as well as discrete labels over time. In detail, we propose a semi-supervised variational recurrent neural network (SVRNN), as described in Section 3.1, which inherits the generative capacities of a variational autoencoder (VAE) [6, 9], extends these to temporal data [2] and combines them with a discriminative model in a semi-supervised fashion. The semi-supervised VAE [5] can handle labeled and unlabeled data. This property allows us to propagate label information over time even during testing and therefore to generate possible future action and motion sequences.

We demonstrate the ability of our model to represent mixed categorical and continuous data in an anticipatory fashion on the UTKinect-Action3D Dataset [10].

2. Background

Our approach builds on three basic ingredients which are introduced below.

2.1. Variational autoencoders

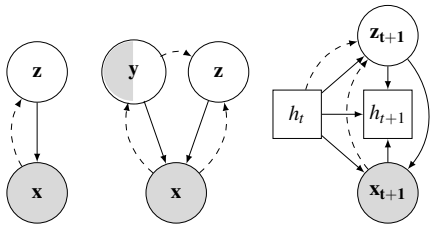
Our model builds on VAEs, latent variable models that are combined with an amortized version of variational inference (VI). Amortized VI employs neural networks to learn a function from the data $\mathbf{x}$ to a distribution over the latent variables $q(\mathbf{z}|\mathbf{x})$ that approximates the posterior $p(\mathbf{z}|\mathbf{x})$. Likewise, they learn the likelihood distribution as a function of the latent variables $p(\mathbf{x}|\mathbf{z})$. This mapping is depicted in Figure 2a). Instead of having to infer N local latent variables for N observed data points, as common in VI, amortized VI requires only the learning of neural network parameters of the functions $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{z})$. We call $q(\mathbf{z}|\mathbf{x})$ the recognition network and $p(\mathbf{x}|\mathbf{z})$ the generative network. To sample from a VAE, we first draw a sample from the prior $\mathbf{z} \sim p(\mathbf{z})$ which is then fed to the generative network to yield $\mathbf{x} \sim p(\mathbf{x}|z)$

2.2. Semi-supervised variational autoencoders

To incorporate label information when available, semi-supervised VAEs (SVAE) [5] include a label $\mathbf{y}$ into the generative process $p(\mathbf{x}|\mathbf{z},\mathbf{y})$ and the recognition network $q(\mathbf{z}|\mathbf{x},\mathbf{y})$, as shown in Figure 2b). To handle unobserved labels, an additional approximate distribution over labels $q(\mathbf{y}|\mathbf{x})$ is learned which can be interpreted as a classifier. When no label is available, the discrete label distribution can be marginalized out, e.g. $q(\mathbf{z}|\mathbf{x}) = \sum_{\mathbf{y}} q(\mathbf{z}|\mathbf{x},\mathbf{y})q(\mathbf{y}|\mathbf{x})$.

2.3. Recurrent variational autoencoders

VAEs can also be extended to temporal data, so called variational recurrent neural networks (VRNN) [2]. Instead of being stationary as in standard VAEs, the prior over the latent variables depends on past observations $p(\mathbf{z_t}|h_{t-1})$, which are encoded in the hidden state of an RNN h_{t-1}. Similarly, the approximate distribution $q(\mathbf{z_t}|\mathbf{x_t}, h_{t-1})$ depends on the history as can be seen in Figure 2c). The advantage of this structure is that data sequences can be generated by sampling from the temporal prior instead of an uninformed prior, i.e. $\mathbf{z_t} \sim p(\mathbf{z_t}|h_{t-1})$.



a) VAE b) SVAE c) VRNN

Figure 2: Model structure of the VAE **a)**, its semi-supervised version SVAE **b)**, and the recurrent model VRNN **c)**. Random variables (circle) and states of RNN hidden units (square) are either observed (gray), unobserved (white) or partially observed (white-gray). The dotted arrows indicate inference connections.

3. Methodology

Equipped with the background knowledge, we will now describe the structure of our proposed model, semi-supervised variational recurrent neural networks (SVRNN), and the inference procedure applied to train them.

