Proceedings of the 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1--8.
Lille (France), 14-18 May 2019.

Facial Expression Editing with Continuous Emotion Labels

Alexandra Lindt[1], Pablo Barros[2], Henrique Siqueira[2] and Stefan Wermter[2]
[1] University of Hamburg, Hamburg, Germany
[2] Knowledge Technology, University of Hamburg, Hamburg, Germany

***Abstract*— Recently deep generative models have achieved impressive results in the field of automated facial expression editing. However, the approaches presented so far presume a discrete representation of human emotions and are therefore limited in the modelling of non-discrete emotional expressions. To overcome this limitation, we explore how continuous emotion representations can be used to control automated expression editing. We propose a deep generative model that can be used to manipulate facial expressions in facial images according to continuous two-dimensional emotion labels. One dimension represents an emotion's valence, the other represents its degree of arousal. We demonstrate the functionality of our model with a quantitative analysis using classifier networks as well as with a qualitative analysis.**

I. INTRODUCTION

Finding an automated way to alter the expression in a facial image is relevant for a variety of different fields, such as face animation, face recognition, emotion recognition or human-computer interaction. It has therefore been the topic of various publications over the last decade. The most recent publications on the task made use of deep generative models [14], [18], [24]. Radford et al. [32] describe a variation of Generative Adversarial Networks (GAN) [14] named Deep Convolutional GAN (DCGAN), which can realize facial expression transfer through arithmetic operations on the latent input vectors of the generator network. However, the generated images are of low resolution and are limited to expressing only one discrete emotion class (happy, sad, excited, etc.).

Another recent solution, proposed by Yeh et al. [40], combines the structure of the Variational Autoencoder [18] with the approach of flow-based image warping [39]. The resulting Flow Variational Autoencoder (FVAE) is trained to create an expression flow map that can be used to transform an image to express a specific emotion. The FVAE is capable of generating high-resolution images and can even synthesize different intensities of a discrete emotion class in a face. Unfortunately, it can only be trained with paired data samples (i.e. pictures of the same face with different emotions), the emotion in the input image must be known and only images in which the face is shown frontally can be processed.

The approach of Song et al., the Geometry-Guided GAN (G2-GAN) [35], consists of two pairs of GAN [14] that form a mapping cycle in which one GAN applies an emotion to a neutral face image and the second GAN neutralizes the expressive face image. The emotion to be synthesized or removed is represented with facial geometry points, which makes the framework flexible in the emotions that can be synthesized and enables it to create different intensities of basic emotions. However, paired training data is necessary to train G2-GAN.

Recently, Zhang et al. introduced the Conditional Adversarial Autoencoder (CAAE) [41] for manipulating the age of a face in an image. The model extends a conventional Adversarial Autoencoder [24] by a second discriminator network that ensures the generation of a photo-realistic output image. The CAAE is able to transform a facial image so that it corresponds to one out of ten discrete age groups. It does not require paired training data samples and has proven to be robust to variations in the input images [41].

Two approaches used the CAAE's structure as a basis. The first one is the Conditional Difference Adversarial Autoencoder (CDAAE) [42], which extends the CAAE by a feedforward connection between encoder and decoder network. Unfortunately, the CDAAE can only be trained with paired training data and the generated images have quite low resolution. The second CAAE-based approach is the Expression Generative Adversarial Network (ExprGAN) [8]. It extends the CAAE by an expression controller module that enables it to create discrete facial expressions of different intensities as well as by a face identity preserving loss function. Both ExprGan and CDAAE are able to synthesize mixed emotions as percentages of the emotion classes of their training set (e.g. 50% sadness and 50% fear).

Although very successful in their specific tasks, the approaches discussed above are in some ways limited in modelling non-discrete emotions because they presume a discrete representation of human emotions. This is probably due to the fact that until the recent release of the AffectNet database [26] only facial expression databases with discrete annotations were available. AffectNet's images, in contrast, are annotated with a two-dimensional vector that represents an depicted emotion as its degree of valence (unpleasant-pleasant) and arousal (relaxed-aroused). With the availability of this data, the motivation arises to investigate the applicability of continuous two-dimensional emotion representation in the field of automated expression editing.

With the goal of providing a controllable face expression editing mechanism that produces high-fidelity and high-quality facial translation, we employ the AffectNet database [26] to train a deep generative model for the manipulation of facial images according to continuous two-dimensional emotion labels. To validate the contribution of our model to the generation of the desired emotional expressions, we

Proceedings of the 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1--8.
Lille (France), 14-18 May 2019.

conduct an objective experiment where individual neural networks are used to measure the arousal and valence of the generated images. Our experiments show that our model does edit the faces with the intended arousal and valence. Finally, we provide an analysis of how the proposed model imposes the facial expressions on the original images in order to better explain our contributions.

