Learning to Generate Customized Dynamic 3D Facial Expressions

Rolandos Alexandros Potamias, Jiali Zheng, Stylianos Ploumpis, Giorgos Bouritsas, Evangelos Ververas, and Stefanos Zafeiriou

Department of Computing, Imperial College London, UK
{r.potamias, jiali.zheng18, s.ploumpis, g.bouritsas, e.ververas16, s.zafeiriou}@imperial.ac.uk

Abstract. Recent advances in deep learning have significantly pushed the state-of-the-art in photorealistic video animation given a single image. In this paper, we extrapolate those advances to the 3D domain, by studying 3D image-to-video translation with a particular focus on 4D facial expressions. Although 3D facial generative models have been widely explored during the past years, 4D animation remains relatively unexplored. To this end, in this study we employ a deep mesh encoder-decoder like architecture to synthesize realistic high resolution facial expressions by using a single neutral frame along with an expression identification. In addition, processing 3D meshes remains a non-trivial task compared to data that live on grid-like structures, such as images. Given the recent progress in mesh processing with graph convolutions, we make use of a recently introduced learnable operator which acts directly on the mesh structure by taking advantage of local vertex orderings. In order to generalize to 4D facial expressions across subjects, we trained our model using a high resolution dataset with 4D scans of six facial expressions from 180 subjects. Experimental results demonstrate that our approach preserves the subject's identity information even for unseen subjects and generates high quality expressions. To the best of our knowledge, this is the first study tackling the problem of 4D facial expression synthesis.

Keywords: Expression Generation, Facial Animation, 4D synthesis, 4DFAB, Graph Neural Networks

1 Introduction

Recently, facial animation has received attention from the industrial graphics, gaming and filming communities. Face is capable to impart a wide range of information not only about the subject's emotional state but also about the tension of the moment in general. An engaged and crucial task is 3D avatar animation, which has lately become feasible [34]. With modern technology, a 3D avatar can be generated by a single uncalibrated camera [9] or even by a self portrait image [33]. At the same time, capturing facial expression is an important task in order to perceive behaviour and emotions of people. To tackle the problem of facial expression generation, it is essential to understand and model facial

muscle activations that are related to various emotions. Several studies have attempted to decompose facial expressions on two dimensional spaces such as images and videos [15, 31, 44, 50]. However, modeling facial expressions on high resolution 3D meshes remains unexplored.

In contrast, few studies have attempted 3D speech-driven facial animation exclusively based on vocal audio and identity information [13, 21]. Nevertheless, emotional reactions of a subject are not always expressed vocally and speech-driven facial animation approaches neglect the importance of facial expressions. For instance, sadness and happiness are two very common emotions that can be voiced, mainly, through facial deformations. To this end, facial expressions are a major component of entertainment industry, and can convey emotional state of both scene and identity.

People signify their emotions using facial expressions in similar manners. For instance, people express their happiness by mouth and cheek deformations, that vary according to the subject's emotional state and characteristics. Thus, one can describe expressions as "unimodal" distributions [15], with gradual changes from the neutral model till the apex state. Similarly to speech signals, emotion expressions are highly correlated to facial motion, but lie in two different domains. Modeling the relation between those two domains is essential for the task of realistic facial animation. However, in order to disentangle identity information and facial expression it is essential to have a sufficient amount of data. Although most of the publicly available 3D datasets contain a large variety of facial expression, they are captured only from a few subjects. Due to this difficulty, prior work has only focused on generating expressions in 2D.

Our aim is to generate realistic 3D facial animation given a target expression and a static neutral face. Synthesis of facial expression generation on new subjects can be achieved by expression transfer of generalized deformations [39, 50]. In order to produce realistic expressions, we map and model facial animation directly on the mesh space, avoiding to focus on specific face landmarks. Specifically, the proposed method comprises two parts: (a) a recurrent LSTM encoder to project the expected expression motion to an expression latent space, and (b) a mesh decoder to decode each latent time-sample to a mesh deformation, which is added to the neutral expression identity mesh. The mesh decoder utilizes intrinsic lightweight mesh convolutions, introduced in [6], along with unpooling operations that act directly on the mesh space [35]. We train our model in an end-to-end fashion on a large scale 4D face dataset. The devised methodology tackles a novel and unexplored problem, i.e. the generation of 4D expressions given a single neutral expression mesh. Both the desired length and the target expression are fully defined and controlled by the user. Our work considerably deviates from methods in the literature as it can be used to generate 4D full-face customised expressions on real-time. Finally, our study is the first 3D facial animation framework that utilizes an intrinsic encoder-decoder architecture that operates directly on mesh space using mesh convolutions instead of fully connected layers, as opposed to [13, 21].

