The Many Variations of Emotion

Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Frédéric Jurie

The Many Variations of Emotion

Valentin Vielzeuf[1,2], Corentin Kervadec[1], Stéphane Pateux[1] and Frédéric Jurie[2]
[1] Orange Labs, Rennes
[2] Normandie Univ., UNICAEN, ENSICAEN, CNRS

***Abstract*— This paper presents a novel approach for changing facial expressions in images. Its strength lies in its ability to map face images into a vector space in which users can easily control and generate novel facial expressions based on emotions. It relies on two main components. The first one learns how to map face images to a 3-dimensional vector space issued from a neural network trained for emotion classification. The second one is an *image to image translator* allowing to translate faces to faces with expressing different emotions, the emotions being represented as 3D points in the aforementioned vector space. The paper also shows that the proposed face embedding has several interesting properties: i) while being a continuous space it allows to represent discrete emotions efficiently and hence enables to use those discrete emotions as targeted facial expressions ii) this space is easy to sample and enables a fine-grained control on the generated emotions iii) the 3 orthogonal axes of this space may be mapped to arousal, valence and dominance – 3 directions used by psychologists to describe emotions – which again is highly interesting to control the generation of facial expressions.**

I. Introduction

Affective computing is a topic of broad interest, finding applications in many fields such as health care, marketing or human-machine interface. Therefore, a great effort has been put in the recognition of emotions in different contents. Several works propose to analyze facial expressions from images [1], [25], from multimodal videos [6], [18], [29], [41] or from multi-view videos [2], [38]. Other works focus more on sentiment expressed by text [4], [26], [30] or audio [14], [32], [33], building a very large and complete set of emotion recognition methods. Nevertheless, some recent works [40] underline that the performance might begin to saturate on the emotion recognition task because of the nature of the used datasets and the subjective representations of emotion.

Thus, understanding and manipulating emotion representations is of tremendous interest to progress towards more complete affective computing abilities. Consequently, the very definition of facial expressions of emotions has to be carefully addressed. The literature came up with three main definitions. First, Ekman *et al.* [9] proposed the concept of discrete emotions, identifying six universal classes (e.g. "Happy", "Sad", "Disgusted", "Fearful", "Surprised" and "Angry"). Later, the arousal-valence system was built by Russell [31], placing emotions in a 2-d continuous space. Finally, the Facial Action Coding Systems [10] allows to objectively represent facial expressions with Action Units (e.g. "raised eyebrows").

As facial expressions are one of the main ways for human to express emotions, a whole branch of the affective computing community focuses on the generation of facial expressions, aiming at better measuring the effect of emotion representations as well as understanding how to simulate emotions. The first works on the synthesis of artificial facial expressions came from the computer graphics community and were focused on the animation of faces with model-based approaches [16], [34], [42]. More recently deep learning approaches and especially Generative Adversarial Networks [5], [7], [12], [28], [37] have been proposed, borrowing ideas from the computer graphics community [27], [35], but also aiming to learn these facial expressions from diverse datasets and representations using machine learning techniques. These recent approaches are generally trained on small corpuses with pronounced emotions and limited annotations [20], [21], [44], leading to limited spaces of possible expressions.

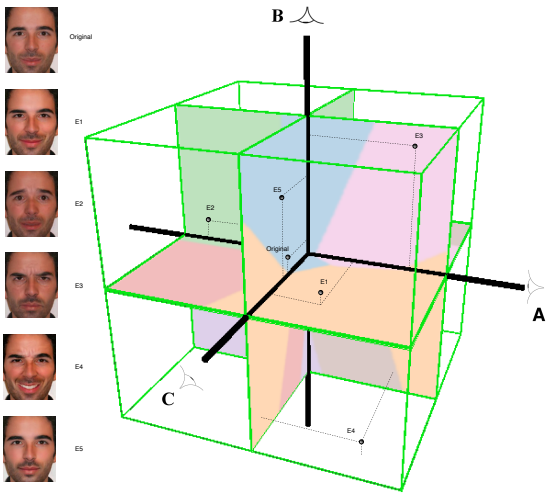

Fig. 1. Illustration of our 3-d representation space of emotion. Each color represents a discrete emotion ('happy', etc.). E_i are some generated facial expressions (left) mapped in our 3-d space (right). Better viewed in color.

