Factorized Variational Autoencoders for Modeling Audience Reactions to Movies

Zhiwei Deng[1], Rajitha Navarathna[2], Peter Carr[2], Stephan Mandt[2], Yisong Yue[3],
Iain Matthews[2], Greg Mori[1]
[1]Simon Fraser University [2]Disney Research [3]Caltech

Abstract

Matrix and tensor factorization methods are often used for finding underlying low-dimensional patterns from noisy data. In this paper, we study non-linear tensor factorization methods based on deep variational autoencoders. Our approach is well-suited for settings where the relationship between the latent representation to be learned and the raw data representation is highly complex. We apply our approach to a large dataset of facial expressions of movie-watching audiences (over 16 million faces). Our experiments show that compared to conventional linear factorization methods, our method achieves better reconstruction of the data, and further discovers interpretable latent factors.

1. Introduction

The desire to uncover compact yet expressive latent representations from raw data is pervasive across scientific and engineering disciplines. The benefits of having such representations are numerous, including de-noising, imputing missing values, reducing label complexity for supervised learning, and interpretable latent spaces for better insight into the phenomena hidden within the data. However, for complex domains, conventional approaches often fail to learn a latent representation that is both semantically expressive and compact.

In this paper, we focus on learning such a representation for a large dataset of facial expressions extracted from movie-watching audiences. Intuitively, we expect audience members to have correlated reactions, since each movie has been specifically crafted to elicit a desired response [34]. Thus, one can view audience analysis as a form of collaborative filtering, which has been popularized for modeling recommender systems (e.g., the Netflix challenge). For instance, we can assume that there are underlying exemplar facial expressions which form a basis to reconstruct the observed reactions of each audience member. The most common approaches for collaborative filtering are factorization techniques [25], which we build upon in our work.

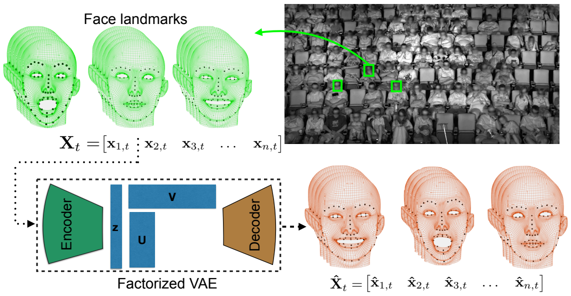
Face landmarks
$\mathbf{X}_t = [\mathbf{x}_{1,t}\ \mathbf{x}_{2,t}\ \mathbf{x}_{3,t}\ \dots\ \mathbf{x}_{n,t}]$
Encoder
z
V
U
Decoder
Factorized VAE
$\hat{\mathbf{X}}_t = [\hat{\mathbf{x}}_{1,t}\ \hat{\mathbf{x}}_{2,t}\ \hat{\mathbf{x}}_{3,t}\ \dots\ \hat{\mathbf{x}}_{n,t}]$

Figure 1. **Factorized VAE**. Facial landmarks are detected on audience members for the duration of a movie. Tensor factorization assumes a linear decomposition. We propose a non-linear version which learns a variational autoencoder such that the latent space factorizes linearly.

Conventional factorization approaches rely on the data decomposing linearly. For complex processes such as time-varying facial expressions, such an assumption is not appropriate. The typical way to address this limitation is to engineer feature representations that lead to a linearly decomposable structure, which requires significant manual trial-and-error and/or domain expertise.

In this paper, we formulate a new non-linear variant of tensor factorization using variational autoencoders, which we call *factorized variational autoencoders* (FVAE), to learn a factorized non-linear embedding of the data. Our approach enjoys several attractive properties, including:

- FVAEs inherit the flexibility of modern deep learning to discover complex non-linear relationships between a latent representation and the raw data.
- FVAEs inherit the generalization power of tensor factorization that exploits known assumptions of the data (e.g., the signal factorizes across individuals and time).
- FVAEs enable compact and accurate modeling of complex data such as audience facial expressions that was

not possible with either variational autoencoders or tensor factorization alone.

- FVAEs leverage existing optimization techniques for variational inference and matrix factorization, and are fully trainable end-to-end.

We demonstrate the effectiveness of our approach on a audience facial expression dataset collected from an instrumented 400 seat theatre that hosted multiple viewings of multiple movies over a twelve month period. We conduct extensive empirical analyses, including both quantitative (e.g., predicting missing data) and qualitative (e.g., interpreting the learned latent space). Our findings are:

- FVAEs can learn semantically meaningful representations that capture a range of audience facial expressions, which are easy to interpret and visualize.
- FVAEs are much more compact than conventional baselines, and require fewer latent dimensions to achieve superior reconstruction of the data.
- FVAEs can accurately predict missing values, and in the case of our audience data, can anticipate the facial expressions of an individual for the entire movie based only on observations at the beginning of the movie. In fact, using only the initial 5% of data as observations, FVAEs can reconstruct an audience member's facial reactions for the entire movie more accurately than conventional baselines using all the data.

2. Related Work

Audience Analysis. The study of facial expressions is an area of great interest in computer vision and computer graphics, and is the basis for applications in affective computing and facial animation. Within this field, self-reports [1, 42] have become the standard approach to understand audience sentiment for long-term stimuli such as movies. This approach is not only subjective and labor intensive but also loses much of the fine-grained temporal detail, as it requires a person to consciously think about and document what they are watching, which means they may miss important parts of the movie. Although subjects could be instrumented with a myriad of wearable sensors such as heart rate or galvanic skin response [5, 11, 29, 37], vision-based approaches are ideal, as they can be applied unobtrusively and allow viewers to watch the stimuli uninhibited.