II. Background

A. Representation of Human Emotion

From a psychological point of view, there are two different approaches to categorizing human emotions and the corresponding facial expressions. A discrete or categorical representation of emotions assumes a set of fundamentally distinct basic emotions [36], [10]. Since several scientists have defined these basic emotions differently [36], [31], [16], it is not clear which exact emotions belong to them. However, there is widespread agreement on the following six emotions: anger, disgust, fear, happiness, sadness and surprise [30]. In contrast, a continuous or dimensional emotion representation is based on the assumption that emotions cannot be divided into distinct groups, but can rather be described within a continuous space [33], [38]. Russell described emotions as points in the two-dimensional space of valence (unpleasant-pleasant) and arousal (relaxed-aroused) and termed this space *The Circumplex Model of Affect* [33].

There is no consensus on which model represents human emotions best. Several publications have investigated whether humans intuitively use a continuous or categorical representation of emotion. For both representation forms there are publications with evidence, some of them even contradict each other directly [11], [4]. Further, some researchers examined the connections between a perceived emotion and the physical state or language of a subject. More scientists were able to show clear correlations for the continuous model [30]. A detailed description and a list of all related experiments can be found in [30]. This indicates that the two-dimensional emotion representation is intuitively used by human beings.

B. Deep Generative Models

Generative modeling currently has two main approaches: Generative Adversarial Networks (GAN) [14] and Variational Autoencoders (VAE) [18]. GAN consist of a discriminator and a generator network that are trained simultaneously in a minimax two-player game. This process results in a generator network that is able to generate high-dimensional data samples similar to those of the training data set. An extension of GAN is the Conditional GAN (CGAN) [25], whose generated output is further influenced by a conditional variable. In contrast to GAN, the VAE is a stochastic model that consists of an encoder and a decoder network. The encoder network maps an input to a latent representation and the decoder network subsequently uses the latent vector to reconstruct the input. A prior distribution is imposed on the latent space through the loss function of the model. After training, the decoder network can be employed to create data samples from latent vectors that are sampled from the imposed distribution. The Adversarial Autoencoder (AAE) [24] integrates the idea of GAN into the VAE by using an adversarial training process to impose the prior distribution to the latent space. For this purpose, an additional discriminator network is employed as adversary.

Our proposed model is a VAE with an additional discriminator network on the generated output. This additional network imposes the distribution of the training data on the generated data samples and, therefore, causes the generation of photo-realistic facial images that express the target emotion.

III. Proposed Model

For automated expression editing according to continuous two-dimensional emotion labels, we propose an updated version of the Conditional Adversarial Autoencoder (CAAE) [41]. We choose this model because it does not require paired data (i.e. pictures of the same face with different emotions) for training and has also proven to be robust against variations in the input images [41].

Inspired by ExprGan [8] and G2-GAN [35], we extend the CAAE by an identity-preserving loss on the reconstructed image. This loss forces the output image to show the same person as the input image. The original CAAE approach trains encoder and generator network with a loss function on total variation minimization [23] to reduce ghosting artifacts in the generated images. Since we have empirically found that these artifacts do not appear in the output images when the CAAE is trained with the identity-preserving loss, the total variation minimization loss is omitted in our approach. In order to improve the quality of the synthesized emotional expressions, we further change the influence of the emotion label on the discriminator network on output and emotion label.

A. Architecture

As illustrated in Fig. 1, our model consists of a total of four networks: an encoder network E, a generator network G and two discriminator networks D_{img} and D_z. It receives a face image x and a continuous two-dimensional emotional label y as input and outputs an image x_{gen} that displays the face from the input image expressing the emotion represented by y. During the model's training, y describes the emotion that is expressed in x.

1) Encoder and Generator: Given an input image x and respective emotion label y, the encoder network E first maps the input image to a lower-dimensional representation $E(x) = z$, which represents the individual properties of the face depicted in input x. This representation is subsequently used by the generator network G to create a reconstruction $G(z, y) = x_{gen}$ of x conditioned on z and the two-dimensional input emotion label y. In order to reconstruct x as good as possible, both networks are trained with an image reconstruction loss L_{rec}, which is defined in (1). L_1 denotes the mean absolute error [5].

$$\min_{E,G} L_{rec} = L_1(x, G(E(x), z)) \tag{1}$$

Proceedings of the 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1--8.
Lille (France), 14-18 May 2019.

