Perceptual Facial Expression Representation

Olga Mikheeva[1] Carl Henrik Ek[2] Hedvig Kjellström[1]
[1] Dept. Robotics, Perception, and Learning, KTH Royal Institute of Technology, Sweden
[2] Dept. Computer Science, University of Bristol, UK
olgamik@kth.se carlhenrik.ek@bristol.ac.uk hedvig@kth.se

***Abstract*— Dissimilarity measures are often used as a proxy or a handle to reason about data. This can be problematic, as the data representation is often a consequence of the capturing process or how the data is visualized, rather than a reflection of the semantics that we want to extract. Facial expressions are a subtle and essential part of human communication but they are challenging to extract from current representations. In this paper we present a method that is capable of learning semantic representations of faces in a data driven manner. Our approach uses sparse human supervision which our method grounds in the data. We provide experimental justification of our approach showing that our representation improves the performance for emotion classification.**

I. INTRODUCTION

During the last decades, there has been a rapid development of methods for machine analysis of human spoken language. Very sophisticated commercial systems for automatic speech understanding and synthesis are now a part of our everyday life, and runs on the smallest devices. However, the information communicated between interacting humans is only to a small part contained in the spoken language; humans transfer huge amounts of information through non-verbal cues: facial expressions, body motion, muscle tensions, and also the voice prosody [1], [17].

For an artificial social agent to be able to interact with humans in a socially capable manner, it is thus essential to implement reasoning about and recognition from such non-verbal signals into the agent's system.

Facial expressions are important in this respect, where a small raise of the eyebrows can alter a reaction from confused to surprised. Facial data can be acquired at large volume from image and video data. However, expressions are very subtle and correspond to a very small part of the variations in a video as most of the information are related to the appearance. This remains true even when removing the visual appearance and considering only the movement of landmark points on a face. Therefore, reasoning directly from raw data is challenging. One approach is to use representations that is explicitly designed to construct a semantic feature space. One such representation is the Facial Action Coding System (FACS) [7] which is based on features corresponding to facial muscle activations. Even though such representations correlate well with emotion expressions, they are cumbersome to extract from video.

In this work we take a different approach and develop a generative model of facial landmarks which is characterized by a low-dimensional latent variable. Our method uses a human notion of similarity to structure this latent representation, providing an interpretable representation.

The main contribution of this work is this latent face representation (Section III). In addition we have collected a data set of partial similarity rankings in the form of triplets (Section IV-B).

978-1-5386-2335-0/18/$31.00

II. RELATED WORK

Classic approaches to feature extraction for facial expression applications (e.g. automatic classification) have traditionally been characterized by implicit hand designed features. One of the most commonly used representations is the Facial Action Coding System [7], [24]. In the FACS system facial expressions are described using a set of facial action units (AU), where each action unit corresponds to different muscle movement (e.g. "Inner Brow Raiser", "Upper Lip Raiser", etc.). There are 28 main action units. The intensity of activations is measured on a discrete 5-level scale.

The system has been very successful with a wide variety of applications and is considered the standard representation for physical expression of emotions. In the traditional approach, action units are defined by hand but there has also been work to learn action units directly from data [8]. FACS models are successful but they lead to representations which are discrete and spanned by the action units who are highly dependent. This makes it challenging to build models which use FACS as an underlying parametrization.

With the increase in available data there has been a movement towards learning representations directly from data. In certain domains such as computer vision the resurgence of neural networks in combination with large volumes of data have completely replaced hand designed features with data-driven features. These methods have recently made inroads also for the task of facial expression recognition [22].

However, these methods do not learn representations but decisions and it is not clear what assumptions underlie their success. In this paper we want to leverage the available data but also learn an *interpretable* representation which corresponds well to human perception. Our goal is to recover and parametrize a latent representation of facial data where the notion of similarity aligns with human perceptual similarity.

The task of learning latent representations in an unsupervised manner is a well studied but highly ill-constrained problem requiring significant assumptions to proceed. Probably the most well known assumption is to assume the

latent representation to be Gaussian. In the linear case this leads to Principal Component Analysis [11] but can also be generalized to the non-linear case leading to what is known as a Gaussian process latent variable model [18]. In the case of non-linear transformation, full Bayesian inference is not analytically tractable but efficient variational approximations [26] have been developed.

The popularity of deep learning methods have led to significant developments in representation learning. The biggest challenge with this approach is that the assumptions that lead to the representations are not clear as there is no clear model specification. One type of network structure that has been very successful is based on auto-encoders. These models are composed of an 'encoder' that maps from the observed data to a latent representation and a 'decoder' that maps from the latent back to the data. This approach has been very successful as the latent space is constrained from "two sides" both to reconstruct the data and to be encoded by it. A very attractive type of model that merges benefits of models with that of neural networks in an auto-encoder structure is the Variational auto-encoder (VAE) [16], [6]. This approach allows to incorporate assumptions over the latent representation. Fundamentally it is a probabilistic directed graphical model with latent random variables and observed random variables. The generating process is modeled as a function of a latent variable $\mathbf{z}$ via a neural network with added Gaussian noise, $\mathbf{x} \sim \mathcal{N}(f(\mathbf{z};\theta),\sigma^2)$, where f is a neural network with parameters θ. The prior over latent space is usually chosen to be a spherical Gaussian, but in principle a different distribution can be chosen if it satisfies some constraints [15]. This characteristic motivates our choice of this model as the latent prior provides us with a "handle" for incorporating structure over the latent space. Given the capability of formulating a prior over human perceptual similarity we can now include such information.