A plethora of work in automatically measuring a person's behavior using vision-based approaches has centered on recognizing an individual's facial expression [50]. Joho *et al*. [18] showed that observed facial behavior is a useful measure of engagement, and Teixerira *et al*. [44] demonstrated that smiling is a reliable feature of engagement. McDuff *et al*. [32] further demonstrated the use of smiles to gauge a test audience's reaction to advertisements, while Whitehill *et al*. [51] used facial expressions to investigate student engagement in a classroom setting. In particular, Whitehill *et al*. [51] showed that human observers reliably conclude that another person is engaged based on head pose, eyebrow position, and eye and lip motions [51]. Higher-level behaviors, such as the number of eye fixations, were used as reliable indicators for measuring engagement in advertisements [48]. However, the use of eye tracking in long-term stimuli such as feature-length movies poses a considerable challenge, not only due to the duration of the stimulus but also the distance at which it is viewed. The first attempt to automate viewer sentiment analysis over long period of time (e:g upto 2 hours) was proposed by Navarathna *et al*. [35] where, they measured the distribution of short-term correlations of audience motions to predict the overall rating of a movie.

Collaborative Filtering. Our work builds upon matrix factorization (MF), which is a common tool in recommender systems [25]. In many applications, one or both of the factorized representations are constrained to be non-negative, which leads to non-negative MF [8, 27]. Tensor factorization approaches have also been used for modeling higher-order interactions [38,54]. Probabilistic versions have been proposed in [36, 41], where the data likelihood is typically Gaussian, but has been recently generalized to Poisson [12] for sparse count data. Our likelihood function uses a neural network as part of the variational approximation, and can therefore more flexibly fit the training data. Deterministic deep hierarchical matrix factorization models have been explored in [45, 47], but these approaches do not serve our purpose of analyzing how a user reacts to a specific stimulus over time, or any interpretable latent representations. We instead use probabilistic variational autoencoders to jointly learn a latent representation of each face and a corresponding factorization across time and identity.

Variational Autoencoders. Bayesian inference for deep generative models was first proposed in [24, 39] and has since become very popular. The latent dimensions of a variational autoencoder (VAE) often lack interpretability, so Higgins *et al*. [14] proposed to disentangle the latent dimensions by upweighting the KullbackLeibler (KL) divergence between prior and variational distribution which relates to variational tempering [31]. Building hierarchical Bayesian models with VAEs has until recently been elusive – one recent example is image generation with multiple layers [16]. Johnson *et al*. [17] combine hierarchical models with neural networks for structured representation but do not discuss factorized priors but rather other types of priors such as dynamical systems. Other recent extensions include incorporating ladder connections [43], recurrent variants for modeling sequence data [6], incorporating attention mechanisms [13], and adversarial training [30]. In contrast, we focus on learning VAEs that have a factorized representation in order to further compress the embedding space and enhance generalization and interpretability.

3. Methods

In this section we discuss the different models that we used to analyze our data. For each movie, we observe N audience members for a duration of T frames. For each audience member i at time t we record a $D = 136$ dimensional vector $\mathbf{x}_{it} \in \mathbb{R}^{136}$ representing the (x, y) locations of 68 facial landmarks. We expect audience members to react in similar but unknown ways, and therefore investigate methods for identifying patterns in the $N \times T \times D$ tensor $\mathbf{X}$. In practice, $\mathbf{X}$ will have missing entries, since it is impossible to guarantee facial landmarks will be found for each audience member and time instant (*e.g.* a person's face may not be sufficiently visible to the camera).

Since we are interested in learning a non-linear low-dimensional encoding of the raw data, our models integrate both matrix/tensor factorization approaches as well as variational autoencoders into a common framework. Although we are interested in discovering spatiotemporal patterns across individuals, we present the formulation in terms of factorizing an arbitrary $N \times T \times D$ tensor.

3.1. Baselines

Tensor Factorization (TF). Tensor factorization is an established technique for identifying an underlying low-dimensional representation of K latent factors. In this case, we use the PARAFAC decomposition and factorize $\mathbf{X}$ into matrices $\mathbf{U} \in \mathbb{R}^{N\times K}$, $\mathbf{V} \in \mathbb{R}^{T\times K}$ and $\mathbf{F} \in \mathbb{R}^{D\times K}$ (see Fig. 2) such that each element of the original tensor is a linear combination of K latent factors from each matrix

$$x_{itd} = \sum_{k=1}^{K} U_{ik} V_{tk} F_{dk}. \tag{1}$$

Equivalently, (1) can be expressed using vector and matrix operations to generate the D dimensional slice of the original tensor at row i and column t

$$\mathbf{x}_{it} = (\mathbf{U}_i \circ \mathbf{V}_t)\, \mathbf{F}^{\mathsf{T}}, \tag{2}$$

where $\circ$ denotes the Hadamard product, $\mathbf{U}_i$ and $\mathbf{V}_t$ are K dimensional vectors from rows i and t of matrices $\mathbf{U}$ and $\mathbf{V}$ respectively (and represent the latent factors corresponding to subject i and time t), and $\mathbf{F}$ represents the latent spatial factors for facial landmarks.

Intuitively, each latent dimension $k = 1, \ldots, K$ corresponds to a separate archetype of how subjects react to the movie. For example, each column of $\mathbf{V}$ represents the time series motion associated with a particular archetype. This motion profile is used to modulate the latent spatial factors $\mathbf{F}$ to generate locations of facial landmarks for a particular time instant. The matrix $\mathbf{U}$ encodes the affinity of audience members to each archetype. For example, $\mathbf{U}_{ik}$ encodes how well the reaction of audience member i is described by archetype k. As is common in factorization approaches, we enforce non-negativity in $\mathbf{U} > 0$ in order to encourage interpretability (e.g., [28, 33, 53]).

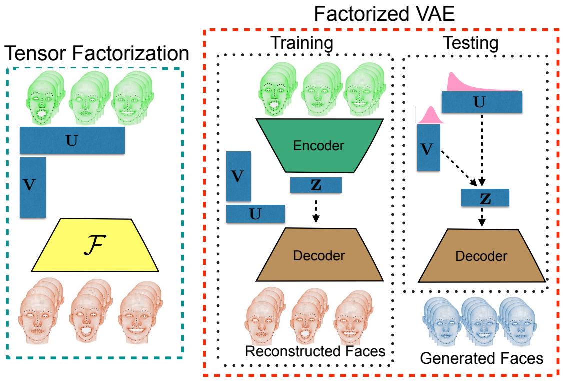

Figure 2. The figure in the left is a standard tensor factorization model. When $\mathcal{F}$ is a single matrix, it corresponds to model 1 (linear tensor factorization). When $\mathcal{F}$ is a multi-layer neural network, it corresponds to model 2 (nonlinear tensor factorization). The model on the right is the factorized variational auto encoder, it explains the training phase and testing phase of our model.