3.1. SVRNN

In the SVRNN, the model is trained on a dataset with temporal structure $D = \{D^L, D^U\}$ consisting of the set L of labeled time steps $D^L = \{\mathbf{x_t}, \mathbf{y_t}\}_{t \in L} \sim \widetilde{p}(\mathbf{x_t}, \mathbf{y_t})$ and the set U of unlabeled observations $D^U = \{\mathbf{x_t}\}_{t \in U} \sim \widetilde{p}(\mathbf{x_t})$. $\widetilde{p}$ denotes the empirical distribution. Further we assume that the temporal process is governed by latent variables $\mathbf{z_t}$, whose distribution $p(\mathbf{z_t}|h_{t-1})$ depends on a deterministic function of the history up to time t: $h_{t-1} = f(x_{<t}, y_{<t}, z_{<t})$. The generative process is as follows

$$\mathbf{y_t} \sim p(\mathbf{y_t}|h_{t-1}),\ \mathbf{z_t} \sim p(\mathbf{z_t}|\mathbf{y_t}, h_{t-1}),\ \mathbf{x_t} \sim p(\mathbf{x_t}|\mathbf{y_t}, \mathbf{z_t}, h_{t-1}),$$

where $p(\mathbf{y_t}|h_{t-1})$ and $p(\mathbf{z_t}|\mathbf{y_t}, h_{t-1})$ are time-dependent priors, as shown in Figure 3a). To fit this model to the dataset at hand, we need to estimate the posterior over the unobserved variables $p(\mathbf{y_t}|\mathbf{x_t}, h_{t-1})$ and $p(\mathbf{z_t}|\mathbf{x_t}, \mathbf{y_t}, h_{t-1})$ which is intractable. Therefore we resign to amortized VI and approximate the posterior with a simpler distribution $q(\mathbf{y_t}, \mathbf{z_t}|\mathbf{x_t}, h_{t-1}) = q(\mathbf{y_t}|\mathbf{x_t}, h_{t-1})q(\mathbf{z_t}|\mathbf{x_t}, \mathbf{y_t}, h_{t-1})$, as shown in Figure 3b). To minimize the distance between the approximate and posterior distributions, we optimize the variational lower bound of the marginal likelihood $\mathcal{L}(p(D))$.

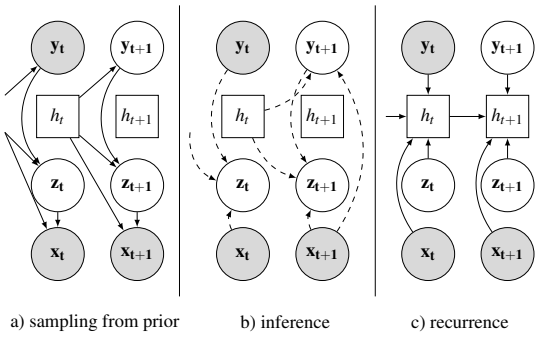


Figure 3: Information flow through SVRNN. a) Passing samples from the prior through the generative network. b) Information passing through the inference network. c) The recurrent update. Node appearance follows Figure 2.

As the distribution over $\mathbf{y_t}$ is only required when it is unobserved, the bound decomposes as follows

$$\mathscr{L}(p(D)) \geq \mathscr{L}^L + \mathscr{L}^U + \alpha\mathscr{T}^L \tag{1}$$

$$-\mathscr{L}^L = \sum_{t\in L} \mathbb{E}_{q(\mathbf{z_t}|\mathbf{x_t},\mathbf{y_t},h_{t-1})}[log(p(\mathbf{x_t}|\mathbf{y_t},\mathbf{z_t},h_{t-1}))] \tag{2}$$
$$- KL(q(\mathbf{z_t}|\mathbf{x_t},\mathbf{y_t},h_{t-1})||p(\mathbf{z_t}|\mathbf{y_t},h_{t-1})) + log(p(\mathbf{y_t}))$$

$$\mathscr{T}^L = -\sum_{t\in L} \mathbb{E}_{\widetilde{p}(\mathbf{y_t},\mathbf{x_t})} log(p(\mathbf{y_t}|h_{t-1})q(\mathbf{y_t}|\mathbf{x_t},h_{t-1})) \tag{3}$$

$$-\mathscr{L}^U = \sum_{t\in U} \mathbb{E}_{q(\mathbf{y_t},\mathbf{z_t}|\mathbf{x_t},h_{t-1})}\big[log(p(\mathbf{x_t}|\mathbf{y_t},\mathbf{z_t},h_{t-1}))] \tag{4}$$
$$- KL(q(\mathbf{z_t}|\mathbf{x_t},\mathbf{y_t},h_{t-1})||p(\mathbf{z_t}|\mathbf{y_t},h_{t-1}))$$
$$- KL(q(\mathbf{y_t}|\mathbf{x_t},h_{t-1})||p(\mathbf{y_t}|h_{t-1})).$$