Exact inference over latent variables in the VAE is intractable but it is feasible to bound the marginal likelihood by a surrogate model. In specific, learning can be approached using a variational approach where we minimize a divergence measure between the exact and an the approximate posterior over the latent space. The approach suggested in the VAE is to formulate the approximate distribution as a deterministic function of the observed data, in specific the VAE framework uses a neural network to describe the parameters of the approximate posterior, e.g. in case of a Gaussian posterior mean and variance, $p(\mathbf{z}|\mathbf{x}) \approx q(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(g(\mathbf{x};\phi))$, where g is a neural network. The model is trained jointly (both generating and inference network) by maximizing the evidence lower bound (ELBO) on the whole data set

$$\log p(\mathbf{X}) \geq \mathscr{L}(\mathbf{X}) = E_{\mathbf{Z}\sim q(\mathbf{Z}|\mathbf{X})}[\log p(\mathbf{X}|\mathbf{Z})] - KL(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z}))$$

via backpropagation. The name of the model is therefore based on the fact that it has inference and generating networks similar to encoding and decoding networks in classic auto-encoders. This approach of representing latent representations have also been used in other types of work [20] and for GP-LVMs [19], [4].

Due to the flexibility of the model there has been several extensions proposed. The adversarial auto-encoder is using an additional adversarial network rather then a Kullback-Leibler divergence to incorporate a prior on the latent space [21]. Importance weighted auto-encoder modifies the model objective to get a tighter lower bound [2]. Ladder variational auto-encoder uses an improved inference mechanism for models with more than one layer of latent random variables [25]. Another recent line of research that the VAE can benefit from is addressing the problem of using simple posterior distributions for variational inference by specifying a more flexible, complex and scalable approximate posterior distributions using normalizing flow and inverse autoregressive flow [14], [23]. In contrast, Higgins et al. [9] are advocating for the importance of disentanglement and modify the VAE loss to force the approximated posterior to be closer to prior by putting much more weight on the KL-divergence term.

Other related work includes metric learning. [12] and [5] focus on metric learning for the purpose of finding the most similar facial expression. We, in contrast, are learning a generative model where the additional topological prior forces Euclidean distances in the latent space to reflect the closeness of facial expressions as perceived by humans.

III. METHOD

The methodology is built on VAE, that has been chosen as a primary framework due to its elegant construction combining Bayesian approach (allowing priors) with neural networks (allowing fast and easy inference). We present two extensions to VAE, *M1* and *M2*.

A. M1: VAE with neutral facial expressions

The main problem of the standard auto-encoder model for this application is that the whole face is generated from the latent space, i.e. latent representation contains information not only about facial expression (of interest here), but also about individuality features (nose shape, eyes, etc.).

To eliminate the effect of individual facial characteristics we propose using a neutral facial expression of a person as additional input to the model and therefore modeling only the transformation from this individual's neutral face to another expression of the same person.

The corresponding probabilistic graphical model is shown in Figure 1, where solid lines show the generative process and dashed lines show the inference process. There $\mathbf{z}$ denotes the hidden variable of dimensionality K, $(\mathbf{X},\mathbf{Y}) = \{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\}_{i=1}^{N}$ is a data set of facial expressions, where $\mathbf{y}^{(i)}$ is a neutral face of a person and $\mathbf{x}^{(i)}$ is any facial expression of the same person. The model factorizes as follows:

$$p_\theta(\mathbf{x},\mathbf{z}|\mathbf{y}) = p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z})p(\mathbf{z}) \tag{1}$$

The prior over the latent variable is the same isotropic Gaussian as before $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I})$, but the likelihood now also depends on the neutral face:

$$p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z}) = \mathcal{N}(\mathbf{x}|f(\mathbf{y},\mathbf{z},\theta),\sigma^2\mathbf{I}) \tag{2}$$

where $f(\mathbf{y},\mathbf{z},\theta)$ is a neural network.

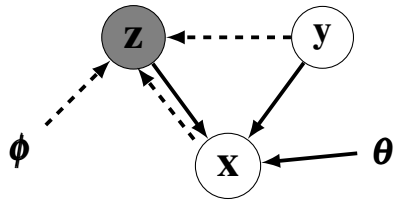


Fig. 1: The graphical representation of *M1*. Solid lines show generative process, dashed lines inference process.

From the PGM (Figure 1) we can see that $\mathbf{y}$ and $\mathbf{z}$ are connected through a "V-structure" and therefore not independent given $\mathbf{z}$. The posterior distribution over the latent variable $p_\theta(\mathbf{z}|\mathbf{x},\mathbf{y})$ is intractable. We approximate it with a distribution, which now also depends on the neutral face $\mathbf{y}$:

$$q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x},\mathbf{y},\phi),\boldsymbol{\sigma}^2(\mathbf{x},\mathbf{y},\phi)\mathbf{I}) \tag{3}$$

where mean and variance functions are neural networks.

The only difference from standard VAE in the computational graph for this model is an additional input to both the reconstruction network and the generative network.