Even if larger 'in the wild' corpuses annotated with action units or arousal valence information exist [11], [19], [24], they are very costly to annotate and their use is only marginal [27].

Last but not the least, the previously mentioned approaches have focused on the quality of the generated faces and not on the way to control the targeted emotions.

This paper proposes an elegant way to control targeted emotions by building a bridge between psychological inter-

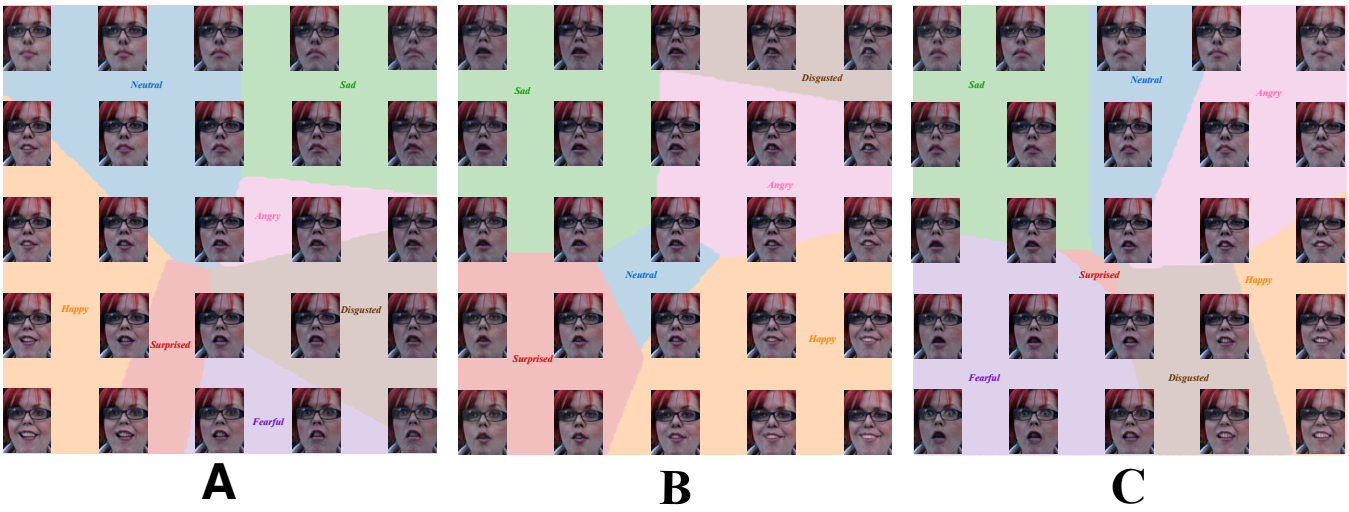

Fig. 2. Illustration of our 3-d representation space of emotion. Each plane is colored with the associated discrete emotion classes (see position of these planes in Figure 1). Furthermore, for each plane we illustrate some generated faces from the associated 3D representation coordinates. The generated sample faces within same color areas show the many possible faces inside a given emotion class. Note that expressions in these three planes are only a small part of the possible samples generated by our model. Better viewed in color.

pretations of emotion representations and what is visually observed. Our contribution is twofold. First, we propose a 3-d representation of emotion based on the latent space of a classification neural network. This model is trained on the task of *discrete emotion classification*, and thus does not require any costly continuous annotations (such as arousal or valence values). Second, once this representation space is settled, we can use it to learn an image-to-image translator based on generative adversarial networks, to generate target faces having controlled expressions, for any source face. We show that not only the generation of faces is more robust than with other representations, but also that we could exhibit complementary directions within the 3-d space representation that are in line with the common psychological definition of arousal, valence and dominance. It enables easy and meaningful ways to control emotions as interesting interpretations of the observed improvements[1].

