The Many Moods of Emotion

Valentin Vielzeuf, Corentin Kervadec, Stéphane Pateux, Frédéric Jurie

The Many Moods 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 to the facial expression generation problem. Building upon the assumption of the psychological community that emotion is intrinsically continuous, we first design our own continuous emotion representation with a 3-dimensional latent space issued from a neural network trained on discrete emotion classification. The so-obtained representation can be used to annotate large in the wild datasets and later used to trained a Generative Adversarial Network.**

We first show that our model is able to map back to discrete emotion classes with a objectively and subjectively better quality of the images than usual discrete approaches. But also that we are able to pave the larger space of possible facial expressions, generating the many moods of emotion. Moreover, two axis in this space may be found to generate similar expression changes as in traditional continuous representations such as arousal-valence. Finally we show from visual interpretation, that the third remaining dimension is highly related to the well-known dominance dimension from psychology.

I. Introduction

Affective computing is a topic of broad interest, finding applications in many fields such as health-care, marketing or human-machine interfaces. Therefore, a great effort has been put in the recognition of emotion across different contents. Indeed, several works propose to analyze facial expressions from images [1], [24], from multimodal videos [6], [17], [28], [40] or from multi-view videos [2], [37]. Other works focus more on sentiment expressed within text [4], [25], [29] or audio [13], [31], [32], finally building a very large and complete set of emotion recognition methods. Nevertheless, some recent works [39] underline that the performance may begin to saturate on the emotion recognition task, because of the nature of the used datasets and of the subjective representation of emotion.

Thus, understanding and manipulating the emotion representation is of tremendous interest to progress towards a more complete affective computing ability. For that, the very definition of the facial expression of the emotion has to be taken into account. The literature comes up with three main definitions. First, Ekman *et al.* [9] proposes discrete emotions, identifying six universal classes of emotion (e.g. "Happy", "Sad" or "Angry"). Later, the arousal-valence system was built by Russell, placing emotions in a 2-d continuous space. Finally, the Facial Action Coding Systems allows to objectively represent facial expression with Action Units (e.g. "raised eyebrow") and thus may be used to infer the emotion.

As the face is one of the main ways of expressing emotion, a branch of the affective computing community proposes to focus on the generation of facial expression,

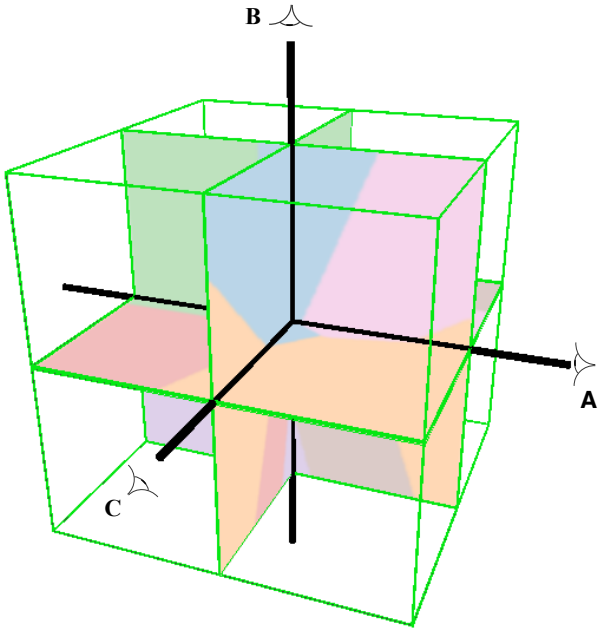

Fig. 1. Illustration of our 3-d representation space of emotion. Better viewed in color.

thus aiming at both a better interpretation of the visual consequences of an emotion representation and a straightforward way to simulate emotion. For generating artificial representation of emotions, first works from computer graphics community focus on the animation of the faces with model-based approaches [15], [33], [41]. More recently deep learning approaches and especially Generative Adversarial Networks [5], [7], [11], [27], [36] have been proposed, often borrowing ideas from the computer graphics community [26], [34] but also aiming to learn these facial expressions from diverse datasets and representations. Nevertheless, these approaches are generally trained on small corpus with pronounced emotions [19], [20], [43], meaning that the space of possible generated expressions is limited.