A key limitation of tensor factorization is that the underlying patterns in the data are modeled as a linear combination of latent factors. This linear assumption is not flexible enough and leads to a poor reconstruction of face reactions.

Nonlinear Tensor Factorization (NLTF). We next consider a non-linear version of probabilistic tensor factorization [41], which can be thought of as the simplest non-linear variant of generative factorization methods, and serves as a conceptual "bridge" to our proposed factorized variational autoencoder.

We first draw $\mathbf{U}_i$ from a log-normal, hence $\mathbf{U}_i \equiv e^{\mathbf{u}_i}$ where $\mathbf{u}_i \sim \mathcal{N}(0, \mathbf{I})$, which ensures positivity of the latent preference matrix $\mathbf{U}$. We also draw $\mathbf{V}_t \sim \mathcal{N}(0, \mathbf{I})$ from a Gaussian prior. We then draw the observations $\mathbf{x}_{ij} \sim \mathcal{N}(\mathbf{f}_\theta(\mathbf{U}_i \circ \mathbf{V}_t), \mathbf{I})$ where $\mathbf{f}_\theta$ is a deep neural network. One can view this model as a straightforward combination of deep learning with matrix factorization.

When the data set is large, the posterior is sharply peaked around its maximum mode. For inference, we can thus simply replace the latent variables by point estimates (MAP approximation). The Gaussian priors then simply become quadratic regularizers, and the objective to maximize is:

$$\begin{aligned}\mathcal{L}(\mathbf{U}, \mathbf{V}, \theta) &= \log p_\theta(\mathbf{x}|\mathbf{U}, \mathbf{V}) + \log p(\mathbf{U}) + \log p(\mathbf{V}) \\ &= \textstyle\sum_{it} ||\mathbf{x}_{it} - \mathbf{f}_\theta(e^{\mathbf{u}_i} \circ \mathbf{V}_t)||_2^2 + \sum_i ||\mathbf{u}_i||_2^2 + \sum_t ||\mathbf{V}_t||_2^2.\end{aligned} \tag{3}$$

As we show in our experiments, this straightforward combination of tensor factorization with deep learning does not provide the flexibility to properly model our audience facial expression dataset. In other words, even after the nonlinear transformation induced by the neural network, the data may not be amenable to linear factorization. This insight motivates our factorized variational autoencoder framework, which we describe next.

3.2. Our Framework

Variational Autoencoders. We first describe variational autoencoders (VAEs) [24], which form the core building block of our factorized VAE framework. A VAE is a generative latent variable model that embeds each $\mathbf{x}_{it}$ separately into a K-dimensional latent space. Its generative process involves drawing a latent variable $\mathbf{z}_{it} \in \mathbb{R}^K$ from a uniform prior distribution $p(\mathbf{z}_{it}) = \mathcal{N}(0, \mathbf{I})$, pushing the result through a decoder network $\mathbf{f}_\theta(\cdot)$ with network parameters θ, and adding Gaussian noise. In other words, we can model the likelihood as $p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{it} \mathcal{N}(\mathbf{f}_\theta(\mathbf{z}_{it}), \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. The generative process is thus:

$$\forall_{i,t} : \mathbf{z}_{it} \sim \mathcal{N}(0, \mathbf{I}), \quad \mathbf{x}_{it} \sim \mathcal{N}(\mathbf{f}_\theta(\mathbf{z}_{it}), \mathbf{I}). \tag{4}$$

We are interested in the posterior distribution over latent variables $p_\theta(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})/p_\theta(\mathbf{x})$, which has an intractable normalization $p_\theta(\mathbf{x})$. Using variational inference [19], we approximate the posterior using a variational distribution $q_\lambda(\mathbf{z}|\mathbf{x}) = \prod_{i,t} \mathcal{N}(\mu_\lambda(\mathbf{x}_{it}), \mathbf{\Sigma}_\lambda(\mathbf{x}_{it}))$. Here, $\mu_\lambda(\cdot)$ and $\mathbf{\Sigma}_\lambda(\cdot)$ are neural networks (encoders) with parameters λ. We call λ *variational parameters* which we optimize to minimize KL divergence between $q_\lambda(\mathbf{z}|\mathbf{x})$ and $p_\theta(\mathbf{z}|\mathbf{x})$. We can thereby also simultaneously optimize the model (decoder) parameters θ. One can show that this optimization is equivalent to maximizing the following variational objective [9, 24], which does not depend on $p_\theta(\mathbf{x})$:

$$\mathcal{L}(\theta, \lambda) = \mathbb{E}_q[\log p_\theta(\mathbf{x}|\mathbf{z})] - KL(q_\lambda(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \tag{5}$$

Note that the expectation can be carried out numerically by sampling from q in each gradient step.

VAEs can learn a very compact encoding of the raw data, which can then lead to low reconstruction error. However, in the context of audience facial analysis, the VAE by itself has extremely limited usefulness, since it does not relate different audience members or a single audience member at different times. For example, one cannot use VAEs to impute missing values when an audience member is not tracked at all times, which prohibits VAEs from being used in many prediction tasks. In order to properly harness the potential of VAEs, we must develop a VAE framework that can effectively capture collective effects of the data.

Factorized Variational Autoencoders (FVAEs). Our primary technical contribution is the factorized variational autoencoder (FVAE). This model jointly learns a non-linear encoding $\mathbf{z}_{it}$ for each face reaction $\mathbf{x}_{it}$, and jointly carries out a factorization in $\mathbf{z}$. In contrast to the non-linear tensor factorization baseline, the FVAE contains both local and global latent variables, which makes it able to learn more refined encodings. As such, the FVAE enables sharing of information across individuals while simultaneously taking into account the non-linearities of face deformations.