$\mathscr{L}^L$ and $\mathscr{L}^U$ are the lower bounds for labeled and unlabeled data points respectively, while $\mathscr{T}^L$ is an additional term that encourages $p(\mathbf{y_t}|h_{t-1})$ and $q(\mathbf{y_t}|\mathbf{x_t},h_{t-1})$ to follow the data distribution over $\mathbf{y_t}$. This lower bound is optimized jointly. We assume the latent variables $\mathbf{z_t}$ to be i.i.d. Gaussian distributed. The categorical distribution over $\mathbf{y_t}$ is determined by parameters $\pi = \{\pi_i\}_{i=1:N_{class}}$). To model such discrete distributions, we apply the Gumbel trick [4, 8]. The history h_{t-1} is modeled with the last of three Long short-term memory (LSTM) units. For more details, we refer the reader to the background work discussed in Section 2.

3.2. Predict, detect, anticipate and generate

To solve the different tasks, we make use of the different components of our model in the following way.

Predict and detect actions: To classify or detect at time t, we choose the largest of the weights $\pi^{q_y} = \{\pi_i^{q_y}\}_{i=1:N_{class}}$ of the categorical distribution $q(\mathbf{y_t}|\mathbf{x_t},h_{t-1})$. Prediction and detection are performed at all time steps.

Anticipate actions: To anticipate a label after time t', we make use of the prior, which does not depend on the current observation $\mathbf{x_t}$. Thus, for time $t'+1$, we choose the largest of the weights $\pi^{p_y} = \{\pi_i^{p_y}\}_{i=1:N_{class}}$ of the categorical distributions $p(\mathbf{y_t}|h_{t-1})$. To anticipate several steps into the future, we need to generate both future observations and future labels as described below.

Predict and generate motion: To sample an observation sequence $\{\mathbf{x_t},\mathbf{y_t}\}_{t>t'}$ after time t', we propagate the sampled observations and generate with help of the approximate distribution $\mathbf{y_t} \sim q(\mathbf{y_t}|\mathbf{x_t},h_{t-1})$, $\mathbf{z_t} \sim q(\mathbf{z_t}|\mathbf{x_t},\mathbf{y_t},h_{t-1})$, $\mathbf{x_t} \sim p(\mathbf{x_t}|\mathbf{y_t},\mathbf{z_t},h_{t-1})$ for each t. This method is used to predict a sequence, by averaging over several samples of the distributions.

4. Experiments

In this section, we describe both experimental design and results. For details about model architectures and the training procdure we refer the reader to [1]. We apply our model to the UTKinect-Action3D Dataset (UTK) [10], which consists of 10 subjects each recorded twice performing 10 actions in a row. The sequences are recorded with a Kinect device (30 fps) and the extracted skeletons consist of 20 joints. Due to high inter-subject, intra-class and view-point variations, this dataset is challenging. The actions in each recording do not immediately follow each other but are disrupted by long periods of unlabeled frames. As our model is semi-supervised, these unobserved data labels can be incorporated naturally and do not require the introduction of e.g. an additional *unknown* label class. We train our model on five subjects and test on the remaining five subjects.

4.1. Action classification, detection and prediction

In this section, we focus on the capabilities of our models to detect and predict semantic labels. As far as we are aware, only one comparable work, based on class templates [3], has attempted to detect actions on the UTK dataset. We assume action a to be detected if the majority of observations within the ground truth time interval are inferred to belong to action a.

In Table 1, we see that the model is able to detect actions with only a short or no delay. This is apparent when we measure the F1 score for partially observed action sequences, namely when the model has observed 25 %, 50 %, 75 % or 100 % of the current. We present results for action detection in context of the previous actions, i.e., on the unsegmented sequence (unseg), and for

Table 1: F1 score for action prediction with history (with H) and without history (without H) on the UTK dataset.