Even if larger in the wild corpus using other emotion representations such as action units or arousal valence exist [10], [18], [23], they are only used by a few authors [26]. Moreover, the annotation cost of such representation is really higher than for discrete emotion.

Last but but not the least, the described approaches are often focusing on the quality of the generated faces. In this paper, we propose to add another concern, building a bridge between psychological interpretations of the emotion

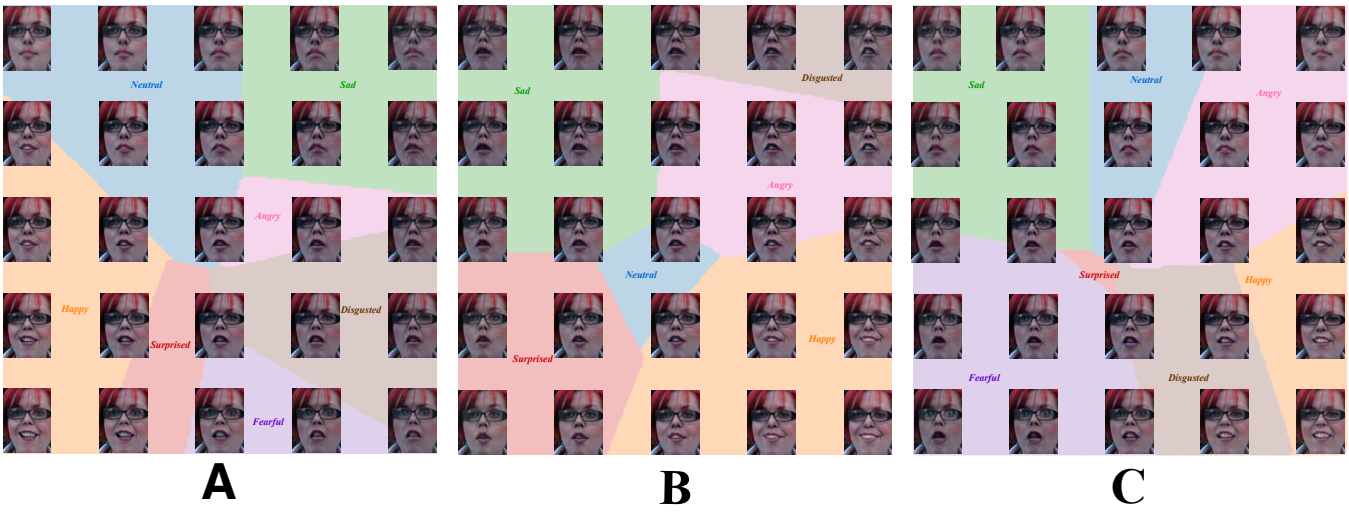

Fig. 2. Illustration of our 3-d representation space of emotion. Each plane is colored with the associated discrete classes of emotions (see position of these planes in Figure 1). Furthermore for each of these planes we illustrate some generated faces from the associated 3D representation coordinates. The generated sample faces within same color areas show the many possible moods 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.

representation and what is visually observed. Our contribution is therefore to propose a 3-dimensional representation of emotion based on the latent space of a classification neural network. This model is trained on discrete emotion classification and thus required less annotation effort than for traditional continuous approaches. Using it to annotate a large in the wild corpus, we then learn a generative adversarial network on the so-obtained dataset. 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 3D space representation that are in line with the common psychological definition of arousal, valence and dominance, enabling easy interpretations of the observed improvements.