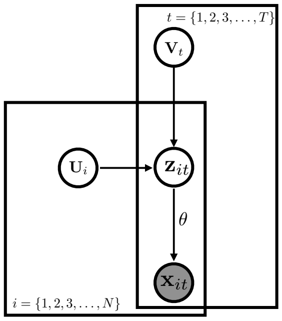

Figure 3. **Graphical model of Factorized VAE**

After drawing $\mathbf{U}_i$ from the standard log-normal and $\mathbf{V}_t$ from Gaussian priors, we draw $\mathbf{z}_{it} \sim \mathcal{N}(\mathbf{U}_i \circ \mathbf{V}_t, \mathbf{I})$, and then draw $\mathbf{x}_{it} \sim \mathcal{N}(\mathbf{f}_\theta(\mathbf{z}_{it}), \mathbf{I})$. In contrast to the non-linear tensor factorization baseline, we replace the hard constraint $\mathbf{z}_{it} = \mathbf{U}_i \circ \mathbf{V}_t$ with a soft constraint by adding Gaussian noise. As before, since $\mathbf{U}_i$ and $\mathbf{V}_t$ are global, we can MAP-approximate them. The objective is then:

$$\begin{aligned} &\mathcal{L}(\theta, \lambda, \mathbf{U}, \mathbf{V}) = \mathbb{E}_q[\log p_\theta(\mathbf{x}|\mathbf{z})] \\ &- KL(q_\lambda(\mathbf{z}|\mathbf{x})||\mathcal{N}(\mathbf{U} \circ \mathbf{V}, \mathbf{I})) + \log p(\mathbf{U}) + \log p(\mathbf{V}). \end{aligned} \tag{6}$$

Intuitively, the FVAE jointly learns an autoencoder embedding of the $\mathbf{x}_{it}$'s while also learning a factorization of the embedding. To see this, fix the embeddings $\mathbf{z} \equiv \mathbf{z}^*$. The remaining non-constant parts of the objective are $\mathcal{L}(\mathbf{U}, \mathbf{V}|\mathbf{z}^*) = \log \mathcal{N}(\mathbf{z}^*; \mathbf{U}_i \circ \mathbf{V}_t, \mathbf{I}) + \log p(\mathbf{U}) + \log p(\mathbf{V})$ which is just a factorization of the embedding space. Instead of having a simple normal Gaussian prior for $\mathbf{z}$, the FVAE prior is a probabilistic factorization model with global parameters $\mathbf{U}$ and $\mathbf{V}$ and log-normal and normal hyperpriors, respectively. Fig. 3 shows the graphical model depiction.

To perform prediction/matrix completion, a factorized variational autoencoder simply takes the Hadamard product $\mathbf{U}_i$ and $\mathbf{V}_t$ to get the corresponding latent factor $\mathbf{z}_{it}$, and then predicts the output by pushing $\mathbf{z}_{it}$ into the decoder. Because the FVAE is a generative model, it is also capable of producing new data, which is accomplished by drawing $\mathbf{U}_i$ and $\mathbf{V}_t$ from the prior distributions and pushing the Hadamard product through the decoder (see Fig. 2).

Optimization. Our approach follows the standard variational inference procedure for VAEs [9, 23]. In the forward

Movie	# Sessions	Time [min]	Genre
Ant Man	29	117	Action
Big Hero 6	11	102	Animation
Bridge of Spies	09	141	Drama
Inside Out	28	94	Animation
Star Wars: The Force Awakens	25	135	Action
The Finest Hours	06	115	Drama
The Good Dinosaur	13	93	Animation
The Jungle Book	17	105	Action
Zootopia	15	105	Animation

Table 1. **Movies**. The number of viewings of each film.

pass, we learn an encoder network which predicts the latent encoding for every data point, and a second network predicts a corresponding variance. In the backward pass, we learn a decoder network and minimizes reconstruction error, where the likelihood is averaged over samples of the stochastic hidden layer. The gradient of both encoder and decoder networks can be backpropped using the standard reparametrization trick as shown in [9]. In addition to these standard steps, we optimize $\mathbf{U}$ and $\mathbf{V}$ jointly along with the network to maximize likelihood using stochastic gradient descent.

4. Audience Data

In order to capture useful video signal in a dark movie theater environment, we employed a setup similar to Navarathna *et al.* [35]. We instrumented a 400 seat movie theater using four infra-red (IR) cameras and four IR illuminators placed above the projection screen. The cameras were outfitted with IR bandpass filters to remove the spill of visible light that reflected off the movie screen. The video was recorded at 12 frames per second with a resolution of 2750×2200 pixels. The resolution of faces ranged from 15×25 (back rows) to 40×55 (front rows). We collected over 150 viewings of 9 mainstream movies released in 2015 and 2016 (see Tab. 1). The length of the movies varies from 90 – 140 minutes. For each viewing, the number of audience ranges from 30 – 120.

Face Detection. Recently, King *et al.* [22] proposed 'Max-Margin Object Detection' (MMOD) which optimizes over all sub-windows to detect objects in images. This approach learns a Histogram of Oriented Gradients (HoG) [7] template on training images using structural support vector machines, which enables it to train on all sub-windows in every training image (efficienctly finding the 'hard negatives' automatically). We trained an MMOD face detector using the implementation in DLib [21] and manually labeled 800 training images. Due to the difference in resolution between the front and back rows of the theater, we created two face detection models: one for seats in the last three rows, and one for the rest of the theater. The face detectors run at 4 – 6 frames per seconds. We ran the detectors in every 0.5 seconds. On a validation set of 10000 frames, the precision and recall values are 99.5% and 92.2% for the front seats model and 98.1% and 71.1% for the back seats face detection model.

Method	Dataset		
	LFPW	HELEN	IBUG
RCPR	0.035	0.065	–
SDM	0.035	0.059	0.075
ESR	**0.034**	0.059	0.075
ERT	0.038	**0.049**	**0.064**

Table 2. **Landmark Performance**. The performance of different landmark detection algorithm across different datasets [20]. The average distance between predicted and groundtruth landmark locations are normalized by the inter-ocular distance. The ERT method of Kazemi *et al.* [20] has good performance on all datasets.