For this model, the evidence lower bound (ELBO) on the whole data set has the following form:

$$\begin{aligned}\mathscr{L}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{N}\Bigg[E_{z\sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})}\Big[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{y}^{(i)},\mathbf{z})\Big] \\ - KL\Big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})||p(\mathbf{z})\Big)\Bigg]\end{aligned} \tag{4}$$

The expectation is approximated with L Monte Carlo samples. The Stochastic Gradient Variational Bayes (SGVB, [15]) estimator of the loss function is based on a mini-batch of size B:

$$\begin{aligned}Loss(\mathbf{X},\mathbf{Y}) &\approx -\widetilde{\mathscr{L}}(\mathbf{X}^B,\mathbf{Y}^B) = -\frac{N}{B}\sum_{i=1}^{B}\widetilde{\mathscr{L}}(\mathbf{x}^{(i)},\mathbf{y}^{(i)}) \\ &= \frac{N}{B}\sum_{i=1}^{B}\Bigg[-\frac{1}{L}\sum_{l=1}^{L}\Big[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{y}^{(i)},\mathbf{z}^{(i,l)})\Big] \\ &\quad + KL\Big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})||p(\mathbf{z})\Big)\Bigg] \\ \text{where } \mathbf{z}^{(i,l)} &= \boldsymbol{\mu}(\mathbf{x}^{(i)},\mathbf{y}^{(i)},\phi) + \boldsymbol{\sigma}(\mathbf{x}^{(i)},\mathbf{y}^{(i)},\phi)\cdot\boldsymbol{\varepsilon}^{(i,l)}, \\ &\quad \boldsymbol{\varepsilon}^{(i,l)} \sim \mathcal{N}(\mathbf{0},\mathbf{I}).\end{aligned} \tag{5}$$

Here we use the reparameterization trick of Kingma and Welling to allow backpropagation. The model can be trained using any SGD algorithm.

B. M2: VAE with neutral facial expressions and topology

Naturally in many applications, there is some prior knowledge about the topology (e.g. smoothness or similarity preservation), and it is clearly beneficial to incorporate this knowledge into the model. Urtasun et al. [27] do this for the human motion modeling with Gaussian processes by putting explicit constraints on the embedding. More specifically they formulate the prior in the form of $p(\mathbf{Z}) \propto e^{-\frac{1}{\gamma}\Phi(\mathbf{Z})}$, where $\Phi(\mathbf{Z})$ is an energy function modeling specific topological constraints and γ is a global scaling of the prior.

In the neural network approach to representation learning it is also possible to impose such constraints by modifying the objective (adding penalizing term for violation of the constraints), but often the most challenging part is to modify a network with minimum loss in computational capacity. For the case of partial similarity constraints in the form of triplets (as in this paper), Hoffer and Ailon [10] propose a "triplet network", that has three parts of the network sharing weights and then an additional layer for distance comparison.

Our hypothesis is that imposing human-like topological constraints on the latent space will result in learning a better representation. We combine ideas of Urtasun et al. [27] and Hoffer and Ailon [10] and propose a new model based on the previous one with a modified prior on the latent space.

Topological constraints on the latent space are represented as a set of T triplets, where each triplet consists of a reference face and two other faces with one of those faces marked as being more similar to the reference one than the other based on human perception. The resulting triplet data set is

$$\mathbf{S} = \left\{(\mathbf{s}^{(t,ref)},\mathbf{s}^{(t,+)},\mathbf{s}^{(t,-)}) : d(\mathbf{h}^{(s_t^{ref})},\mathbf{h}^{(s_t^{+})}) \le d(\mathbf{h}^{(s_t^{ref})},\mathbf{h}^{(s_t^{-})})\right\}_{t=1}^{T}$$

where each of $\mathbf{s}^{(t,ref)},\mathbf{s}^{(t,+)},\mathbf{s}^{(t,-)}$ corresponds to some index $i \in \{1,..N\}$ in the original data set of facial expressions, d is the Euclidean distance and $\mathbf{h}^{(i)}$ is some (human-like) representation of the facial expression $\mathbf{x}^{(i)}$.

To fulfill these topological constraints over triplets on the latent representation $\mathbf{z}$ we want to minimize:

$$\Phi(\mathbf{Z},\mathbf{S}) = \sum_{i=1}^{T} max\big(0; d(\mathbf{z}^{(s_t^{ref})},\mathbf{z}^{(s_t^{+})}) - d(\mathbf{z}^{(s_t^{ref})},\mathbf{z}^{(s_t^{-})})\big)$$

Instead of using $f(x) = max(0;x)$ to penalize incorrect distances, we will use its smooth approximation $f(x) = ln(1+e^x)$ called "softplus" to force a small margin on the distance difference. For additional flexibility, each triplet can have a weight w_t (e.g. corresponding to a reliability level for each triplet if the triplets are collected from people).

$$\Phi(\mathbf{Z},\mathbf{S}) = \sum_{i=1}^{T} w_t \ln\big(1+\exp(d(\mathbf{z}^{(s_t^{ref})},\mathbf{z}^{(s_t^{+})}) - d(\mathbf{z}^{(s_t^{ref})},\mathbf{z}^{(s_t^{-})}))\big)$$

This can be interpreted as a prior [27] $p_T(\mathbf{Z}|\mathbf{S}) \propto e^{-\frac{1}{\gamma}\Phi(\mathbf{Z},\mathbf{S})}$ that forces to fulfill as much constraints as possible, where γ is a "topological variance". The smaller the value, the larger the penalty for an incorrect topology.