II. Related Works

a) Emotion Representation: Emotion representation is a well-explored topic in the psychological community, as mentioned in Section I. Therefore, an easy way to build a taxonomy of the different emotion representations is to categorize them along two directions: semantic meaning and power of description. The higher semantic meaning comes with discrete emotion [9]. Each class is associated with one word, but at the cost of loosing a lot of power of description. Indeed behind one word many variations may be found. Proposing compound emotion [8] (*e.g.* happily surprised) is a way to reach a more fine-grained representation while keeping a high-level semantic meaning. Nevertheless even with a large vocabulary of words, the whole space of emotion may not be completely described. Indeed, as shown by Russel *et al*, emotion is a continuum, thus requiring a continuous system to obtain a fine-grained description. To keep a semantic meaning, several interpretable axis were proposed to build continuous spaces, such as arousal, valence or even dominance [22]. At a much lower semantic level, but with a perfect depiction of the facial expression, the computer graphics community tends to propose a Facial Action Coding System, allowing to objectively represent facial expression with Action Units (*e.g.* "raised eyebrow").

Using datasets annotated by representations with a great power of description allows to train more efficient model, but it implies a higher annotation cost. Our method is aiming to reach a compromise between having a great power of description and a low-cost annotation process.

b) Computer Graphics: The face animation task has already been actively explored by the computer graphics community, some early works proposing 3D model-based approaches [3], [43]. More recently, Soladié *et al.* [34] uses a 4-d emotion representation space to animate face and Active-Appearance Models features. In a more general fashion, Weber *et al.* [42] propose an unsupervised person-specific model which easily adapts to the targeted subject. Finally, hybrid approaches mixing deep learning and model-based method are also proposed. Susskind *et al.* [36] first propose to train a deep belief network based on both action units and identity information to generate facial expression. More recently, another approach [35] is using fiducial points to geometrically control the face animation while Tulyakov *et al.* [37] is learning to directly generate sequences of images, based on a "content and motion" method. Quia *et al.* [28] use facial landmarks to improve the animation smoothness of a changing emotion. Kim *et al.* [16] enable to generate video face animation using another portrait video as an example. These approaches are working on the very shape of the face. Therefore it implies complex modifications of the model to adapt to "in the wild" conditions where important illumination changes and occlusions are common.

[1]Additional generated facial expressions and supplementary works will be shared at https://github.com/vielzeuf/The-Many-Variations-of-Emotion/ .

c) Generative Neural Networks: To fulfill the previous requirement of robustness towards real "in the wild" conditions, an interesting path of research for image synthesis using neural networks is Generative Adversarial Networks(GAN) [12] and Variational AutoEncoders(VAE) [17]. Focusing on GANs, many extensions exist, such as Conditional GAN [23] where a condition variable allows to control the generation or more recently StarGAN, where Choi *et. al* [5] propose a multi-domain method, learning both facial attribute transfer and facial expression generation. Interestingly, the targeted facial expression is fed with the input face to modify, allowing an end-to-end approach. Extending the previous works, Ding *et al* [7] propose a new GAN framework enabling to learn the intensity of a facial expression by a specific encoding of the emotion label. The covered domain of possible facial expression is then larger than for classical discrete approaches, each class containing many variations along an intensity criteria. Nevertheless, this approach does not allow to generate all the possible facial expressions such as compound emotions. Pumarola *et al.* [27] propose a more general approach, coupling GAN and Action Units to continuously generate facial expressions from a large dataset. This implies a lot of labeling work, as action units are costly to annotate. Moreover the constructed space has a high dimension (15 action units) leading to non direct analysis of the organization of the generated faces.

III. Proposed Method

The proposed approach relies on two components. First, a method allowing to map face images into a 3-d vector space describing facial expressions. Second, an *image-to-image* translator capable of generating faces with controlled facial expressions. The generator represents facial expressions as 3-d points in the vector space.

A. Representing Facial Expressions in a 3-d Vector Space

As argued by [31], continuous annotations can be more subtle and more accurate than discrete labels to represent emotions. The rationale behind this idea is to continuously describe several features of facial expressions, such as intensity (arousal) or pleasure (valence). A mapping to discrete representations of emotion is always possible [24], [31], enabling to take benefits from both discrete and continuous representations. The number of continuous features necessary to describe emotions is debatable. However, some psychological studies, e.g. [22], suggest that the two usual dimensions (arousal and valence) might not be sufficient to represent the whole spectrum of emotions. Despite the advantages of such continuous representations, building annotated datasets costs much more than annotating them with discrete labels. Consequently, large datasets with such continuous annotations do not exist, preventing us to learn a direct mapping.