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 classes being associated with one word, but at the cost of loosing a lot of power of description, as behind one word many variations may be found. Proposing compound emotion [8] (*e.g.* happily surprise) 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 really fine-grained description. To keep a semantic meaning, several interpretable axis were proposed to build continuous spaces, such as arousal, valence or even dominance [21]. 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 at reaching 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], [42]. More recently, Soladié *et al.* [33] uses a 4-d emotion representation space to animate face and Active-Appearance Models features. In a more general fashion, Weber *et al.* [41] proposes an unsupervised person-specific model which easily adapted to the targeted subject. Finally, hybrid approaches mixing deep learning and model-based method are also proposed. Susskind *et al.* [35] first proposes to train a deep belief network based on both action units and identity information to generate facial expression. More recently, another approach [34] is using fiducial points to geometrically control the face animation while Tulyakov *et al.* [36] is learning to directly generate sequences of images, based on a "content and motion" approach. Quia *et al.* [27] use facial landmarks to improve the animation smoothness of a changing emotion. Kim *et al.* [15] enable to generate video face animation using another portrait video as an example. These approaches are working on the very shape of the face and it therefore implies complex modifications of the model to adapt to "in the wild" conditions, where important illumination changes and occlusions are common.

c) Generative Neural Networks: To fulfill the previous requirement of robustness towards real "in the wild" conditions, an interesting path of research for image syn-

thesis using neural networks is Generative Adversarial Networks(GAN) [11] and Variational AutoEncoders(VAE) [16]. Focusing on GANs, many extensions exist, such as Conditional GAN [22] where a condition variable allows to control the generation or more recently StarGAN, where Choi *et. al* [5] propose a multi-domain approach, 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 intensity of an emotion 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.* [26] 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. Methods

This section describes our proposed approach. First we present how to exhibit a continuous representation of the emotions. The aim of this continuous representation being to enable compound emotions but also other possible variations within a given emotion. Then from this representation we adapt a GAN to allow for continuous emotion editing.

A. Facial Expression Representation

As argued by [30] continuous annotations may be a more subtle and accurate emotion representation than discrete basic emotion. The idea behind is to continuously quantify several features of a facial expression, such as intensity (arousal) or pleasure (valence). A mapping to the discrete representation of emotion may be done [23], [30], enabling to take benefits from both discrete and continuous annotations. Nevertheless, as underlined by psychological study [21], two dimensions might not be sufficient to represent the whole variations of emotions. Moreover, annotation cost is high compared to discrete emotions. But a recent approach [14] shows that a compact latent space issued from intermediate hidden layers of a convolutional neural network trained on discrete emotion classification may lead to an arousal-valence like topology.

We therefore propose to use a latent space of a convolutional neural network. In order to assess and understand the benefit of using additional dimensions in emotion representation [33], we propose to define a 3-d representation from the latent space. For that, we train a modified ResNet-18 [12] to classify discrete emotions, as in Figure 3. The modification consists in adding a bottleneck fully-connected before classification, forcing the classifier to use only three

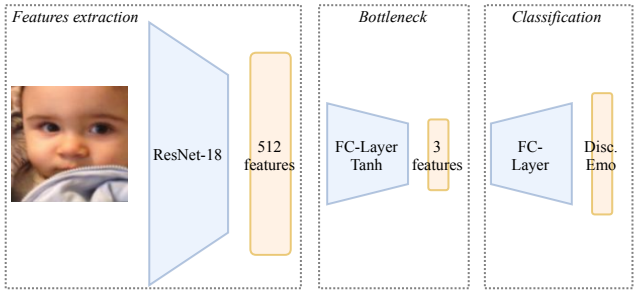

Fig. 3. Generating a 3-d compact representation, following a similar approach to [14]

features to predict the discrete emotion. An hyperbolic tangent activation is applied on these three features to ensure that the 3-d representation range is kept between -1 and 1.

From an initial corpus annotated only in discrete emotions, we train such an aforementioned network. This network is then used to provide continuous annotations to any provided dataset. In our case we only consider the dataset initially annotated in discrete emotion, but we could also have use any additional faces without any need of annotation.

We may observe in the upper left of the Figure 1 how the discrete emotions are associated to the 3-d representation in our representation space and especially on three planes cutting it. The many variations observed in these planes are also reported in the Figure 2 as examples of the diversity of possible generated expressions.

B. Facial Expression Generation

To achieve facial expression generation, we choose to build a Generative Adversarial Neural Network (GAN). Among the various GAN approaches, we retain the StarGAN [5] architecture, which allows to take both the face and the targeted emotion as input of the generator and has already prove to be efficient on discrete emotion generation. As a reminder, the model is composed of a discriminator D and a generator G. Thus, as in the original approach, we use a loss composed with different terms. Nevertheless, we need to adapt them to the continuous labeling case. We therefore rewrite the following terms.