This topological prior can be factorized over triplets:

$$\begin{aligned}p_T(\mathbf{Z}|\mathbf{S}) &= \prod_{t=1}^{T} p_T\left(\mathbf{z}^{(s_t^{ref})},\mathbf{z}^{(s_t^{+})},\mathbf{z}^{(s_t^{+})}|\mathbf{s}^{(t)}\right) \\ &\propto \prod_{t=1}^{T}\exp\Big(-\frac{1}{\gamma}\Phi(\mathbf{z}^{(s_t^{ref})},\mathbf{z}^{(s_t^{+})},\mathbf{z}^{(s_t^{+})})\Big)\end{aligned} \tag{6}$$

The topological prior on the latent variable $\mathbf{z}$ can be added to the standard Gaussian prior we used in the previous model:

$$\begin{aligned}p(\mathbf{Z}) &= p_T(\mathbf{Z}|\mathbf{S})p_{\mathcal{N}}(\mathbf{Z}) \\ &= \prod_{t=1}^{T} p_T\left(\mathbf{z}^{(s_t^{ref})},\mathbf{z}^{(s_t^{+})},\mathbf{z}^{(s_t^{+})}|\mathbf{s}^{(t)}\right)\prod_{i=1}^{N}\mathcal{N}(\mathbf{z}^{(i)}|\mathbf{0},\mathbf{I})\end{aligned} \tag{7}$$

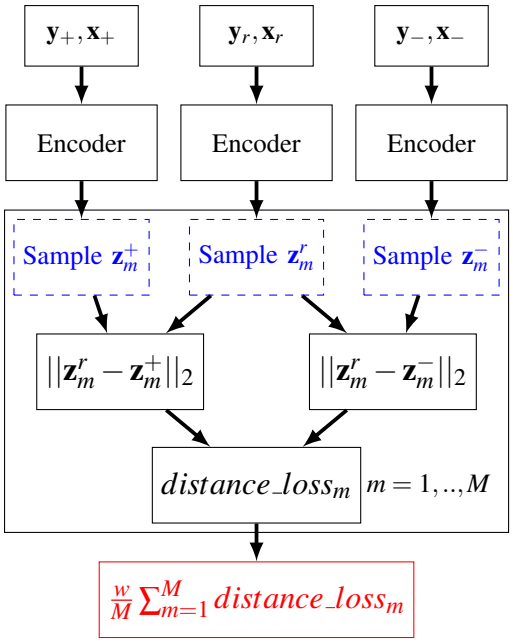


Fig. 2: Architecture of the "triplet" part of *M2*.

Given that the prior is now not factorizable over the data points, we derive ELBO on the whole data set:

$$\begin{aligned}\log p_\theta(\mathbf{X}|\mathbf{Y}) &= \log \int p_\theta(\mathbf{X},\mathbf{Z}|\mathbf{Y}) d\mathbf{Z} \\ &= \log \int q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y}) \frac{p_\theta(\mathbf{X},\mathbf{Z}|\mathbf{Y})}{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y})} d\mathbf{Z} \\ &\geq E_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\left[\log \frac{p_\theta(\mathbf{X},\mathbf{Z}|\mathbf{Y})}{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\right] = \mathscr{L}(\mathbf{X},\mathbf{Y})\end{aligned} \tag{8}$$

The last line was derived by applying Jensen's inequality.

$$\begin{aligned}&\mathscr{L}(\mathbf{X},\mathbf{Y}) = E_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\left[\log \frac{p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{Y}) p_T(\mathbf{Z}|\mathbf{S}) p_\mathscr{N}(\mathbf{Z})}{q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\right] \\ &= E_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{Y})}\Bigg[\log \prod_{i=1}^{N} p_\theta(\mathbf{x}^{(i)}|\mathbf{z}^{(i)},\mathbf{y}^{(i)}) \\ &\qquad + \log \prod_{t=1}^{T} p_T\left(\mathbf{z}^{(s_t^{ref})}, \mathbf{z}^{(s_t^+)}, \mathbf{z}^{(s_t^+)}\right) \\ &\qquad + \log \prod_{i=1}^{N} p_\mathscr{N}(\mathbf{z}^{(i)}) - \log \prod_{i=1}^{N} q_\phi(\mathbf{z}^{(i)}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})\Bigg] \\ &= \sum_{i=1}^{N} E_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})}\left[\log p_\theta(\mathbf{x}^{(i)}|\mathbf{z},\mathbf{y}^{(i)})\right] \\ &- \sum_{i=1}^{N} KL\Big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})||p_\mathscr{N}(\mathbf{z})\Big) \\ &+ \sum_{t=1}^{T} E_{\left\{\begin{array}{ll}\mathbf{z}_\mathbf{r} & \sim q_\phi(\mathbf{z}_r|\mathbf{x}^{(s_t^{ref})},\mathbf{y}^{(s_t^{ref})}) \\ \mathbf{z}_+ & \sim q_\phi(\mathbf{z}_+|\mathbf{x}^{(s_t^+)},\mathbf{y}^{(s_t^+)}) \\ \mathbf{z}_- & \sim q_\phi(\mathbf{z}_-|\mathbf{x}^{(s_t^-)},\mathbf{y}^{(s_t^-)})\end{array}\right\}}\left[\log p_T\left(\mathbf{z}_r,\mathbf{z}_+,\mathbf{z}_-\right)\right]\end{aligned} \tag{9}$$

We further write the exact density functions and use the reparameterization trick to derive a differentiable lower bound:

$$\begin{aligned}\mathscr{L}(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{N}\Bigg[&-\frac{1}{2\sigma^2 L}\sum_{l=1}^{L}\left(\mathbf{x}^{(i)} - f(\mathbf{z}^{(i,l)},\mathbf{y}^{(i)})\right)^2 \\ &- KL\Big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})||p_\mathscr{N}(\mathbf{z})\Big)\Bigg] \\ -\frac{1}{\gamma M}\sum_{t=1}^{T} w_t \sum_{m=1}^{M} \ln\Big(1 + \exp\big(&d(\mathbf{z}^{(s_t^{ref},l)}, \mathbf{z}^{(s_t^+,l)}) \\ &- d(\mathbf{z}^{(s_t^{ref},l)}, \mathbf{z}^{(s_t^-,l)})\big)\Big)\end{aligned} \tag{10}$$

where $f(\mathbf{y},\mathbf{z})$ is the generative neural network.