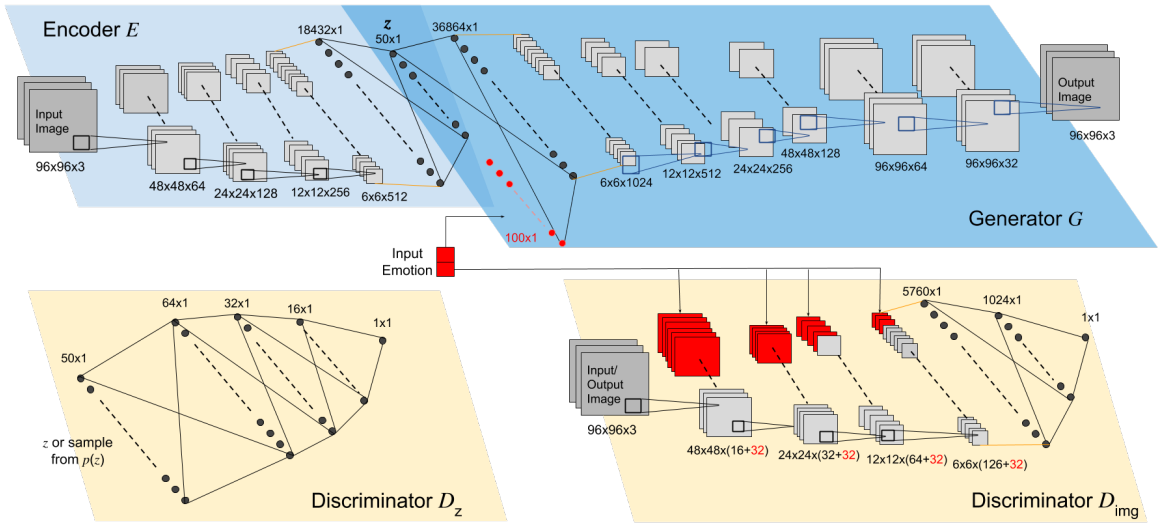

Fig. 1. Detailed Architecture of the CAAE used for expression editing according to continuous two-dimensional emotion labels. The encoder network E maps an input image x to a smaller representation z through four convolution layers and one fully connected layer. Note that the orange line denotes a reshape operation. The vector z is concatenated to the enlarged input emotion label y (bright red) and the resulting vector serves as input for the generator network G. G consists of one fully connected layer followed by six transposed convolution layers and outputs an image x_{gen} of the same size as x that shows the face from x expressing y. The discriminator D_z imposes the prior distribution $p(z)$ on the generated z while the discriminator D_{img} ensures that x_{gen} is photo-realistic and expresses y.

We compute the identity-preserving loss in the same way as ExprGan [8] and therefore use the pre-trained VGG face model by Parkhi et al. [28]. The VGG face model originally classifies the identity of a face in an image. To compute whether it assumes two facial images to show the same person, the activations of five of its convolutional layers are compared. The calculation of the identity-preserving loss is defined as

$$\min_{E,G} L_{iden} = \sum_{l} L_1(\phi_l(x), \phi_l(G(E(x), z)). \tag{2}$$

In this context, ϕ_l denotes the activation of the lth layer. The considered layers are the *conv1_2*, *conv2_2*, *conv3_2*, *conv4_2* and *conv5_2* layer of the VGG face model.

2) Discriminator D_z : The discriminator network D_z is imposed on the encoder network's output z and fosters it to be uniformly distributed. To this end, it receives either $z = E(x)$ or a sample from the uniform distribution $p_{prior}(z)$ as input and is trained to distinguish between both. The adversarial loss function between E and D_z is defined in (3). $\mathbb{E}$ denotes the likelihood [13], $p_{prior}(z)$ the prior distribution imposed on the internal representation z and $p_{data}(x)$ the distribution of the training images.

$$\begin{aligned}\min_{E} \max_{D_z} L_z = \quad & \mathbb{E}_{z_{prior} \sim p_{prior}(z)}[\log D_z(z_{prior})] \\ + \quad & \mathbb{E}_{x \sim p_{data}(x)}[\log(1 - D_z(E(x)))].\end{aligned} \tag{3}$$

3) Discriminator D_{img} : The second discriminator network D_{img} is employed to ensure that the generator network G produces a photo-realistic output image x_{gen} that expresses the input emotion y. Therefore, D_{img} receives the generated image $x_{gen} = G(E(x), y)$ and the input emotion label y or an image and corresponding emotion label from the database as input. D_{img} is trained to distinguish between generated and true pairs, while the generator network G is trained to trick D_{img} with its generated outputs. The adversarial loss function between G and D_{img} is defined in (4), where $p_{data}(x, y)$ denotes the distribution of the training data.

$$\begin{aligned}\min_{G} \max_{D_{img}} L_{img} = \quad & \mathbb{E}_{x,y \sim p_{data}(x,y)}[\log D_{img}(x, y)] \\ + \quad & \mathbb{E}_{x,y \sim p_{data}(x,y)} \\ & [\log(1 - D_{img}(G(E(x), y), y)].\end{aligned} \tag{4}$$

4) Overall Loss Function: The overall loss-function L_{total} of the model is described in (5) as the weighted sum of all loss functions. The coefficients λ_1, λ_2, λ_3 and λ_4 balance the resolution of the generated images, the quality of the generated emotions and the obtained identity features in the generated images.

$$\min_{E,G} \max_{D_z, D_{img}} L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{iden} + \lambda_3 L_z + \lambda_4 L_{img} \tag{5}$$

B. Implementation Details

1) Structural Details: The following details extend on Fig. 1. The encoder network E is a Convolutional Neural Network [20] that consists of four convolution layers and one fully connected output layer. Inspired by DCGAN [32], a convolution of stride 2 is employed instead of pooling. This allows the encoder network to learn its own spatial downsampling [41].