The *adversarial loss* aims at making the generated fake expressions not distinguishable from real facial expressions. It may be written following:

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

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

The usual classification loss, which we denote here as a *regression loss*, is itself composed of two terms. The first term, namely L_{reg}^{real}, forces D to correctly regress the ground truth associated to the real image. While the second term, namely L_{reg}^{fake}, forces G to generate facial expressions with

a representation close to the target. More formally, we use a Mean Squared Error term:

$$L_{reg}^{real} = \mathbb{E}_{x,r}[D(x) - r]^2 \tag{2}$$

$$L_{reg}^{fake} = \mathbb{E}_{x,r}[D(G(x,r)) - r]^2 \tag{3}$$

The *reconstruction loss*, namely L_{rec} ensures that the generated faces preserved contents not relative to the expression. It is defined as:

$$L_{rec} = \mathbb{E}_{x,r_1,r_2}[||x - G(G(x,r_2),r_1)||_1] \tag{4}$$

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 can then write the generator and discriminator losses as:

$$L_D = -L_{adv} + \lambda_{reg} L_{reg}^{real} \tag{5}$$

$$L_G = L_{adv} + \lambda_{reg} L_{reg}^{fake} + \lambda_{rec} L_{rec} \tag{6}$$

For efficient implementation, the Equation 1 can be reformulated with a Wasserstein GAN objective with gradient penalty, as done in the original paper [5].

C. Implementation Details

To build the training set, we use the recent AffectNet dataset [23] which provides both discrete emotions and arousal valence annotations. We sanitize the dataset by discarding samples where there is no face or where the provided annotation does not exist. The resulting practical dataset then consists of 297000 annotated faces recorded in the wild. We preprocess them, using 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 augmentation transformations (scale jittering, rotation and flip).

We train (on the training set) our modified ResNet-18 and thus obtain our 3-d representation for each face. Then, we train three GANs (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]. In the case of continuous annotations, the regression loss ponderation λ_{reg} is changed to 3 instead of 1 because of the scale difference with a classification cross-entropy loss. The other ponderations are the same as in [5]. The parameters are optimized with the Adam method during 300000 iterations.

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

IV. Results

This section evaluates the benefits of the proposed continuous approach. We first detail our experimental protocol and then present our results. We study the ability of our approach to be applied to discrete emotion generation, then we enlighten the relations between our learned representation and the arousal valence representation. Finally, we build a bridge towards psychological interpretations, finding back a third dimension visually similar to the dominance [21].

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

A. Discrete Results

We propose to evaluate the impact of the different representations on the quality of the generated expressions. Therefore, the same previously described GAN architecture is used – we are not comparing different architectures of GAN – and we only vary on the use of emotion representation. For that, we first generate the seven discrete emotions, using the discreteGAN as baseline. Then, we choose to compute the coordinates of the centroid in the continuous representation spaces for each emotion class. We thus simply generate the expression associated to these centroids and we label them with the emotion classes. More formally,

$$C_{continuous}^{i} = \sum_{k \in C_{discrete}^{i}} \frac{r_k}{\#C_{discrete}^{i}} \tag{7}$$

where $C_{continuous}^{i}$ 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 continuous representation of the sample k. As we