To maximize the evidence lower bound with batch-wise stochastic gradient descend, the objective is reformulated. We use separate data batches and triplet batches. The approximation of the loss function based on a data batch of size B and a triplet batch of size V can be computed as follows:

$$\begin{aligned}Loss(\mathbf{X},\mathbf{Y},\mathbf{S}) &\approx Loss(\mathbf{X}^B,\mathbf{Y}^B,\mathbf{S}^V) = -\widetilde{\mathscr{L}}(\mathbf{X}^B,\mathbf{Y}^B,\mathbf{S}^V) \\ &= \frac{N}{B}\sum_{i=1}^{B}\Bigg[\frac{1}{2\sigma^2 L}\sum_{l=1}^{L}\left(\mathbf{x}^{(i)} - f(\mathbf{z}^{(i,l)},\mathbf{y}^{(i)})\right)^2 \\ &\qquad + KL\Big(q_\phi(\mathbf{z}|\mathbf{x}^{(i)},\mathbf{y}^{(i)})||p_\mathscr{N}(\mathbf{z})\Big)\Bigg] \\ &+ \frac{T}{V}\frac{1}{\gamma M}\sum_{t=1}^{V} w_t \sum_{m=1}^{M} \ln\Big(1 + \exp\big(d(\mathbf{z}^{(s_t^{ref},m)}, \mathbf{z}^{(s_t^+,m)}) \\ &\qquad - d(\mathbf{z}^{(s_t^{ref},m)}, \mathbf{z}^{(s_t^-,m)})\big)\Big)\end{aligned} \tag{11}$$

Note, that in fact the batch-wise objective also depends on the data corresponding to the triplet batch.

The architecture of the computational graph for this model is a combination of the one used in the *M1* with neutral faces and the part shown in Figure 2, which is responsible for the triplet term in the loss function and was inspired by the "triplet network" of Hoffer and Ailon [10]. All the parts of the loss function are shown in red in the pictures.

IV. DATA

Two data sets will be used in this work, BU-3DFE [29] containing posed static facial expressions, and BP4D-Spontaneous [30] containing dynamic spontaneous expressions. Our models also require neutral expressions as additional inputs. This is described in Section IV-A.

M2 imposes a human-like similarity metric on the latent space and needs triplet data as constraints. Triplet data set was collected using a crowd-sourcing service (Amazon Mechanical Turk)[1], as described in Section IV-B.

A. Data sets

The data format in both data sets used in the experiments is 83 3D points for each face (facial landmarks).

[1]The triplet data set is available upon request to the main author. Please note, that it only contains frame IDs for comparison; the main BP4D-Spontaneous database should be acquired separately.

1) Static posed data set with stereotypical facial expressions: This data set consists of 100 individuals, each posed with 4 degrees of stereotypical facial expressions (”angry”, “disgust”, “sad”, “happy”, “surprised”, “fear”) and a “neutral” one [29]. Each person has 25 data points and the total size of the dataset is $2,500$ facial expressions with dimensionality 249. This data set is quite small and does not have enough variability, but has labels which can be useful for evaluation. The data set is split person-wise into training, validation and test as 80/10/10.

2) Dynamic spontaneous data set: For this data set, 41 individuals were asked to participate in 8 tasks, each task has an intended emotion (e.g. ”sing a song” for “embarrassment”) [30]. This data set is much larger than the first one and has more variability. The total number of data points is $367,492$. We split the data set into 3 subsets person-wise with 25/8/8 individuals for training, validation and test subsets respectively. In this dataset, for each video 20 seconds were manually annotated with action units (AU) by specialists.

B. Triplets

M2 uses our hypothesis that incorporating knowledge about human perception of facial expression will help to learn a better latent representation. In order to formalize this knowledge we collect triplet data, where people choose which facial expressions are more similar.

Artificial triplets: As a preparation for the collection of real triplets and a proof of concept, we conduct experiments on the artificial triplets on the posed data set. For the training data subset, 8000 triplets were generated. For the validation and test sets, 1000 triplets per set were generated. Every data point is present in at least one triplet. The rules for generating triplets from the true labels we used are:

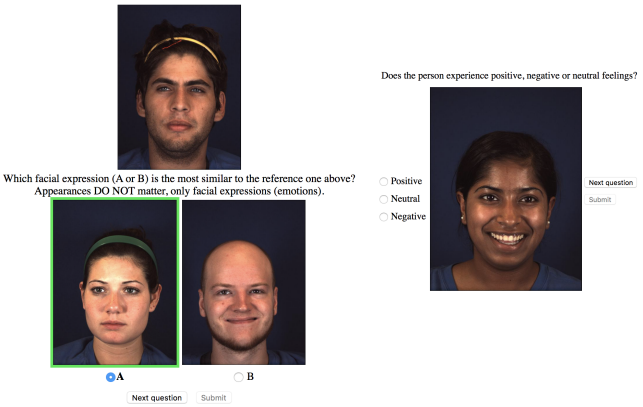
- Expressions from the same class are closer than from different classes.
- Within the same class: the closer the degree of expression the smaller the distance.
- A neutral expression is closer than a different class.

Triplet collection with Amazon MT: Triplets are collected for the spontaneous data set using Amazon MT. Participants are asked to chose which of the 2 facial expressions looks more similar to the reference one (Figure 3 (a)).

A subset of data to collect triplets on should have a reasonable size and ideally cover the true latent space evenly. To select this subset we exploit the fact that the most expressive part of each sequence was annotated with AUs and only use unique expressions for each person in terms of labeled AUs. The number of images is further minimized by removing blinking based on the distances between eyelids.