Interestingly, a recent approach [15] shows that a compact latent space issued from the hidden layers of a convolutional neural network trained for discrete emotion classification can lead to an arousal-valence like topology. We build on this work, proposing to use the latent space of a neural network as the 3-d vector space for representing facial expressions.

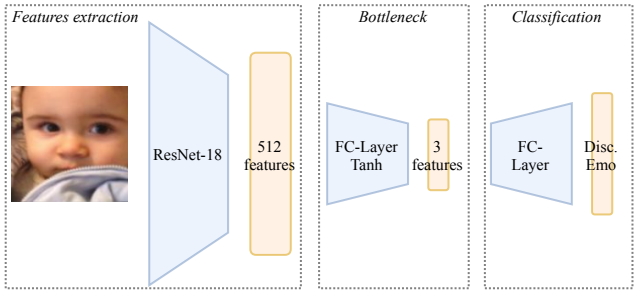

Fig. 3. Learning a 3-d compact representation of emotions, inspired by the work of [15]

In practice, we trained a modified ResNet-18 [13] to classify discrete emotions, as in Figure 3. The modification consists in adding a bottleneck – which is a fully connected layer – before the classification layer, forcing the classifier to use only three features to predict discrete emotions. A hyperbolic tangent activation is applied on these three features to ensure that the 3-d features are in the range $[-1, 1]$. While this network is trained using a corpus annotated with discrete emotions, it can be used in a second time to map any new face image to the 3-d vector space.

We can observe in Figure 1 how discrete emotions are related with their 3-d representations. Figure 2 gives another illustration of the embedding in the 3 planes of the reference frame of Figure 1.

B. Facial Expression Modification as an Image-to-Image Translation problem

We consider the task of facial expression modification as an image-to-image translation problem. We build on the StarGAN [5] algorithm, which allows to take both a face and a targeted expression as input of the generator and has already proven to be efficient on discrete emotion generation. As a reminder, the model is composed of a discriminator D and a generator G, trained by minimizing two loss functions. Our main contribution on this point is to adapt these loss functions to the continuous labelling case.

The *adversarial loss* aims at making the generated fake images not distinguishable from real one:

$$L_{adv} = \mathbb{E}_x[logD(x)] + \mathbb{E}_{x,r}[1 - logD((G(x, r)))] \quad (1)$$

where x is the input image and r is the 3-d representation. The generator and the discriminator respectively aim at minimizing and maximizing the term.

The StarGAN classification loss, which we replaced by a *regression loss*, is itself composed of two terms. The first term, namely L_{reg}^{real}, forces D to correctly regress the emotion associated with the original image. While the second term, namely L_{reg}^{fake}, forces G to generate facial expressions with a representation close to the targeted emotion. More formally,

we use Mean Squared Error terms as follows:

$$L_{reg}^{real} = \mathbb{E}_{x,r}[D(x)-r]^2 \text{ and } L_{reg}^{fake} = \mathbb{E}_{x,r}[D(G(x,r))-r]^2 \quad (2)$$

The *reconstruction loss*, namely L_{rec} ensures that the generated faces conserve other information than emotion (identity, orientation, etc.). It is defined as:

$$L_{rec} = \mathbb{E}_{x,r_1,r_2}[||x - G(G(x,r_2),r_1)||_1] \quad (3)$$

where r_1 is the original facial expression representation and r_2 is the representation of the facial expression the generator has to generate.

Finally, we write the generator and discriminator losses as in the StarGAN [5]:

$$L_D = -L_{adv} + \lambda_{reg} L_{reg}^{real} \quad (4)$$

$$L_G = L_{adv} + \lambda_{reg} L_{reg}^{fake} + \lambda_{rec} L_{rec} \quad (5)$$

C. Controlling the Facial Expressions of Targeted Images

We recall that our goal is to allow users to generate face images whose facial expressions are controlled. Within our framework, this is equivalent to computing the 3-d vector characterizing the target expression (denoted as r in previous Section). Except for the case where we want to mimic the expression of another face – in this case we can directly obtain the expression vector of the target face – this is not straightforward as the vector space might look arbitrary and not user-friendly. To make the control possible, we imagined two scenarios: one consists in specifying the target expression by discrete emotions. Another one consists in specifying the target emotions in terms of arousal, valence and dominance, with respect to the original face.