E obtains an image x as input and transforms into a vector z of size 50×1. The two-dimensional input emotion label y is scaled to 50 times its size and subsequently

Proceedings of the 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1--8.
Lille (France), 14-18 May 2019.

concatenated to z. The resulting vector serves as input for the generator network G, which up-samples its input via one fully connected layer and six transposed convolution layers [9] into an output image of the same size as the input image. The first four layers use a stride of 2, the last two layers use a stride of 1. All convolution layers in E and transposed convolution layers in G use a kernel size of 5. All values of x, y and z are in $[-1, 1]$.

D_{img} consists of four convolution layers and two fully connected layers. The convolutions have a stride of 2 and kernel size of 5. Batch normalization [15] is applied after each convolution. Following each block of convolution and batch normalization, the enlarged emotion label y is concatenated to the block's output. We found this repeated concatenation of y crucial for the performance of our CAAE, as it leads to a higher quality of the emotions synthesized in the output image x_{gen}. D_{img} outputs a value from $[0, 1]$, which represents the estimated probability of its input being a pair of image and emotion label from the original data set.

Finally, D_z receives a 50×1 vector as input that is either the output z of the encoder network or a 50-dimensional vector sampled from the uniform distribution over $[-1, 1]^{50}$. By using four fully connected layers, of which the first three are followed by batch normalization [15], the discriminator converts its 50-dimensional input into a one-dimensional output in $[0, 1]$ that represents the probability that D_z's input was sampled from the uniform distribution.

2) Hyperparameters: We trained the proposed model with the overall loss function defined in (5). The model has empirically been found to produce the best outputs for $\lambda_1 = 1$, $\lambda_2 = \frac{1}{3}$, $\lambda_3 = 0.01$ and $\lambda_3 = 0.01$. At the same time, the batch size should not be smaller than 49. To reduce computational cost, we only calculated the identity preserving loss for 16 out of 49 images in a batch. The normal distribution with mean 0 and standard deviation 0.02 is employed for the initialization of the weights of all layers. All biases are initially set to 0. For optimization, the Adam Optimizer [17] with learning rate $\alpha = 0.0002$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$ is employed. We implemented the model using the machine learning framework TensorFlow [1].

Our particular architecture design and hyper parameters were found by using an empirical exploratory search. Our goal was to maximize the quality of the generated faces, and thus not represent an objective search. The observation of the generated faces were made purely subjective to the author's opinion, and yet, present an impressive performance on facial expression editing as demonstrated bellow.

C. Facial Expression Editing

After training, our encoder and generator network can be employed to manipulate emotions in facial images according to arbitrary two-dimensional emotion labels. To modify an image x to express an emotion y, we feed x to our trained encoder network E, obtain the identity representation $E(x)$ and use the trained generator network G to create an output image $x_{gen} = G(E(x), y)$. The synthesized image x_{gen} shows the face from x expressing y.

IV. Experimental Evaluation

Following, we evaluate our model for its performance in automated expression editing according to continuous two-dimensional emotion labels. To enable an appropriate interpretation and analysis, we first explain three evaluation metrics. Then we take a closer look at the data set on which the model is trained. Finally, the capabilities of the model are investigated in two experiments.

A. Evaluation Metrics

1) Root Mean Squared Error: The Root Mean Squared Error (RMSE) [5] is a common evaluation metric for the difference between two vectors of the same size. Given two vectors x and y, both of size n, it is defined as

$$RMSE(x, y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

2) Concordance Congruence Coefficent: The Concordance Congruence Coefficent (CCC) [37] is a statistical measure of the agreement between the values of two equally sized vectors x and y. It combines the Pearson's correlation coefficient with the squared difference. The CCC is defined as

$$CCC(x, y) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where ρ denotes the Pearson's correlation coefficient between the two vectors, σ^2 describes the variance of the respective vector and μ its mean value. The CCC can take values between -1 and 1 where 1 stands for a strong similarity and -1 for oppositeness. Unlike the Pearson Correlation Coefficient, the CCC penalizes predictions that are well correlated with the ground truth but shifted in value in proportion to their deviation. This property makes the CCC metric a meaningful metric for the evaluation of our two-dimensional emotion labels $\in [-1, 1]^2$. Not only the correlation between the predicted emotion and the true emotion is considered, but also the prediction value's divergence from the real value.

3) Sign Agreement: The Sign Agreement (SAGR) [27] is defined for two vectors x and y of equal length n as

$$SAGR(x, y) = \frac{1}{n}\sum_{i=1}^{n}\delta(sign(x_i), sign(y_i))$$

where $\delta(x, y)$ denotes the Kronecker delta

$$\delta(x, y) = \begin{cases} 1 & \text{for } x = y \\ 0 & \text{for } x \neq x \end{cases}$$

The SAGR measures how much the signs of the individual values of two vectors x and y match. It takes on values in $[0, 1]$, where 1 represents complete agreement and 0 represents complete contradiction. The SAGR metric is well suited to evaluate the predictions of our emotion labels [26]. Let's consider a face image that expresses a valence of +0.3 (i.e. a slightly positive valence). Although the two predictions +0.7 and -0.1 would have the same RMSE, the prediction of +0.7 is better suited, since +0.7 also denotes a positive valence [26].