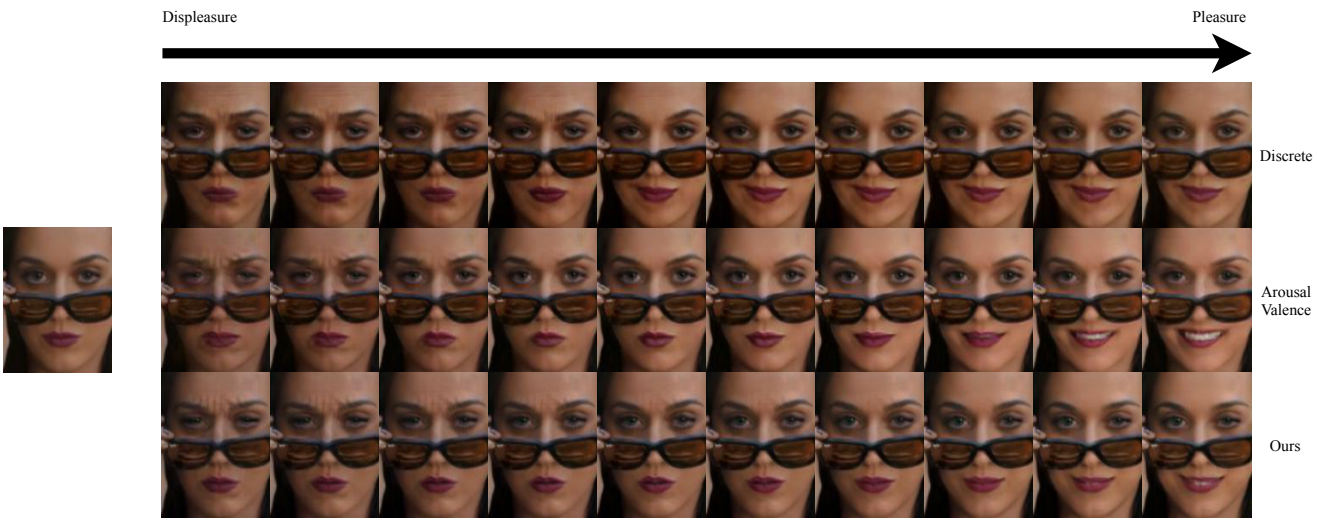

Fig. 5. Generation of expression along the valence axis, from displeasure to pleasure. First row is discreteGAN using the same approach as in [7], second row is avGAN and third row is ours, using a linear regression to find a similar axis to valence. The input image is on the left.

can observe in the column of Figure 4, the generated emotion classes are relatively similar for all GANs. Nevertheless, we may denote some interesting differences.

For the happy class (second column), the continuous models tend to add teeth to the faces, with a better success for our approach than for avGAN (some artifacts are visible for the second face in the case of avGAN).

In the case of the disgust class, we note the presence of artifacts in the faces generated by the discreteGAN, while avGAN tends to generate relatively similar expression for both disgusted and angry faces. The artifacts may be explained by the small number of disgust occurrences in the training dataset (less than 2%), meaning that the discreteGAN did not see many examples of this class. Furthermore disgust and anger classes are relatively closed in arousal valence space, leading to very similar values for their centroids and thus very similar generated expressions. Our GAN, using a third dimension as shown in Figure 1, allows to improve the separation of these two classes and thus leads to a real difference between the generated faces.

The neutral class is also interesting, especially for the third 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 may be explained by the fact that we choose the centroids for continuous representation, which thus are not extreme samples of these classes.

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 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 mean of the mean colors of the seven generated images on the whole test set (5000 faces) for each GAN in Table I. 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 to a darker context and some expressions may imply a color modification, such as teeth showing during a smile. Finally, the last column of Table I reports the L_{rec} evaluated on the test set for the different GANs and objectively show that the face is better preserved in the case of our GAN.

	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 I

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

B. Continuous Results

We are now focusing on the ability of the different methods to generate transitions between different expressions. To be able to compare the different methods, we choose to evaluate the transitions on arousal and valence axes, which are easy to interpret and often used in the psychological community [30]. For avGAN, the transitions is therefore straightforward, we only need to browse through the different values of one dimension. To generate continuous transition with the discreteGAN, we encode the emotion in a one hot vector and