1) Controlling Facial Expression by Discrete Emotions: Controlling facial expressions by discrete emotions is possible if we map the discrete emotion in our new 3-d space. We did it by taking a dataset with images annotated with discrete emotions (AffectNet [24]) and computed the centroid of the 3-d coordinates of each image of each one of the 7 emotions. More formally, we compute $r^i = \sum_{k \in C_{discrete}^i} \frac{r_k}{\#C_{discrete}^i}$, where r^i is the coordinates of the centroid of the class i, $C_{discrete}^i$ is the set of all elements of the class i, and r_k is our 3-d representation of the sample k.

2) Generation of Emotions based on Arousal, Valence and Dominance: The previous method does not allow a fine control of the generated facial expressions. Controlling expression directly within our 3-d space would allow more flexibility, but the raw coordinates are not meaningful to users. Interestingly, we are going to show how our 3-d space can be aligned with the arousal/valence (a/v) representation, which is easy to understand. For aligning our representation with the a/v space we use a dataset in which images are annotated in terms of a/v (AffectNet [24]). For the images of this dataset we have both their coordinates in our 3-d space (denoted as r_k for the k^{th} image) as well as their coordinates in the a/v space (denoted as (av_k)), given by the annotations. It allows us to learn a mapping from the 3-d space to the a/v space. Let i_a and i_v the two vectors pointing in the arousal and in the valence direction of the vector space. We obtain i_a and i_v by:

$$\min_{i_a,i_v} \sum_k ||av_i - [i_a; i_v] \times r_k||^2 \quad (6)$$

This estimation is validated in the results section (see Table I). The third vector of this new basis is obtained by vector product: $i_d = i_a \bigotimes i_v$. We show in the section IV that i_d is related to the *dominance* factor.

Finally, we can generate a novel expression r' from r by computing $r' = a_a \times i_a + a_v \times i_v + a_d \times i_d$ where (a_a, a_v, a_d) are the arousal, valence and dominance-like of the targeted expression.

	Baseline [24]	Human [24]	Ours
Arousal RMSE	0.40	0.36	0.34
Valence RMSE	0.394	0.34	0.36

TABLE I

PROJECTION OF OUR REPRESENTATION TO THE AROUSAL VALENCE SPACE.

IV. RESULTS

This section evaluates the benefits of the proposed continuous representation. After detailing our experimental protocol, we evaluate the ability of our approach to generate discrete and continuous emotions. Then we enlighten the relations between our learned representation and arousal/valence, and we build a bridge towards psychological interpretations, finding back a third dimension visually similar to the concept of dominance [22].

A. Implementation Details

Our experiments use the recent AffectNet dataset [24], which provides both discrete emotions and arousal valence annotations. We sanitized it by discarding images without faces or without annotations, resulting in 297000 annotated faces that are 'in the wild'. We preprocess them, with a face detector and a landmark aligner. To fit our convolutional neural networks requirements, faces are then resized to 256x256x3. During training time, we also apply data augmentation (scale jittering, rotation and flip). It gives us a modified ResNet-18 (see explanations before), mapping each face to the 3-d vector space.

We then train 3 image-to-image translators denoted as: discreteGAN, avGAN, and ours, one for each annotation type (resp. discrete, arousal-valence, and ours). We use a batch size of 16, a learning rate of 1e-4 with exponential decay factor of 0.996. The architectures of both generator and discriminator are similar to the one described in [5]. For continuous annotations, the regression loss weight λ_{reg} is changed to 3 instead of 1 because of the scale difference with a classification cross-entropy loss. Other weights are same as in [5]. Parameters are optimized with the Adam method during 300,000 iterations.

To evaluate the generation task, we use the validation set faces of AffectNet and generate faces of size 128x128x3.

Fig. 4. The 7 emotion classes generated with the three different approaches: discreteGAN (first row), avGAN (second row) and ours (third row). The three source images are randomly taken from the test set, to ensure a fair comparison between the approaches.