For each subset of data (train, validation, test), separate sets of triplets were generated in a way that each image is in 6 triplets.

Given that the data we are collecting are opinions, and often there is no obviously correct answer, we collect 5 answers from different people in order to get statistics for the trustworthiness of each triplet. The difference in the number of votes for each answer (5, 3 or 1) give a reliability statistic that we use as a weighting coefficient in the *M2*. The distribution of differences between two possible answers in the collected data set is shown in Table I.



(a) Triplets (b) Labels

Fig. 3: Data collection with Amazon Mechanical Turk

C. Labels

In the posed data set, all data points are labeled with stereotypical facial expressions by design. For the spontaneous data set only intended emotions are given which is not accurate enough for a classification task. To test the latent representation of that data set we use Amazon MT to collect labels for $2,500$ data points. Each participant is asked to access whether a given facial expression corresponds to neutral, positive or negative emotions (Figure 3 (b)). For each data point 5 answers were collected. Only those where the number of votes was at least 3 for positive or negative are used later. The total number of labeled points is $1,401$ ($1,040$ “positive” and 361 “negative”).

V. EXPERIMENTS

A. Preprocessing

We conducted experiments separately on each data set due to incompatibility of landmark placement. For each data point, the point cloud was rotated to neutral pose using pose estimates given in data sets, centered with respect to the point between eyes and scaled so the the distance between the eyes is 1. The position of each eye was computed as a center of points corresponding to the lower eyelid. For the training procedure the data was further centered using the mean of the training portion of the corresponding data along each dimension for numerical stability of the neural network.

To eliminate the effect of appearance and only encode facial expressions in the latent representation, the neutral

TABLE I: Agreement statistics for collected triplets

Data subset	**Difference**			**Total**
	5	**3**	**1**	
Train	2,426	2,364	2,206	6,996
Validation	652	686	643	1,981
Test	728	767	719	2,214
Whole data set	**3,806**	**3,817**	**3,568**	**11,191**

facial expression is needed for each individual. In the posed data set each person has a corresponding neutral face. The spontaneous dataset contains no specifically marked neutral facial expressions. We decided to leverage the action units labeling available in the dataset to select one neutral face for each person. For each person we selected time frames which have AU labels and all AU are marked as non present. We further select a single face with the minimal sum of distances to the other neutral faces for each individual.

B. Evaluation

To evaluate quality of the learned latent representation we use classification as the target task due to its interpretability and availability of the labels in the data sets. Labels in the posed data are emotions and the task is a multi-class classification. In the spontaneous data set, collected labels are "positive" and "negative" emotions and the task is a binary classification.

Another evaluation technique that will be used is the number of satisfied triplets. This measure will reflect the topological coherency of the representation space.

Baseline: Our methods fall into the category of non-linear dimensionality reduction techniques. One natural baseline in representation learning is the original data space with no transformation. As a comparison, linear PCA and standard VAE is also used. The method of choice for the classification tasks is SVM with linear kernel.

Classification of stereotypical facial expressions on the posed data set: The posed data set contains 4 degrees of all 6 stereotypical facial expressions and a neutral expression. The task of classifying these facial expressions is a multi-class classification problem and is implemented as a set of one-vs-rest SVM classifiers, one for each class. The standard accuracy is used as a performance measure.

Classification of positive and negative emotions on spontaneous data set: $1,401$ data point in the spontaneous data set is labeled using crowd-sourcing. We use them as a binary classification task with linear SVM as the classifier of choice.

Distance preservation: Given the hypothesis about topology another test we will use is the number of satisfied triplets. It indicates the degree of similarity of the learned latent space topology and that of an assumed internal human representation, which can be useful for a number of application. This evaluation will be conducted on both data sets ("artificial" triplets will be used on the posed data set).

A satisfied triplet is defined as follows:

$$sat((i_{ref}, i_+, i_-)) = I(||z_{i_{ref}} - z_{i_+}||_2 \leq ||z_{i_{ref}} - z_{i_-}||_2) \quad (12)$$

The collected triplet data for the spontaneous data set have 3 levels of "confidence". Since the triplets for the posed data set were generated according to the reasonable rules, they all considered to have a 100% reliability level (weight 1).

C. Training

Architecture: Both the reconstruction and the generative parts of the model are approximated with neural networks. All the layers are fully connected with exponential linear unit non-linearity (ELU, [3]). Different layer configurations are used during training. In the standard VAE, the encoding and decoding parts are symmetrical, in the models *M1* and *M2*, they are asymmetrical as the additional neutral face is added to both encoder and decoder input.

Adam algorithm is used for optimization [13]. All models are trained for $100,000$ iterations. Each $2,500$ iterations the model is evaluated by computing ELBO on the validation data set. An iteration with the highest validation ELBO is considered the final model. Fixed parameters are the learning rate of $1e-4$, dropout 0.9, sample size $L = 3$ and triplet sample size $M = 10$, batch size $B = 200/250$ (posed, spontaneous) and triplet batch size $V = 10$.

Annealing of the divergence term: It is typical for the VAE-model to "overregularize", more specifically, to turn off some latent dimensions early, so that the model does not use the full allowed capacity. The most common way to improve the learning is to use a modified objective function [28]:

$$\mathscr{L}(\mathbf{X}) = -E_{q_\phi(Z|X)}\Big[\log p_\theta(X|Z)\Big] + \beta \cdot KL(q_\phi(Z|X)||p(Z)) \quad (13)$$

and slowly increase β from 0 to 1 over a number of iterations. When $\beta = 0$ the objective is the ML estimation and equivalent to that of the standard auto-encoder model. $\beta = 1$ corresponds to the normal VAE objective. In all experiments β is increased linearly with the number of iterations.