Proceedings of the 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1--8.
Lille (France), 14-18 May 2019.

B. Data Set

The AffectNet database [26] is an extensive database of facial images that are annotated with both a categorical and a continuous two-dimensional emotion representation. The two-dimensional emotion labels are $\in [-1, 1]^2$, where one dimension represents the emotion's valence (unpleasant-pleasant) and one represents its degree of arousal (relaxed-aroused). In the categorical representation of emotions, a distinction is made between eleven discrete emotion classes. The database contains about 450.000 manually annotated images. A further 550.000 included images were labeled with a classifier network trained on the hand-annotated images. The images were collected using search queries of emotion-related keywords in different languages in three major search engines. The pictures are therefore very diverse in terms of lighting, colors, camera angle and background as well as in head position, age, gender and ethnicity of the subjects. The facial expressions shown in the images are mostly natural and spontaneous.

Although the images were labeled by skilled annotators, there are considerable variations in the annotations of different individuals. To measure the agreement between the valence and arousal values chosen by different annotators, 36.000 images were annotated by two people. Table I shows the annotations' agreement measured in RMSE, SAGR and CCC. It is apparent that there are noticeable differences in annotation, especially with regard to the arousal values.

C. Experiments

1) Quantitative Experiment: For a quantitative evaluation of the images generated by our model, we need a measure of the quality of the synthesized emotions. For this purpose, we train two Convolutional Neural Networks (CNN) [20] on the classification of valence and arousal in face images of size $96 \times 96 \times 3$, which is the size of images generated by our model. It should be noted that the analysis of valence and arousal in natural face images is not an easy task. As already demonstrated by table I, even human experts struggle to produce consistent annotations. Both classifier networks have the same architecture specified in table II. To determine this structure, we trained a variety of networks on the hand-annotated AffectNet [26] images and selected the one with the best performance on our validation set of 4500 omitted AffectNet images. The final structure is oriented towards a CNN proposed for valence and arousal classification in [19], which is itself based on the VGG-16 model [34]. Both of our classifier networks output a value in $[-1, 1]$ which expresses the estimated valence or arousal respectively. The classifiers were trained using stochastic gradient descent [3] with learning rate 0.001 and a momentum of 0.9. The mean absolute error [5] was employed as loss function. After training, our classifier networks achieve an accuracy that corresponds to that of other networks trained for the identical or similar task [19], [26], [29]. Table III presents the classifiers networks' accuracy on the validation data measured by the previously introduced evaluation metrics. For the quantitative evaluation of our model, we employ the classifier networks for the classification of 220500 images generated from our model.

2) Qualitative Experiment: To investigate which facial attributes are changed by our model in order to make a face image express a particular emotion, we apply our model to 200 uniform face images for 49 different emotion labels each. The images used in this experiment are taken from the

Classifier Network Layer
Convolution(Kernel 3x3, Stride 1, 64 Filters) Max-Pooling(Kernel 2x2, Stride 2)
Convolution(Kernel 3x3, Stride 1, 128 Filters) Max-Pooling(Kernel 2x2, Stride 2)
Convolution(Kernel 3x3, Stride 1, 256 Filters) Convolution(Kernel 3x3, Stride 1, 256 Filters) Max-Pooling(Kernel 2x2, Stride 2)
Convolution(Kernel 3x3, Stride 1, 512 Filters) Convolution(Kernel 3x3, Stride 1, 512 Filters) Max-Pooling(Kernel 2x2, Stride 2)
Fully Connected (4096 units) Dropout(dropout probability 0.5)
Fully Connected (2622 units) Dropout(dropout probability 0.5)
Fully Connected 2622 units) Dropout(dropout probability 0.5)
Fully Connected (1 unit)

TABLE II

DETAILED ARCHITECTURE OF THE CLASSIFIER NETWORKS FOR THE QUANTITATIVE ANALYSIS

	Valence	Arousal
RMSE	0.340	0.362
SAGR	0.815	0.667
CCC	0.821	0.551

TABLE I

AGREEMENT BETWEEN TWO ANNOTATORS OF THE AFFECTNET DATABASE FOR VALENCE AND AROUSAL LABELS RESPECTIVELY. MEASURED IN RMSE, SAGR AND CCC. [26]

	Classifier on Valence	Classifier on Arousal
RMSE	0.450	0.411
SAGR	0.676	0.708
CCC	0.484	0.405

TABLE III

ACCURACY OF THE CLASSIFIER NETWORKS EMPLOYED FOR THE QUANTITATIVE EVALUATION OF OUR MODEL MEASURED IN RMSE, SAGR AND CCC.

Proceedings of the 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1--8.
Lille (France), 14-18 May 2019.

CelebA data set [21], which contains portraits of prominent people. The images are aligned in such a way that the nose of the depicted person is located at the center of the image and the depicted faces are about the same size. We select the first 200 CelebA images that show a frontal face and cut them quadratically around their center. With this method, we obtain 200 quite precisely aligned facial images.