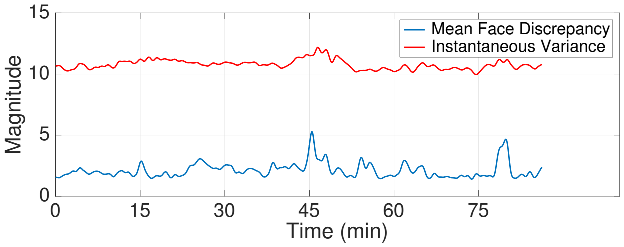

Figure 4. **Synchronicity**. The discrepancy between the global and instantaneous mean faces reveals moments where the audience (on average) is displaying a facial expression that is significantly different from the neutral face. The variance is mostly constant across time, which implies all audience members are exhibiting similar facial expressions.

Landmark Detection. After detecting faces, we use the DLib [21] implementation of ensemble of regression trees (ERT) [20] to detect facial landmarks. This method performs well compared with the state-of-the art approaches such as RCPR [3], SDM [52] and ESR [49] on standard data sets such as LFPW [2], HELEN [26] and IBUG [40] (see Tab. 2).

Face Frontalization. For each detected face in the video, we associate the 68 fitted 2D face landmark locations to a 3D face mesh from Face Warehouse [4]. We calculate the 3D rotation matrix $\mathbf{R}$ to estimate the roll, pitch and yaw of the detected face. Then rotation matrix $\mathbf{R}$ is used to generate a frontalized view of the 68 landmarks, which we denote as $\mathbf{x}_{it} = [x_{1,t}, y_{1,t}, \ldots, x_{68,t}, y_{68,t}]$, where i indicates the audience member and t indicates the frame number. The frontalized landmarks capture significant information such as overall face shape and instantaneous expression. For the most part, the geometric variation from head orientation is removed from the data. In practice, some skew-like residuals are present because of error fitting landmarks as well as warping across extreme pose changes.

Movie	Reconstruction MSE			Prediction MSE		
	TF	NLTF	FVAE	TF	NLTF	FVAE
Ant Man	1287.7	1261.5	**292.7**	1897.7	1349.1	**1325.7**
Big Hero 6	1394.4	1371.7	**275.3**	1505.6	1557.3	**1424.9**
Bridge of Spies	942.8	910.9	**184.3**	1288.3	1062.9	**960.0**
The Good Dinosaur	1156.5	1132.1	**275.4**	1328.4	1416.1	**1244.7**
Inside Out	1214.7	1132.7	**262.1**	1977.9	1240.3	**1161.4**
The Jungle Book	1150.0	1115.3	**186.4**	1622.1	1200.2	**1099.8**
Star Wars: TFA	1080.7	1047.4	**201.4**	1519.0	1192.2	**1085.8**
The Finest Hours	1015.3	962.9	**223.5**	1101.4	1114.0	**1038.8**
Zootopia	1181.0	1153.8	**189.9**	1277.4	1407.5	**1153.7**
Average	1158.1	1120.9	**232.3**	1502.0	1282.2	**1166.1**

Table 3. **Performance.** The performance of all three models with their best K values. FVAEs acheive the lowest training and testing error for all movies.

Data Inspection. We expect face landmarks will contain a strong signal that is conditioned on the audiovisual stimulus of the movie. To test this hypothesis, we calculated the global mean face $\bar{\mathbf{X}}$ across all audience members over all times, as well as the instantaneous mean face $\bar{\mathbf{X}}_t$ across all audience members at a particular time. The discrepancy between $\bar{\mathbf{X}}_t$ and $\bar{\mathbf{X}}$ reveals a signal with significant spikes (see Fig. 4). The variance of audience reactions is generally consistent across time, which implies viewers are all displaying similar, but temporarily varying, facial expressions.

5. Experiments

Our goal is to learn a compact and expressive representation of audience reactions that is semantically meaningful, so that we can identify patterns in the data and succinctly summarize observed behavior. To this end, we conduct a wide range of empirical analyses to broadly test the usefulness of FVAEs. We first evaluate how well FVAEs perform at matrix completion compared to our baseline models to establish the efficiency and accuracy of FVAEs. We then inspect the learned latent factors and show how some have semantic interpretations. Finally, we demonstrate the predictive power of FVAEs in the challenging task of anticipating the facial reactions of audience members for an entire movie, using only observations from the first few seconds/minutes. Our suite of results suggests that FVAEs can capture significantly more expressive representations than conventional baselines.

Dataset. The frontalized face landmarks are arranged into per-movie $N \times T \times D$ tensors. The overall missing data rate was approximately 13%. Since audience reactions are mostly synchronized, we analyze the data at one second intervals to compensate for individual reaction times. Based on the initial raw audience data, we cleaned the data to deal with false positive detections. We remove data with only short trajectories in the tensor. The tensors of the 9 movies finally have approximately 16 million total face landmarks from 3179 audience members.

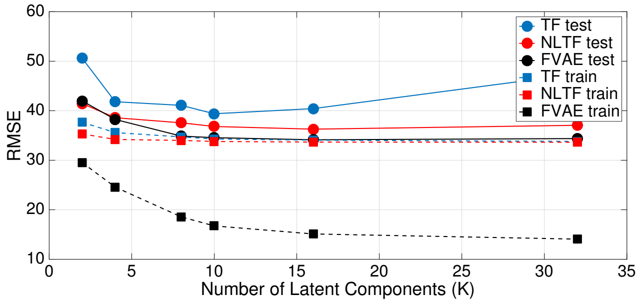

Figure 5. **Compactness and Expressiveness**. The training and testing performance of each model is shown for varying sized latent spaces. FVAEs are both compact (requiring few latent dimensions) and expressive (lowest testing error).

Implementation. During training, we set *mini-batch* size to 10 and use the *Adam* algorithm [23] for optimization. We used 3 stacked fully connected layers with *ReLU* for NLTF and encoder/decoder FVAE. The models are trained for 16 epochs.