M2 has an additional parameter, the topological variance. The smaller this parameter the higher the penalty for not satisfying the triplets. There are two possible ways to modify the objective function to perform the annealing in this case, either only anneal the KL-divergence term or anneal the whole prior. We try both to compare.

D. Results on the static posed data set

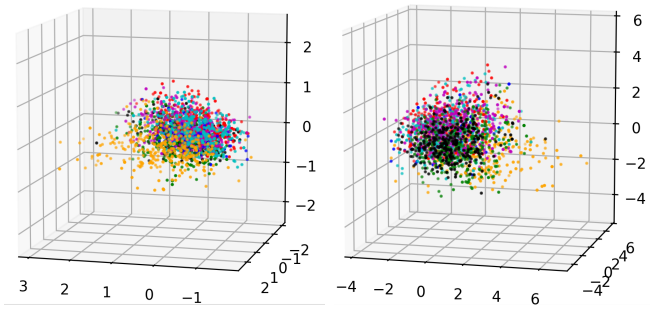
We vary the layer architecture, reconstruction variance and the number of annealing iterations for VAE and *M1*, and also topological variance for *M2*.

The best model on validation data set is *M2* with annealing and only the KL-divergence term with the following configuration: encoding [498, 480, 240, 120, 60, 30], decoding [279, 300, 400, 300, 249], reconstruction variance 0.001, topological variance 0.04, annealing over 100 000 iterations.

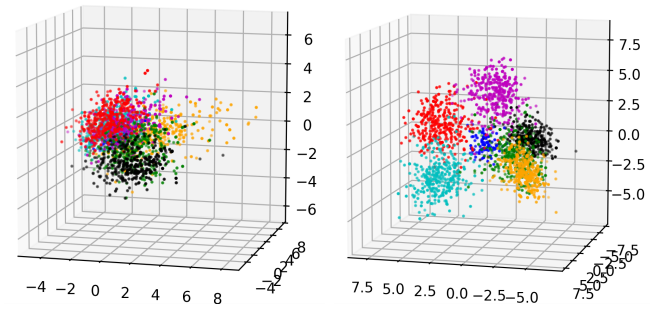
As we can see in the comparison in Table II, *M2* outperforms both baselines (the original data and linear PCA) and also standard VAE. Both, classification accuracy and the number of satisfied triplets, improved with adding neutral face and incorporating topology. Figure 4 shows latent space for different representations projected onto 3 principal components. We can see that *M2* provides better linear separation between classes compared to other representations.

E. Results on spontaneous data set

Our models are also trained on the spontaneous data set. The main advantage of these spontaneous data is the more natural variations in facial expressions as opposed to the stereotypical and static expressions in the posed data set. For this data set we collected triplet data and labels using crowd-sourcing.

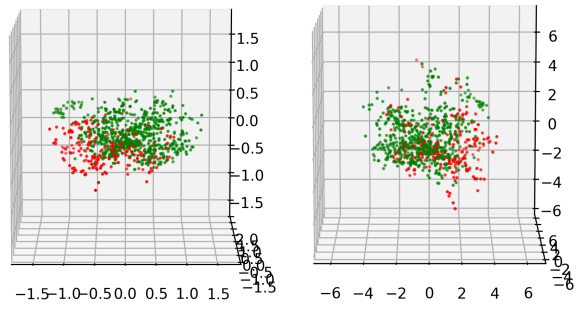


(a) Original (b) VAE

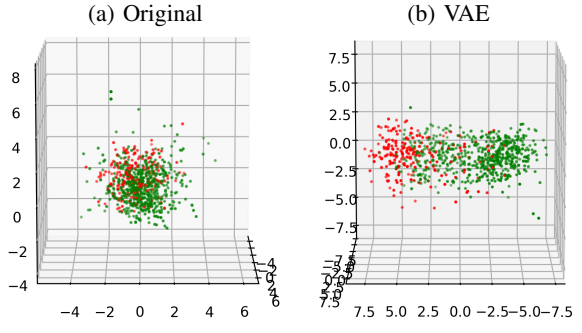


(c) *M1* (d) *M2*

Fig. 4: Visualization of different representations on the static posed data set in 3D using linear PCA. Facial expression color codes: angry, disgust, fear, happy, neutral, sad, surprised.



(a) Original (b) VAE



(c) *M1* (d) *M2*

Fig. 5: Visualization of different representations on the spontaneous data set in 3D using linear PCA. Emotion color codes: negative, positive.

Performance comparison with the baseline models on the test part of the spontaneous data set is given in the Table III. Classification accuracy on this task is quite high even for a linear dimensionality reduction, and very similar for all representations (original, PCA, *M1*, *M2*). As we can see from the 3D projection of the latent space (Figure 5), the classes are already reasonably separated even in the original representation (Figure 5(a)), so the task is not very challenging. Lower dimensional representations provide comparable classification accuracy, but *M2* (Figure 5(d)) also pulls classes apart in the latent space a bit more as opposed to VAE and *M1* which employ only Gaussian prior.

The triplet coherency for *M2* is also much higher than for the baseline models for the high and medium confidence triplets (Table II), which indicates that the latent space of *M2*, to a larger degree than the baselines, is structured similarly to how humans perceive changes in facial expression.