Each test image is processed by our trained network for 49 different two-dimensional emotion labels in $[-1, 1]^2$. One of the emotion labels is the neutral emotion [0,0] (i.e. neutral valence and arousal). Therefore, we receive an emotionally neutral version for every input image. Each of the 48 non-neutral generated images is compared with this neutral image. We deliberately choose to not compare the generated images with the input image, as the CelebA images are mainly red carpet images in which the subject laughs or smiles.

For every non-neutral generated image, a grayscale heatmap that displays the differences to the corresponding neutral image is calculated. With $R(p)$ defining the red channel, $G(p)$ the green channel and $B(p)$ the blue channel of a pixel p in in an image I, the heatmap between two same-sized images I_1 and I_2 is computed pixel-wise for every $p_1 \in I_1$ and corresponding $p_2 \in I_2$ with

$$max(abs(R(p_1) - R(p_2)), \\ abs(G(p_1) - G(p_2)), abs(B(p_1) - B(p_2))).$$

In the resulting heatmap, the brightest areas show the biggest differences between I_1 and I_2. Since all faces in the test images have approximately the same position and the same size, we can add all 200 heatmaps and normalize the outcome to get a general heatmap for each of the 48 non-neutral emotion labels.

D. Results

1) Quantitative Analysis: Table IV shows the results of the evaluation of 220500 images generated by our model for various emotion labels. Of these 220500 images, 18000 express extreme emotions (i.e. very low/high arousal and very low/high valence). These extreme images are also evaluated separately, to enable a comparison between images generated for an average emotion label and those generated for extreme emotion labels. The CCC values indicate a positive correlation between the emotion labels of the generated images and the emotion labels assigned to them by the classifier networks. It is higher for images generated for extreme emotions. Although the values do not indicate a strong similarity of the generated and classified labels, the similarity can be assumed if one considers that this relatively low correspondence is also present among human annotators (see table I). The SAGR metric shows that only half of the emotion label's signs match for all images. For extreme emotions, however, SAGR shows a clear connection between the signs of the emotion labels of the generated images and the emotion labels estimated by the classifiers. The RMSE is lower for images of all emotions than for images of extreme emotions. This suggests that for extreme images, larger errors were made in the emotion classification. When looking at the RMSE, however, it should be considered that the classifications for images with valence/arousal levels close to neutral cannot deviate as much from their true value as those for extreme emotions. To exemplify, the classification of the neutral emotion value 0 must be in [-1,1] and can, therefore, deviate by a maximum of 1 from the true value, while for an extreme emotion value (-1 or 1) the classified emotion can deviate up to 2. Furthermore, very large errors for single images (outliers) are heavily weighted in the RMSE [5].

	All Images		Extreme Images	
	Valence	Arousal	Valence	Arousal
RMSE	0.528	0.607	0.671	0.720
SAGR	0.567	0.483	0.691	0.624
CCC	0.312	0.210	0.353	0.267

TABLE IV
RESULTS OF THE CLASSIFICATION OF 220500 GENERATED IMAGES (THEREOF 18000 EXTREME IMAGES) WITH OUR CLASSIFIER NETWORKS. MASURED IN RMSE, SAGR AND CCC.

2) Qualitative Analysis: Fig. 2 shows the overall grayscale heatmaps for 48 different emotion labels. Note that the changes in the images with weaker emotions (i.e. emotions close to the neutral emotion) are as expected smaller than in the images with at least one extreme emotion value (i.e. emotions of very high or low valence and/or very high or low arousal). Furthermore, it can be noticed that

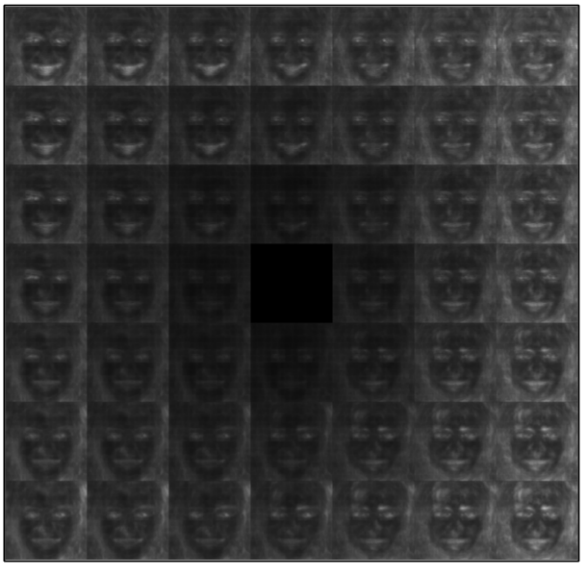

Fig. 2. Computed overall heatmaps for 48 different emotion labels of the form [valence, arousal] $\in [-1, 1]^2$, from high arousal (left) to low arousal (right) and from high valence (top) to low valence (bottom). The brightest areas in the heatmaps show the greatest overall differences between the images generated for the respective emotion label and the images generated for the neutral emotion label [0, 0]. See the digital version for details.