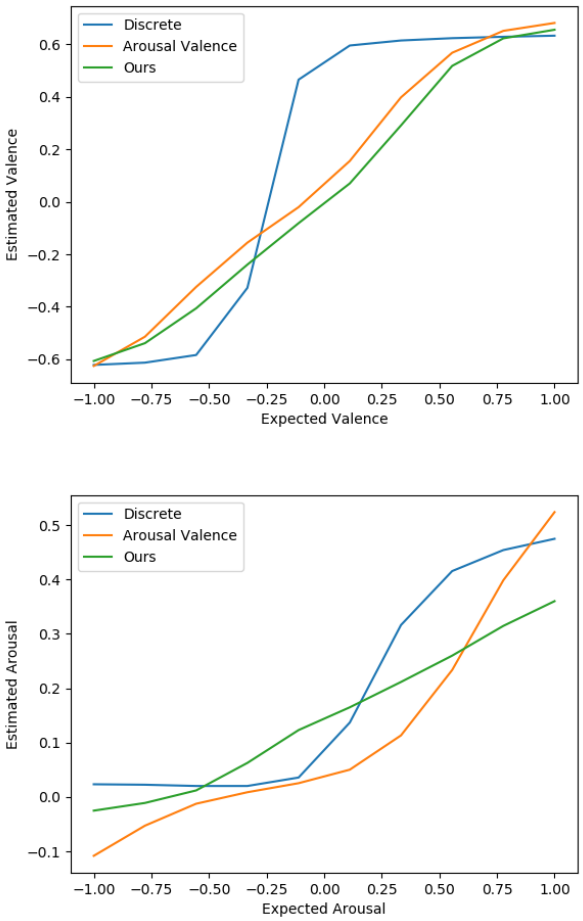

Fig. 6. Plots of respectively the estimated valence (up) and arousal (down) of the generated faces as a function of the targeted valence (up) and arousal (down) by the model for the different approaches. This plots are average plots on the whole test set. Better viewed in colors.

therefore create variations between two one hot vectors, as previously proposed by [7] who are using this concept 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)[1]. For our GAN, we find the 3-d coordinates of the transition axis by applying a linear regression between our representation and arousal valence.

From Figure 5, 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 also can 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.

[1]so it means there that we are not ranging in the complete [-1,1] interval but in [0,1])

Another important point is about the smoothness of the transition. Looking at the first line of Figure 5, 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. In the contrary, for both avGAN and ours, the transitions 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.

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 [23]. 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 6. 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 5. 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 highlight 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 from a not excited face to a really aroused expression. 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 also observe again that the discreteGAN transitions is not very smooth in arousal. To assess this idea, we apply the same process used for valence and plot in Figure 6 the estimated arousal as a function of the targeted arousal. This plot first enlightens that the discreteGAN (blue), as for the valence, is not fitting an identity function. We also need to note the range of the estimated axis: the arousal values should be between -1 and 1 and are between -0.1 and 0.6 in the best case (avGAN). This may be explained by the fact that there are almost no sample with a negative arousal in the training set, leading both the estimator and the GANs to be inefficient on negative values.

C. Interpreting the Third Dimension

Even if the psychologists' community proposes arousal valence for emotion representation, several works show the limitations of using only two dimensions. Therefore, supplementary dimensions have been proposed and one is especially used. It is called the dominance [21] and may 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 may therefore brings interesting information. To dig into this idea,

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. Better viewed in color.