B. Discrete Expression Generation

We evaluate in this section the impact of the different representations on the quality of the generated expressions, when targeted expressions are discrete. We compared the 3 image-to-image translators introduced in the previous section, namely: discreteGAN, avGAN, and ours.

Fig. 4 shows, for 3 source image, generated facial expressions for the 7 discrete emotion classes. Generated images bear similarity for the 3 generators. Nevertheless, we note some interesting differences. For the 'happy' class (second column), teeth look more natural with our continuous model than with discreteGAN or avGan. For the 'disgust' class, we note the presence of artifacts in the faces generated by the discreteGAN, while avGAN tends to generate relatively similar expressions for both 'disgusted' and 'angry' faces. These artefacts can be explained by the small number of 'disgust' occurrences in the training set (less than 2%), meaning that the discreteGAN did not see many examples of this class. Furthermore 'disgust' and 'anger' classes are relatively close in the arousal valence space, leading to very similar values for their centroids and thus very similar generated expressions. Our GAN, using a 3rd dimension as shown in Fig. 1, improves the separation of the 2 classes and leads to more visible differences between generated faces.

	RMSE on mean color				L_{rec}
GAN	Red	Green	Blue	All	
Discrete	4.5	6.3	10.2	7	0.22
AV	6.4	7.8	5.7	6.7	0.14
Ours	**3.7**	**3.1**	**3.2**	**3.4**	**0.12**

TABLE II

EVALUATION OF THE RECONSTRUCTION QUALITY AND COLOR CONSERVATION OF THE DIFFERENT APPROACHES ON THE TEST SET. LOWER IS BETTER.

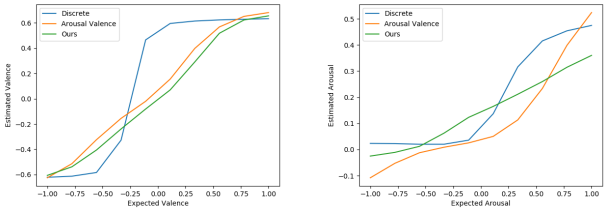

Fig. 5. Estimated valence (left) and arousal (right) of the generated faces as a function of the targeted valence (left) and arousal (right) for the 3 different models. These plots are average plots on the whole test set.

The neutral class is also interesting, especially for the 3rd face, where we can note the ability of our GAN to improve the control of the mouth closing.

We also can note that the intensity of the expressions is higher in the discreteGAN faces. It can be explained by the fact that we choose the centroids for continuous representations, which are not the most pronounced expressions.

Finally, we may think from the third face that the discreteGAN is changing the mean color of the faces it generates. Another good qualitative example can be seen on Figure 7. To be more objective and assess if this visual observation is true, we compute the mean color value of the original image and the mean values of the generated images of the seven emotion classes. We measure the root mean square error between the original mean color and the means of the mean colors of the seven generated images on the whole test set (5000 faces) for each GAN in Table II. We observe that the error is really lower in our case and that there is a clear difference on the blue channel between discreteGAN and continuous GANs. These observations are in line with reconstruction losses (cycle-consistency loss with L1 norm, as in the original StarGAN [5]) obtained by the different GANs. Nevertheless, when generating a new expression, the observed color change may be explained by a bias learned by the model. For instance, negative emotions are often associated with a darker context and some expressions may imply a color modification, such as teeth showing during a smile. Finally, the last column of Table II reports the L_{rec} evaluated on the test set and objectively shows that faces are better preserved with our approach.

C. Continuous Sampling of the 3-d Space of Emotions

We are now focusing on the ability of the different generators to deal with smooth transitions between expressions. To be able to compare the different methods, we choose to evaluate the transitions on arousal and valence axes, which

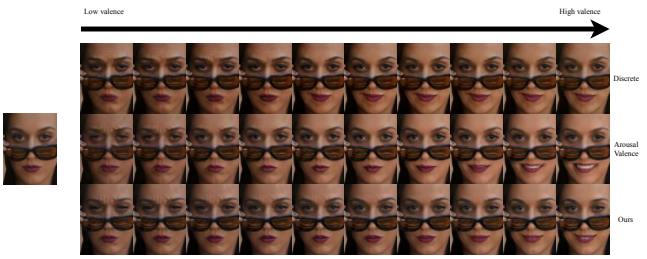

Fig. 6. Generation of expression along the valence axis, from displeasure to pleasure. First row is discreteGAN using the same approach as in [7], second row is avGAN and third row is ours, using a linear regression to find a similar axis to valence. The input image is on the left.