5.1. Matrix Completion for Missing Data

For each movie, we split the observations of each audience member into training and testing data using random sampling (5:1 ratio). We compared against conventional TF and the NLTF described in Section 3.1. To determine the optimal number of latent dimensions, we first select four movies (Inside Out, The Jungle Book, The Good Dinosaur, Zootopia) and compare the training/testing performance for $K = \{2, 4, 8, 10, 16, 32\}$ (see Fig. 5). We measure the RMSE/MSE distance between predicted and actual landmark locations. The reported results are calculated per face instead of per dimension. For all values of K, TF has the worst performance. Moreover, there is a sharp increase in test error for $K = 32$ suggesting the model is unable to capture subtle dynamics within the data. NLTF achieves similar training performance, and slightly better testing performance. FVAE achieves significantly better performance in both training and testing. Similar to NLTF, the performance saturates around $K = 16$. These results suggest that the representation learned by FVAE is significantly more capable of imputing missing values than conventional baselines. The same trends hold up across all movies for three models with their best K values respectively ($K = 10$ for TF, $K = 16$ for NLTF and FVAE). (See Table 3).

FVAEs exhibit a significant discrepancy between training and testing error. This deviation arises because FVAEs automatically estimate a noise component for each training example to maximize the generalization capabilities. Each training $\mathbf{x}_{it}$ is encoded into a $\mathbf{z}_{it}$ which approximately factors into $\mathbf{u}_i$ and $\mathbf{v}_t$. At training time, $\mathbf{z}_{it}$ is pushed through the decoder, but at test time $\mathbf{u}_i \circ \mathbf{v}_t$ is decoded. Time complexity for predicting faces during the testing process using

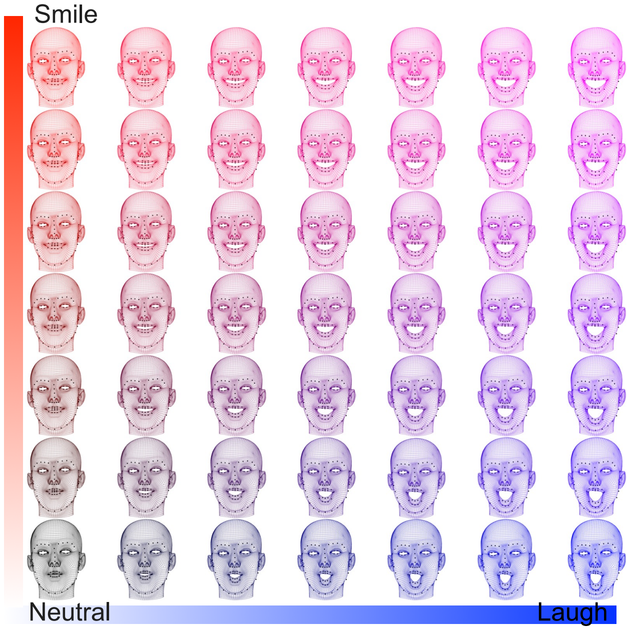

Figure 6. **Emotive Interpretation**. Two dimensions of the FVAE model learned for *Inside Out* resemble the facial expressions of smiling and laughing. We sample the latent space and show the resulting landmark locations generated by the decoder.

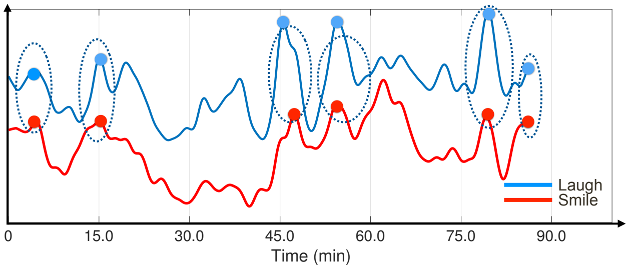

Figure 7. **Latent Temporal Factors**. An example of the $\mathbf{V}_t$ elements that modulate the smiling and laughing factors correlate with humorous scenes in the movie *Inside Out*.

TF and NLTF/FVAE are $\sim$ 9200 faces/s and $\sim$ 7700 faces/s respectively. The models are evaluated on an NVIDIA K40 GPU.

5.2. Visualizing the Latent Factors

We are interested in FVAEs having not only predictive power but also capturing semantically interesting concepts in the learned representation. Having semantically meaningful representations provides strong evidence that FVAEs can be used for a wide range of prediction tasks. Fig. 6 depicts two components of the FVAE model for $K = 16$. We see that these components correspond to smiling and laughing. The latent factor resembling smiling focuses on an upwards curved deformation of the mouth without opening it, whereas the factor resembling laughing captures the mouth opening in a big laugh.

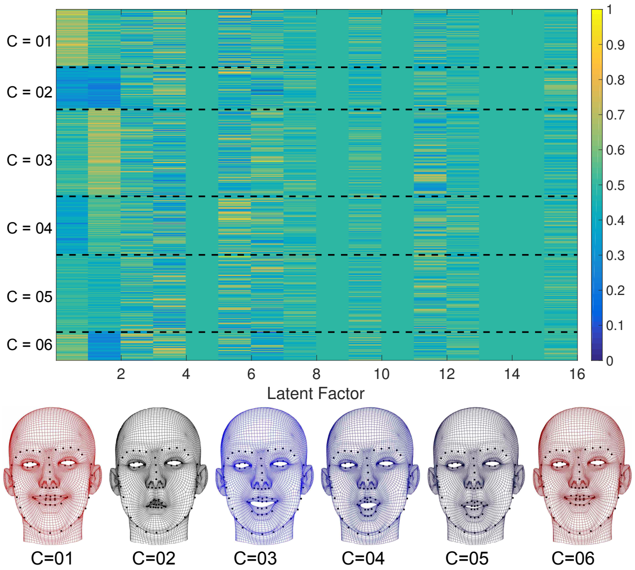

Figure 8. **Group Behavior**. The **U** matrix for *Inside Out* arranged to illustrate 6 clusters. Because the latent space encodes both rigid face shape and dynamic face expressions, clustering over all dimensions may lead to an over-segmentation (e.g. people who exhibit similar dynamic face expressions may be partitioned based on different rigid face shapes). Therefore, we cluster across only the smiling and laughing dimensions. Additionally, the regularization term in the loss function results in some latent dimensions having coefficients near 0.