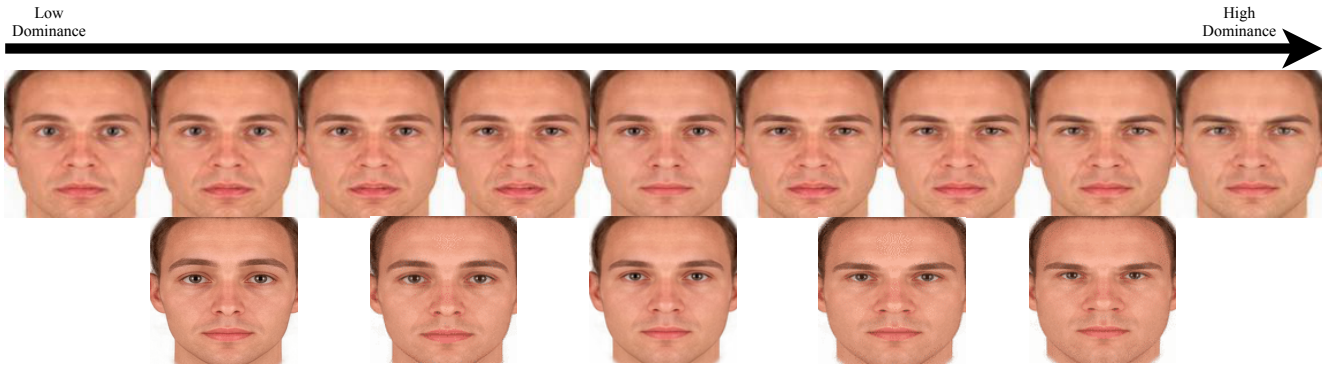

Fig. 8. Illustration of the third dimension found from our representation and used to generate expressions in the first row. Second row is a manual work [38] illustrating what dominance is.

we propose to take the two directions of arousal and valence in our 3-d representation that we previously regress. From the two so-obtained vectors, we compute a third orthogonal vector (by vector product) and generate the expressions along this new axis.

The Figure 8 displays in the first row the generated expression on the dominance axis. As our corpus is not annotated with the dominance value, we propose to compare our generated expressions to manually generated expressions proposed by Allen Grabo [38] [2] (second line). Even if the reconstruction is not perfect, we visually can see the same evolution in the facial expression, 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.

V. Conclusion

We propose a solution to the facial expression generation based on a specific emotion representation containing 3 dimensions. This continuous representation is obtained from the constrained latent space of a neural network trained on the discrete emotion classification task. The so-obtained neural network can be used to annotate every face corpus such as in the wild datasets, allowing to learn from continuous representation. We therefore train a Generative Adversarial Network with these annotations. For that, we modify a well-known StarGAN architecture to fit the requirement of a regression approach. The obtained generated faces are compared to the same architecture trained in the same conditions but with other representations. We show qualitatively and quantitatively that not only our generated faces have a better reconstruction quality than a GAN trained on discrete emotion, but also that we are able to uniformly fit arousal and valence axis, as a GAN that would have been trained on real arousal valence labels. Moreover, we exhibit a third dimension close to the concept of dominance, building a bridge with psychological interpretations of emotion.

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

References

[1] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool. Covariance pooling for facial expression recognition. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition Workshops, pages 367–374, 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 & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 866–871. IEEE, 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. *arXiv preprint*, 1711, 2017.

[6] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 653–656. ACM, 2018.

[7] H. Ding, K. Sricharan, and R. Chellappa. Exprgan: Facial expression editing with controllable expression intensity. *arXiv preprint arXiv:1709.03842*, 2017.

[8] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, page 201322355, 2014.

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

[10] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[14] C. Kervadec, V. Vielzeuf, S. Pateux, A. Lechervy, and F. Jurie. Cake: Compact and accurate k-dimensional representation of emotion. *arXiv preprint arXiv:1807.11215*, 2018.

[15] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *arXiv preprint arXiv:1805.11714*, 2018.

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

[17] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. *arXiv preprint arXiv:1711.04598*, 2017.

[18] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 2584–2593. IEEE, 2017.

[19] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[20] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.

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

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

[23] 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.

[24] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM, 2015.

[25] 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*, 174:50–59, 2016.

[26] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[27] 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.

[28] 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 *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9. ACM, 2017.

[29] 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.

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

[31] 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 *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[32] 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 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. *Proceedings of INTERSPEECH, Hyderabad, India*, 5, 2018.

[33] C. Soladié, N. Stoiber, and R. Séguier. Invariant representation of facial expressions for blended expression recognition on unknown subjects. *Computer Vision and Image Understanding*, 117(11):1598–1609, 2013.

[34] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan. Geometry guided adversarial facial expression synthesis. *arXiv preprint arXiv:1712.03474*, 2017.

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

[36] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. Mocogan: Decomposing motion and content for video generation. *arXiv preprint arXiv:1707.04993*, 2017.

[37] 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 & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 839–847. IEEE, 2017.

[38] 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.

[39] 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. *arXiv preprint arXiv:1808.02668*, 2018.

[40] V. Vielzeuf, S. Pateux, and F. Jurie. Temporal multimodal fusion for video emotion classification in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 569–576. ACM, 2017.

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

[42] 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.

[43] 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.