VI. CONCLUSIONS AND FUTURE WORK

We developed a methodology for learning latent representation of facial expressions consistent with human perception of similarity. The methodology was iteratively built on the basis of variational auto-encoder. To eliminate the individual features classic VAE was modified to include neutral faces, so that the model can focus on learning only the deviation of a facial expression from the corresponding neutral one. We incorporated topological constraints as an additional component in the prior distribution of the latent variable.

The models were trained and tested on two data sets. While the data formats in the data sets are the same, the size, the labeling, and the variation in the data are different. On the posed data we saw that the standard VAE was performing worse than the baselines, but adding a neutral face increased classification accuracy up to the baseline levels. Including the topological prior helped to structure the latent space in a way that is coherent with a human similarity assessment and raised the classification accuracy above the baseline.

The latent representation of the posed data set showed lin-

TABLE II: Results for static posed dataset

Dimensionality reduction	**Dim**	**Accuracy**	**Triplets**
None	249	0.72	0.593
PCA	30	0.676	0.592
VAE	30	0.64	0.592
M1	30	0.72	0.67
M2	30	**0.736**	**0.798**

TABLE III: Results for spontaneous data set

Dimensionality reduction	**Dim**	**Accuracy**	**Triplets**		
			5	**3**	**1**
None	249	0.842	0.747	0.615	0.549
PCA	30	0.824	0.743	0.608	0.549
VAE	30	0.838	0.706	0.593	0.494
M1	30	**0.853**	0.629	0.575	0.512
M2	30	0.835	**0.830**	**0.701**	0.559

ear separation between classes of stereotypical expressions. Even thought the triplets for that data set were artificially generated based on the labels, the assumptions were mild.

On the spontaneous data set the labels were collected using crowd-sourcing. On this data set the classification performance of our models was very similar to the baseline. This can be explained by the facts that, firstly, label quality is not very high, and secondly, the task of classifying positive/negative emotions given the labels is not very complex given good classification even on the original data. Most likely the labels highly correlate with obvious features in the original data (e.g. a wide smile).

On both data sets, topological prior increased the number of satisfied triplets on the test part of the data indicating that the model does not just overfit the training triplets, but rather uses them to structure the latent space in a way consistent with human similarity assessments.

One of the benefits of the generative probabilistic model is the ability to explore the role of each dimension through generations of new samples. It it possible to identify dimensions responsible for very a specific facial action, such as closing and opening the mouth, smiling, blinking, etc.

The learned representations perform better of the same on classification tasks, are structured in the latent space consistently with human perception, and are interpretable through the means of the generative process.

A. Future work

So far we only worked with static facial expressions (even when using the dynamic data set) while they are dynamic by nature. Therefore incorporating temporal dynamics in the model is the most obvious next step that can help learning smooth trajectories in the latent space.

Another option to explore is different prior on the latent space. For example, each latent dimension could have an independent Beta distribution as a prior to mimic action unit activations, but with continuous features.

This work can also be improved by testing the learned latent representation on more evaluation tasks.

Though we worked with 3D facial landmarks, it is also possible to use different type of input data and modify the networks accordingly, e.g. we could use images as input and CNN as encoder and decoder.

VII. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of reviewers' comments. This work is partially funded by Stiftelsen för Strategisk Forskning.

REFERENCES

[1] M. Argyle. *Bodily Communication*. Routledge, 1988.

[2] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[3] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016.

[4] Z. Dai, A. Damianou, J. González, and N. Lawrence. Variational auto-encoded deep gaussian processes. *arXiv preprint arXiv:1511.06455*, 2015.

[5] A. Dhall, A. Asthana, and R. Goecke. A ssim-based approach for finding similar facial expressions. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 815–820. IEEE, 2011.

[6] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[7] P. Ekman and W. V. Friesen. Facial action coding system. 1977.

[8] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic. Variational gaussian process auto-encoder for ordinal prediction of facial action units. In *Asian Conference on Computer Vision*, pages 154–170. Springer, 2016.

[9] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.

[10] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[11] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, Sept. 1933.

[12] I. Kemelmacher-Shlizerman, A. Sankar, E. Shechtman, and S. M. Seitz. Being john malkovich. In *European Conference on Computer Vision*, pages 341–353. Springer, 2010.

[13] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[14] D. P. Kingma, T. Salimans, and M. Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

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

[16] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

[17] M. L. Knapp, J. A. Hall, and T. G. Horgan. *Nonverbal Communication in Human Interaction*. Wadsworth, 2013.

[18] N. D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816, 2005.

[19] N. D. Lawrence and J. Quiñonero Candela. Local distance preservation in the gp-lvm through back constraints. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 513–520, New York, NY, USA, 2006. ACM.

[20] D. Lowe and M. E. Tipping. NeuroScale: Novel topographic feature extraction using RBF networks. *Advances in Neural Information Processing Systems*, pages 543–549, 1997.

[21] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[22] C. Pramerdorfer and M. Kampel. Facial expression recognition using convolutional neural networks: State of the art. *arXiv preprint arXiv:1612.02903*, 2016.

[23] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[24] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697, 2012.

[25] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 3738–3746, 2016.

[26] M. K. Titsias and N. D. Lawrence. Bayesian gaussian process latent variable model. In *International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

[27] R. Urtasun, D. J. Fleet, A. Geiger, J. Popović, T. J. Darrell, and N. D. Lawrence. Topologically-constrained latent variable models. In *Proceedings of the 25th international conference on Machine learning*, pages 1080–1087. ACM, 2008.

[28] S. Yeung, A. Kannan, Y. Dauphin, and L. Fei-Fei. Tackling over-pruning in variational autoencoders. *arXiv preprint arXiv:1706.03643*, 2017.

[29] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[30] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.