Proceedings of the 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), pp. 1--8.
Lille (France), 14-18 May 2019.

there are strong changes in the background of the images. This background noise might have been caused by the fact that the only penalty the model receives for changing the background is the pixel wise reconstruction loss. In contrast, changing the face in the right way is much more rewarding. Therefore, quite specific changes can be observed within the facial region. Our model primarily alters the mouth, eyebrows and eye area.

For a high valence, the corners of the mouth are pulled upwards, the mouth is opened, the upper side of the eyes rises and a slight wrinkle forms under the eyes. These modifications might be associated with an emotion of positive value, such as happiness or delight. A lowered valence causes changes on the inside of the eyebrows as well as on the downside of the eye. These facial movements can be related to a depressed mood, where the inner sides of the eyebrows are moved down a little, pressing the eyes slightly downwards. A high degree of arousal affects the upper lip, the upper part of the eyes and the rear eyebrow arch. All these changes are signs of an excited or surprised facial expression. Looking at the heatmaps for a low degree of arousal, there is clearly more background noise for which there is no discernible reason. However, within the facial region we can see changes in the area of the mouth as well as at the inner eyebrows and the eyes. These changes can be interpreted as signs of sleepiness, calmness or relaxation. The mouth becomes more closed, the eyes become narrower and the eyebrows move down a little.

Overall it can be found that the heatmaps combine the characteristics of their individual dimension values. For instance, the heatmap for a face with [high valence, high arousal] combines the heatmaps of [high valence, neutral arousal] and [neutral valence, high arousal]. Both high valence and high arousal cause a widening of the eyes. On the heatmap for combined high valence and high arousal, these changes seem to have added up to an even larger eye opening. Furthermore, the eyebrows are raised, just as it is characteristic for a high arousal value. The raised corners of the mouth, which were previously caused by high valence, can also be observed.

Fig. 3. Examples of images generated by our model for one input image (top) and 25 different emotion labels $\in [-1, 1]^2$, from high arousal (left) to low arousal (right) and from high valence (top) to low valence (bottom). See the digital version for details.

Fig. 4. Examples of images generated by our model for 25 different emotion labels $\in [-1, 1]^2$ and 25 different input images, from high arousal (left) to low arousal (right) and from high valence (top) to low valence (bottom). See the digital version for details.

V. Discussion, Conclusion And Future Work

A. Discussion

Our proposed model is well suited for editing facial expressions according to emotion labels from a two-dimensional emotion representation of valence and arousal. The generated images can be described as photo-realistic, the synthesized emotions appear natural and realistic and the identity of the person in the input image is largely maintained in the output image (see Fig. 3). Our qualitative evaluation showed that the model ascribes certain characteristics to the values of the individual emotion dimensions and combines them sensibly for mixed emotion labels. As depicted in Fig. 4, our model can be successfully applied to images that differ in terms of lighting, colors, the position of the depicted face or the characteristics of the subject, such as ethnicity, age or gender. The synthesized emotions are rather subtle, which is probably due to the fact that our training data contains almost only natural facial expressions which are in many cases less extreme than posed expressions or expressions observed in a laboratory environment for the reaction to emotional stimuli.

The model produces overall consistent outputs. However, our observations and the quantitative evaluation indicate that extreme emotions in the input images can influence our model's output. The synthesized face images are less representative for their emotion label than the original face images and the emotions in images created for an extreme emotion label are better recognizable than in images created for an arbitrary emotion label. This suggest that a target emotion applied to an image changes its emotion only to a certain extent. We further find that some images from our validation data set cannot be processed correctly by our model. These pictures typically have very low contrasts, an very unusual composition, coloring or illumination or parts

of the depicted face are hidden or cut off.

B. Conclusion

In this paper, we introduced a neurocomputational model for facial expression editing according to continuous two-dimension emotion labels. The model is capable of generating high-quality and overall consistent outputs. From our observations we can conclude that our model changes faces in pictures in an understandable and plausible way. Our quantitative evaluation also demonstrates that our proposed model is able to generate facial expressions with continual conditions. The combination of the quantitative evaluation and the observations demonstrates how different conditions impact our solution, and how robust it is for different image conditions.

C. Future Work

Since our model can generate natural appearing facial images with realistic emotional expressions, its produced images could be used to train machine learning models for tasks like emotion recognition or face recognition. Our solution could be used as a boostrap for one-shot-based learning models. Also, the exploration of sequential data generation would be of fundamental importance for real-world applications.

ACKNOWLEDGMENT

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169) and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 721619 (SOCRATES).

REFERENCES

[1] M. Abadi et al., Tensorflow: A system for large-scale machine learning, *in Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, November 2016.

[2] T. Blanz and T. Vetter, A morphable model for the synthesis of 3d faces, *in Proceedings of the 26th annual Conference on Computer Graphics and Interactive Technique*, August 1999.

[3] L. Bottou, Large-scale machine learning with stochastic gradient descent, *in Proceedings of COMPSTAT2010*, 2010.