To reinforce our interpretation that these latent factors are semantically meaningful, we plot the learned factors for the movie *Inside Out* (see Fig. 7). The plot illustrates how peaks in the smiling/laughing components correlate with significant moments in the film.

The baselines, TF and NLTF, on the other hand, do not learn interesting representations. In fact, the facial reconstructions do not vary much at all as one traverses the latent representation, which suggests that the encoding is not flexible enough to capture semantically meaningful variations of audience faces.

5.3. Analyzing Group Behavior

We can also use the learned representation to analyze phenomena like correlated group behavior. Because movies are crafted to elicit a desired response from the audience, we expect strong similarities between individual reactions. By clustering the rows of **U**, we can discover groups of audience members that exhibit similar behaviors (see Fig. 8).

The bottom row of Fig. 8 depicts exemplar faces for a humorous moment in the movie. Clusters 01 and 06 correspond to smiling (strong and weak affinity), and clusters 03, 04 and 05 correspond to laughing (from strong to weak affinity). Cluster 02 represents the small fraction of the audience which isn't exhibiting either laughing or smiling behaviour.

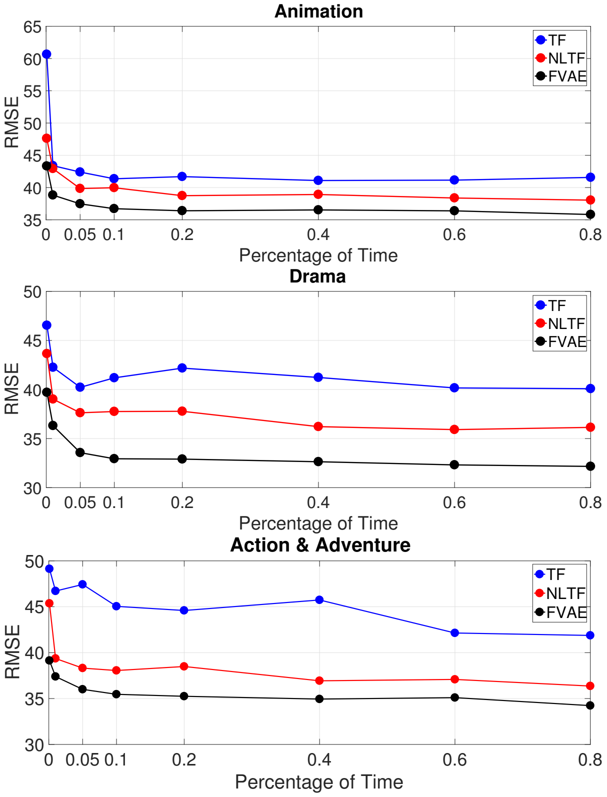

Figure 9. **Predicting Reactions**. We predict the future facial landmarks of each audience member for the remainder of the movie based how they react during the first few minutes of the movie for (top row) Animation, (middle row) Drama and (bottom row) Action & Adventure genre. After observing an audience member for ten minutes, our factorized VAE model typically has enough information to accurately predict the behavior of audience members for the remainder of the movie.

5.4. Predicting Reactions

Finally, we demonstrate the strong generalization performance of FVAEs by tackling an extremely difficult inference problem: predicting an audience member's facial reactions for an entire movie given only a subset of observations. Specifically, we estimate future facial landmarks locations for the remainder of the movie after making initial observations during the first few seconds/minutes of the film. The initial observations are used to estimate values of $\mathbf{U}_i$. Then, the Hadamard product $\mathbf{U}_i \circ \mathbf{V}_t$ is used to generate the predicted facial landmarks $\mathbf{x}_i$ for the entire movie.

Predicting future events/observations is well studied [15, 46], but previous work primarily focused on anticipating immediate events. In contrast, we evaluate the much more challenging task of predicting more than 60 minutes into the future, despite being only given a few minutes of initial observations.

For each movie, we train an FVAE model on 80% of the audience members, and use the remaining 20% to test long term predictions. We investigate different fractions $\lambda \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$ of the test data to use as observations to estimate $\mathbf{U}_i$.

Fig. 9 indicates the variation of performance for different fraction of λ for *Animation, Drama* and *Action & Adventure* movie genres. We observe that FVAE significantly outperforms both TF and NLTF. The non-linearity of NLTF results in better performance than TF, but the additional power of FVAEs to specifically search for generalizable patterns during training allows it to achieve substantially lower error.

The strength of these results for FVAE is quite striking. For example, using only the initial 5% of the data, FVAE can out perform both NLTF and TF when they have access to 100% of the data (i.e., FVAE outperforms the full-information encoding/reconstruction of TF and NLTF). Note that simply predicting the mean face has an order of magnitude higher error.

For all methods, we see that the prediction error drops quickly and saturates after observing the first 10% of data. Moreover, the long term prediction error is consistent with the testing error from matrix completion. Intuitively, the fact that prediction error saturates after observing audience reactions for 10% of the movie agrees with established guidelines that a film has roughly ten minutes to pull the audience into the story [10].

6. Summary

We have presented the factorized variational autoencoder (FVAE) for modeling audience facial expressions when watching movies. As our experiments demonstrated, tensor factorization (an established solution for collaborative filtering) failed to capture a compact and expressive latent representation because our data is complex and does not decompose linearly. Instead, the FVAE applies a non-linear variant of tensor factorization using deep variational autoencoders to learn a latent representation that factors linearly. Our formulation combines the compactness and interpretability of VAEs with the generalization performance of TF.

FVAEs are end-to-end trainable and demonstrated very strong predictive performance. After observing an audience member for a few minutes, FVAEs are able to reliably predict that viewer's facial expressions for the remainder of the movie. Furthermore, FVAEs were able to learn concepts of smiling and laughing, and that these signals correlate with humorous scenes in a movie. These results strongly suggest that learning factorized non-linear latent representations offers dramatically more expressiveness and generalization power than either factorization or autoencoders alone. Finally, our approach did not incorporate generic forms of domain knowledge, which may be useful in constraining the model when tackling even more complex settings.