Observed	25 %	50 %	75 %	100 %
CT [3]	-	-	-	81.8
SVRNN (unseg)	61.0	78.0	84.0	**89.0**
SVRNN (seg)	29.0	48.0	67.0	74.0

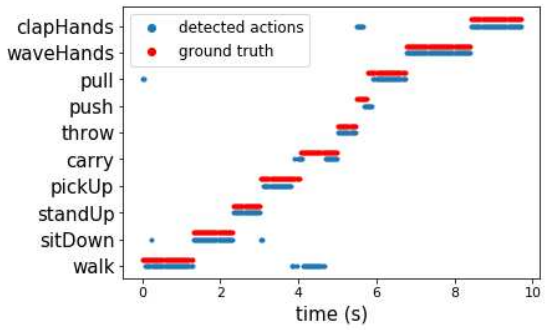


Figure 4: The detected and ground truth actions of a single test recording from the UTK dataset over time. We only display the labeled frames of the test sequence.

action prediction based only on the current action segment (seg). On average, this corresponds to having observed 8, 16, 25 or 33 frames of the ongoing action. As listed in Table 1, the F1 score increases continuously the more of the action has been observed. At 75 % the SVRNN outperforms the results reported in [3] which are based on 100 % of the action interval. When segmented, the performance is lower as our model has not been trained to predict actions without history. Further, we visualize the detected and ground truth action sequence of one unsegmented test sample in Figure 4 and in form of a video in here https://www.youtube.com/watch?v=XfgztgOhuCk. In this test sequence, the action *carry* is partially confused with *walking* which might be caused by the lack of meta-data such as that the subject is holding an object.

4.2. Action anticipation and motion prediction

In the previous section, we focused on early action detection based on an observed data stream. Here we present the ability of the model to anticipate both future semantic actions and motion trajectories. As described in Section 3.2, we feed an observed continuous data sequence to the model up to time t' and let the model infer labels and joint positions for a certain number of future frames. In Figure 5 we visualize the action transition from *sitting* to *standing up*. The green labels and skeletons are inferred by the model, which is able to anticipate the upcoming semantic action and to generate a matching movement.

5. Conclusion

We presented a principled probabilistic approach to fuse mixed categorical and continuous data for predictive human activity modeling. Our model can be used to anticipate human behavior in real-time based on noisy observation, such as e.g. skeletal Kinect recordings. For a more in-depth discussion and additional experiments, we would like to point the reader to [1].

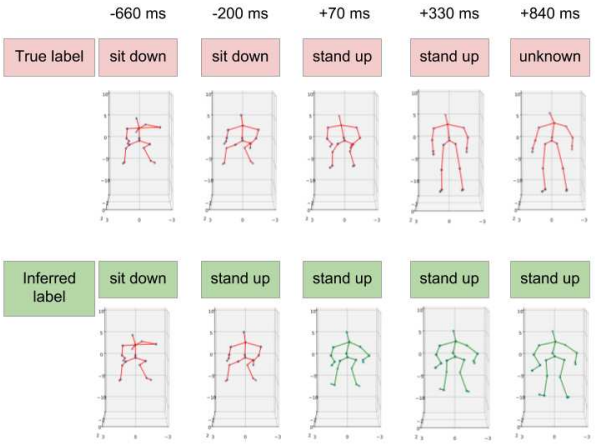


Figure 5: The ground truth actions and trajectories (top) and the inferred labels and trajectories (bottom). From time 0 ms and onward, the model propagates its own predictions and does not receive any ground truth data anymore.

References

[1] J. Bütepage, H. Kjellström, and D. Kragic. A probabilistic semi-supervised approach to multi-task human activity modeling. *arXiv preprint arXiv:1809.08875*, 2018.

[2] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Annual Conference on Neural Information Processing Systems*, 2015.

[3] K. Gupta and A. Bhavsar. Scale invariant human action detection from depth cameras using class templates. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[4] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

[5] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Annual Conference on Neural Information Processing Systems*, 2014.

[6] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:*1312.6114, 2013.

[7] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[8] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[9] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[10] L. Xia, C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012.