Fig. 7. Generation of expression along the arousal axis, from displeasure to pleasure. First row is discreteGAN using the same approach as in [7], second row is avGAN and third row is ours, using a linear regression to find a similar axis to arousal. The input image is on the left.

are easy to interpret and often used in the psychological community [31]. For avGAN, the sampling of the space is straightforward as the two coordinates of the expressions are arousal and valence.

To generate continuous transition with the discreteGAN, we represent emotions with a one hot vector and create variations between two one hot vectors, as proposed by [7], who used this idea to vary "in intensity" between neutral class and another emotion. For the valence axis, we choose a transition between sadness (valence equals -1 and arousal close to 0) and happiness (valence equals 1 and arousal close to 0), while for the arousal axis, the transition is between neutral (arousal and valence to 0) and surprise (valence close to 0 and arousal equals 1).

From Figure 6, we first can observe that all the generated faces respect the valence axis, expressing displeasure at the extreme left and pleasure at the extreme right. We note that the expression from one GAN to another are not exactly similar, which may also be explained by the fact that the chosen axis for discreteGAN and for our GAN are not perfectly fitting the valence axis.

Another important point is about the smoothness of the transition. Looking at the first line of Figure 6, we observe four very similar expressions of displeasure, followed by one or two expressions mixing both displeasure and pleasure and finally five close expressions of pleasure. On the contrary, for both avGAN and ours, the transition is smoother, the expression being modified at each face. So this would mean that the discreteGAN is not able to uniformly fit the axis of valence and tends to generate less variety in the expressions.

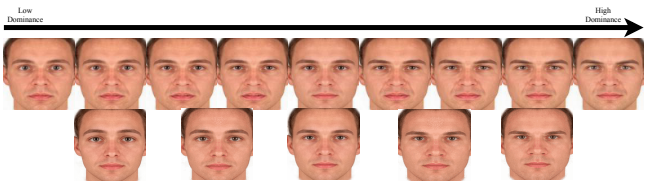

Fig. 8. Illustration of the third dimension found from our representation and used to generate expressions in the first row. Second row is a manual work [39] illustrating what dominance is.

To verify this hypothesis, we propose to use a more objective process. First, we train a ResNet-18 to predict arousal and valence of the faces of AffectNet [24]. Second, we use this model to estimate the valence of the generated faces. Thus, we can plot the estimated valence in function of their targeted valence. We report the mean plots on the whole test set in Figure 5. The avGAN's curve (in orange) should be the identity if both the arousal valence estimator and the avGAN were perfect. It is not the case, but we nevertheless can check that the allure of the curve is coherent with this idea. The curve of our GAN (in green) has a similar allure, validating the smoothness of the expression transition observed on Figure 6. Finally, the discreteGAN's curve (in blue) has an allure which is closer to a step function than to the identity. It is also in line with what has been visually observed and highlights the fact that the discreteGAN is not suited to build a uniformly sampled space of representation.

From Figure 7, we note that all the GANs are able to generate a transition between low and high facial arousal. As observed for the valence axis, the expressions are not totally similar from one GAN to another, as they are not generating expressions exactly on the same axis of arousal. We can observe again that the discreteGAN transition is not very smooth in arousal. To assess this idea, we plot in Figure 5 the estimated arousal as a function of the targeted arousal.

D. Interpreting the Third Dimension

Even if the psychologists' community proposes arousal valence for emotion representation, several works show its limitations. Supplementary dimensions have been proposed and one is especially used: *dominance* [22], which can be seen as a measure of self-confidence. In the previous sections, we show that our representation allows to map back to both discrete emotions and arousal valence. As already observed in Figure 4, it is difficult to distinguish disgust from anger with arousal valence representation, which is not the case with our 3-d representation. The third dimension can therefore bring interesting information. To dig into this idea, we propose to obtain this following representation as previously described (Ea. (6)).