References

[1] R. Bales. Social inteaction system: Theory and measurement. *New Brunswick, NJ:Transaction Publishers*, 1999. 2

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 35, pages 2930–2940, December 2013. 5

[3] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 1513–1520, Washington, DC, USA, 2013. IEEE Computer Society. 5

[4] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, Mar. 2014. 5

[5] M. Chaouachi, P. Chalfoun, I. Jraidi, and C. Frasson. Affect and mental engagement:towards adaptability for intelligent systems. *In FLAIRS*, 2010. 2

[6] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in neural information processing systems*, pages 2980–2988, 2015. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *In CVPR*, pages 886 –893, 2005. 5

[8] C. H. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2010. 2

[9] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 4, 5

[10] S. Field. *Screenplay: The Foundations of Screenwriting*. Random House Publishing Group, 2007. 8

[11] B. Goldberg, R. Sottilare, K. Brawner, and H. Holden. Predicting learner engagement during well-defined and ill-defined computer based intercultural interactions. *In Proceeding of the International Conference on Affective Computing and Intelligent Interaction*, 2011. 2

[12] P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013. 2

[13] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of the International Conference on Machine Learning*, 2015. 2

[14] 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. 2

[15] M. Hoai and F. De la Torre. Max-margin early event detectors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 8

[16] J. Huang and K. Murphy. Efficient inference in occlusion-aware generative models of images. *arXiv preprint arXiv:1511.06362*, 2015. 2

[17] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams. Structured VAEs: Composing probabilistic graphical models and variational autoencoders. *arXiv preprint arXiv:1603.06277*, 2016. 2

[18] H. Joho, J. Staiano, N. Sebe, and J. Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *In Multimedia Tools and Applications*, 2011. 2

[19] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. Springer, 1998. 4

[20] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014. 5

[21] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 5

[22] D. E. King. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015. 5

[23] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 4, 6

[24] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 4

[25] Y. Koren, R. Bell, C. Volinsky, et al. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. 1, 2

[26] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 679–692, Berlin, Heidelberg, 2012. Springer-Verlag. 5

[27] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 2

[28] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 2001. 3

[29] S. Makeig, J. Westerfield, J. Townsend, T. Jung, E. COurchesne, and T. Sejnowski. Functionally independent components of early event-related potentials in a visual spatial attention task. *Philosophical Transactions of the Royal Society: Biological Science*, 1999. 2

[30] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow. Adversarial autoencoders. In *Proceedings of the International Conference on Learning Representations*, 2015. 2

[31] S. Mandt, J. McInerney, F. Abrol, R. Ranganath, and D. Blei. Variational tempering. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 704–712, 2016. 2

[32] D. McDuff, R. Kaliouby, and R. Picard. Crowdsourcing facial responses to online videos. *In IEEE TOAC*, 2012. 2

[33] A. Miller, L. Bornn, R. Adams, and K. Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. In *ICML*, 2014. 3

[34] W. Murch. *In the Blink of an Eye: A Perspective on Film Editing*. new world (for sure) Part 5. Silman-James Press, 2001. 1

[35] R. Navarathna, P. Lucey, P. Carr, E. Carter, S. Sridharan, and I. Matthews. Predicting movie ratings from audience behaviors. *In IEEE Winter Conference on Applications in Computer Vision*, 2014. 2, 5

[36] J. Paisley, D. Blei, and M. I. Jordan. Bayesian nonnegative matrix factorization with stochastic variational inference. *Handbook of Mixed Membership Models and Their Applications. Chapman and Hall/CRC*, 2014. 2

[37] A. Pope, E. Bogart, and D. Bartolome. Biocybernetic system evalutes indices of operator engagement in automated task. *Biological Psychology*, 1995. 2

[38] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010. 2

[39] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 2

[40] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge. *Image Vision Comput.*, 47(C):3–18, Mar. 2016. 5

[41] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, volume 20, pages 1–8, 2011. 2, 3

[42] N. Schwarz and F. Strack. Reports of subjective well-being: Judgmental processes and their methodological implications. *Well-being: The foundations of hedonic psychology*, 1999. 2

[43] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder variational autoencoders. In *Advances in neural information processing systems*, 2016. 2

[44] T. Teixerira, M. Wedel, and R. Pieters. Emotion-induced engagement in internet video advertisements. *Journal of Marketing Research*, 2011. 2

[45] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller. A Deep Semi-NMF Model for Learning Hidden Representations. In *International Conference on Machine Learning (ICML)*, 2014. 2

[46] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Proceedings of IEEE European Conference on Computer Vision*, 2016. 8

[47] S. Wang, J. Tang, Y. Wang, and H. Liu. Exploring implicit hierarchical structures for recommender systems. In *Proceedings of 24th International Joint Conference on Artificial Intelligence*, 2015. 2

[48] M. Wedal and R. Pieters. Eye fixations on advertisements and memory for brands: a model and findings. *Marketing Sciences*, pages 297–312, 2000. 2

[49] Y. Wei. Face alignment by explicit shape regression. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 2887–2894, Washington, DC, USA, 2012. IEEE Computer Society. 5

[50] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Towards practical smile detection. *In TPAMI*, 2009. 2

[51] J. Whitehill, Z. Serpell, Y. Lin, A. Foster, and J. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing,*, 5(1):86–98, Jan 2014. 2

[52] X. Xiong and F. De la Torre Frade. Supervised descent method and its applications to face alignment. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2013. 5

[53] Y. Yue, P. Lucey, P. Carr, A. Bialkowski, and I. Matthews. Learning fine-grained spatial models for dynamic sports play prediction. In *2014 IEEE International Conference on Data Mining*, 2014. 3

[54] Y. Yue, C. Wang, K. El-Arini, and C. Guestrin. Personalized collaborative clustering. In *Proceedings of the 23rd international conference on World wide web*, pages 75–84. ACM, 2014. 2