[4] J. M. Carroll and J. A. Russell, Do Facial Expressions Signal Specific Emotions? Judging Emotion From the Face in Context, *Journal of Personality and Social Psychology*, 1996.

[5] T. Chai and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?, *Geoscientific Model Development Discussions*, February 2014.

[6] R. Cowie et al., What a neural net needs to know about emotion words, *in Proceedings of the CSCC99*, 1999.

[7] R. Cowie et al., Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*, 2001.

[8] H. Ding, K. Sricharan, and R. Chellappa, ExprGAN: Facial Expression Editing with Controllable Expression Intensity, *AAAI Conference on Artificial Intelligence*, February 2018.

[9] V. Dumoulin and F. Visin, A guide to convolution arithmetic for deep learning, *ArXiv e-prints*, March 2016 .

[10] P. Ekman, An argument for basic emotions, *Cognition and Emotion*, 1992.

[11] P. Ekman and W. V. Friesen, Constants across cultures in the face and emotion, *Journal of Personality and Social Psychology*, 1971.

[12] P. Ekman and W. V. Friesen, Measuring facial movement, *Environmental psychology and nonverbal behavior*, 1976

[13] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016, pp.174 .

[14] I. J. Goodfellow et al., Generative Adversarial Networks, *Conference on Neural Information Processing Systems*, December 2014.

[15] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, *International Conference on Machine Learning*, July 2015.

[16] C. E. Izard, D. Z Libero, P. Putnam, and O. M. Haynes, Stability of emotion experiences and their relations to traits of personality, *Journal of Personality and Social Psychology*, 1993

[17] D. Kingma and J. Ba, Adam: A method for stochastic optimization, *International Conference on Learning Representations*, May 2015.

[18] D. P. Kingma and M. Welling, Auto-Encoding Variational Bayes, *International Conference on Learning Representations*, April 2014.

[19] D. Kollias and S. Zafeiriou, A Multi-component CNN-RNN Approach for Dimensional Emotion Recognition in-the-wild, *ArXiv e-prints*, May 2018.

[20] Y. Lecun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, May 2015.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang, Deep Learning Face Attributes in the Wild, *in Proceedings of International Conference on Computer Vision*, December 2015.

[22] Z. Liu, G. Song, J. Cai, T. Cham, and J. Zhang, Conditional adversarial synthesis of 3d facial action units, *ArXiv e-prints*, March 2018.

[23] A. Mahendran and A. Vedaldi, Understanding deep image representations by inverting them, *Conference on Computer Vision and Pattern Recognition*, June 2015.

[24] A. Makhzani et al., Adversarial Autoencoders, *International Conference on Learning Representations*, May 2016.

[25] M. Mirza and S. Osindero, Conditional Generative Adversarial Nets, *Deep Learning and Representation Learning Workshop: NIPS*, December 2014.

[26] A. Mollahosseini, B. Hasani, and M. H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild, *IEEE Transactions on Affective Computing*, 2017.

[27] M. A. Nicolaou, H. Gunes, and M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space, *IEEE Transactions on Affective Computing*, April 2011.

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman, , *in British Machine Vision Conference*, September 2015.

[29] S. Peng, L. Zhang, Y. Ban, M. Fang, and S. Winkler, A Deep Network for Arousal-Valence Emotion Prediction with Acoustic-Visual Cues, *ArXiv e-prints*, May 2018.

[30] C. Peter and A. Herbon, Emotion representation and physiology assignments in digital systems, *Interacting with computers*, 18(2), 2006, pp.139-170.

[31] Plutchik, R., A general psychoevolutionary theory of emotion, *in Approaches to Emotion*, Erlbaum, 1984, pp. 197-219.

[32] A. Radford, L. Metz, and S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, *International Conference on Learning Representations*, May 2016.

[33] J. A. Russell, A circumplex model of affect, *Journal of Personality and Social Psychology*, 1980.

[34] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *ArXiv e-prints*, June 2014.

[35] L. Song et al., Geometry guided adversarial facial expression synthesis, ArXiv e-prints, January 2017.

[36] S. S. Tomkins, *Affect Imagery Consciousness: Volume I II, The Positive Affects*, Springer, 1962.

[37] M. Valstar et al., Avec 2016: Depression, mood, and emotion recognition workshop and challenge, *in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, October 2016.

[38] D. Västfjäll and M. Friman, The measurement of core affects: a Swedish self-report measure derived from the affect circumplex, *Göteborg Psychological Reports*, 2000.

[39] F. Yang et al., Expression flow for 3D-aware face component transfer, *AACM Transactions on Graphics*, July 2011.

[40] R. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala, Semantic Facial Expression Editing using Autoencoded Flow, *ArXiv e-prints*, November 2016.

[41] Z. Zhang, Y. Song, and H. Qi, Age progression/regression by conditional adversarial autoencoder, *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

[42] Y. Zhou and B. E. Shi, Photorealistic Facial Expression Synthesis by the Conditional Difference Adversarial Autoencoder, *International Conference on Affective Computing and Intelligent Interaction*, October 2017.