Figure 8-1st row displays generated expressions on the dominance axis. As our corpus is not annotated with the dominance value, we propose to compare our generated expressions to the manually generated expressions proposed by Allen Grabo [39][2]. We observe the same evolution in

[2]https://allengrabo.myportfolio.com/shifting-personality

the facial expressions, the self-confidence growing from left to right. This a first hint to show that our representation contains the dominance information, which has been learned from discrete labels. More examples may be found on our github repository.

V. Conclusions

This paper proposes a method for facial expression generation based on a specific 3-d emotion space. This continuous representation is the latent space of a neural network trained for discrete emotion recognition. Face generation is done by an image to image translator, based on StarGAN, allowing to modify face images according to targeted expressions given as points in the proposed 3-d space. We qualitatively and quantitatively show that not only the generated faces are visually better but also that the user can control the arousal and the valence of the generated faces. Moreover, the proposed 3-d space has a third dimension, close to the concept of dominance, building a bridge with psychological interpretations of emotion.

References

[1] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool. Covariance pooling for facial expression recognition. In *CVPR Workshop*, 2018.

[2] J. C. Batista, V. Albiero, O. R. Bellon, and L. Silva. Aumpnet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. In *Automatic Face and Gesture Recognition*, 2017.

[3] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library, 2003.

[4] E. Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[5] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CVPR*, 1711, 2018.

[6] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *ICMI*, pages 653–656, 2018.

[7] H. Ding, K. Sricharan, and R. Chellappa. Exprgan: Facial expression editing with controllable expression intensity. *aaai*, 2018.

[8] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 2014.

[9] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *J. of personality and social psychology*, 17(2):124, 1971.

[10] R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[11] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *CVPR*, 2016.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[14] J. Huang, Y. Li, J. Tao, and Z. Lian. Speech emotion recognition from variable-length inputs with triplet loss function. *Interspeech*, pages 3673–3677, 2018.

[15] C. Kervadec, V. Vielzeuf, S. Pateux, A. Lechervy, and F. Jurie. Cake: Compact and accurate k-dimensional representation of emotion. *BMVC Workshop*, 2018.

[16] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *Siggraph*, 2018.

[17] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *NIPS*, 2014.

[18] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *Automatic Face and Gesture Recognition*, 2018.

[19] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*. IEEE, 2017.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshop*, 2010.

[21] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (jaffe) database. In *Automatic Face and Gesture Recognition*, 1998.

[22] A. Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[23] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

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

[25] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ICMI*. ACM, 2015.

[26] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174, 2016.

[27] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, pages 818–833, 2018.

[28] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang. Emotional facial expression transfer from a single image via generative adversarial nets. *Computer Animation and Virtual Worlds*, 29(3-4):e1819, 2018.

[29] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Workshop on Audio/Visual Emotion Challenge*, 2017.

[30] S. Rosenthal, N. Farra, and P. Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, 2017.

[31] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[32] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Interspeech*, 2013.

[33] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, et al. The interspeech 2018 comput. paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. *Interspeech*, 5, 2018.

[34] C. Soladié, N. Stoiber, and R. Séguier. Invariant representation of facial expressions for blended expression recognition on unknown subjects. *CVIU*, 117(11), 2013.

[35] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan. Geometry guided adversarial facial expression synthesis. *ACMM*, 2018.

[36] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson. Generating facial expressions with deep belief nets. In *Transactions on Affective Computing*. IEEE, 2008.

[37] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *CVPR*, 2018.

[38] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition*, pages 839–847, 2017.

[39] M. Van Vugt and A. E. Grabo. The many faces of leadership: an evolutionary-psychology approach. *Current Directions in Psychological Science*, 24(6):484–489, 2015.

[40] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie. An occam's razor view on learning audiovisual emotion recognition with small training sets. *ICMI*, 2018.

[41] V. Vielzeuf, S. Pateux, and F. Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *ICMI*. ACM, 2017.

[42] R. Weber, V. Barrielle, C. Soladié, and R. Séguier. Unsupervised adaptation of a person-specific manifold of facial expressions. *Transactions on Affective Computing*, 2018.

[43] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas. Expression flow for 3d-aware face component transfer. *ACM Transactions on Graphics (TOG)*, 30(4):60, 2011.

[44] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.