2 Related Work

Facial animation generation Following the progress of 3DMMs, several approaches have attempted to decouple expression and identity subspaces and built linear [2, 5, 45] and nonlinear [6, 25, 35] expression morphable models. However, all of the aforementioned studies are focused on static 3D meshes and they cannot model 3D facial motion. Recently, a few studies attempted to model the relation between speech and facial deformation for the task of 3D facial motion synthesis. Karras et al. [21] modeled speech formant relationships with 5K vertex positions, generating facial motion from LPC audio features. While this was the first approach to tackle facial motion directly on 3D meshes, their model is subject specific and cannot be generalized across different subjects. Towards the same direction, in [13], facial animation was generated using a static neutral template of the identity and a speech signal, used along with DeepSpeech [20] to generate more robust speech features. A different approach was utilized in [32], where 3D facial motion is generated by regressing on a set of action units, given MFCC audio features processed by RNN units. However, their model is trained on parameters extracted from 2D videos instead of 3D scans. Tzirakis et al. [41] combined predicted blendshape coefficients with a mean face to synthesize 3D facial motion from speech, replacing also fully connected layers, utilized in previous studies, with an LSTM. Blendshape coefficients are also predicted from audio, using attentive LSTMs in [40]. In contrast with the aforementioned studies, the proposed method aims to model facial animations directly on 3D meshes. Furthermore, although blendshape coefficients might be easily modeled, they rely on predefined face rigs, a factor that limits their generalization to new unseen subjects.

Geometric Deep Learning Recently, the enormous amount of applications related to data residing in non-Euclidean domains motivated the need for the generalization of several popular deep learning operations, such as convolution, to graphs and manifolds. The main efforts include the reformulation of regular convolution operators in order to be applied on structures that lack consistent ordering or directions, as well as the invention of pooling techniques for graph downsampling. All relevant endeavours lie within the new research area of Geometric Deep Learning (GDL) [7]. The first attempts defined convolution in the spectral domain, by applying filters inspired from graph signal processing techniques [37]. These methods mainly boil down to either an eigendecomposition of a Graph Shift Operator (GSO) [8], such as the graph Laplacian, or to approximations thereof, by using polynomials [14, 23] or rational complex functions [24] of the GSO in order to obtain strict spatial localization and reduced computational complexity. Subsequent attempts generalize conventional CNNs by introducing patch operators that extract spatial relations of adjacent nodes within a local patch. To this end, several approaches generalized local patches to graph data, using geodesic polar charts [27] anisotropic diffusion operators [4] on manifolds or graphs [3]. MoNet [28], generalized previous spatial approaches

by learning the patches themselves with Gaussian kernels. In the same direction, SplineCNN [17] replaced Gaussian kernels with B-spline functions with significant speed advantage. Recent studies focused on soft-attention methods to weight adjacent nodes [42, 43]. However, in contrast to regular convolutions, the way permutation invariance is enforced in most of the aforementioned operators, inevitably renders them unaware of vertex correspondences. To tackle that, Bouritsas et al. [6] defined local node orderings, instantiated with the spiral operator of [26], by exploiting the fixed underlying topology of certain deformable shapes, and built correspondence-aware anisotropic operators.

Facial Expression datasets Another major reason that 4D generative models have not been widely exploited is due to the limited amount of 3D datasets. During the past decade, several 3D face databases have been published. However, most of them are static [19,29,36,38,49], consisted of few subjects [10,12,35,46], and have limited [1,16,36] or spontaneous expressions [47,48], making them inappropriate for tasks such as facial expression synthesis. On the other hand, the recently proposed 4DFAB dataset [11] consists of six 3D dynamic facial expressions (from 180 subjects), which is ideal for subject independent facial expression generation. In contrast with all previously mentioned datasets, 4DFAB, due to the high range of subjects, can be a promising resource towards disentangling facial expression from the identity information.

3 Learnable Mesh Operators: Background

3.1 Spiral Convolution Networks

We define a 3D facial surface discretized as triangular mesh $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{F})$ with $\mathcal{V}$ the set of N vertices, $\mathcal{E}$ and $\mathcal{F}$ the sets of edges and faces, respectively. Let also, $X \in \mathbb{R}^{N \times d}$ denote the feature matrix of the mesh. In contrast to regular domains, when attempting to apply convolution operators on graph-based structures, there does not exist a consistent way to order the input coordinates. However, in a *fixed topology* setting, such an ordering is beneficial so as to be able to keep track of the existing correspondences. In [6], the authors identified this problem and intuitively order the vertices by using spiral trajectories [26]. In particular, given a vertex $v \in \mathcal{V}$, we can define a *k-ring* and *k-disk* as:

$$\begin{aligned} ring^{(0)}(v) &= v, \\ ring^{(k+1)}(v) &= \mathcal{N}(ring^{(k)}(v)) - disk^{(k)}(v), \\ disk^{(k)}(v) &= \bigcup_{i=0,\dots,k} ring^{(i)}(v) \end{aligned} \tag{1}$$

where $\mathcal{N}(S)$ is the set of all vertices adjacent to at least one vertex $\in S$.

Once the $ring^{(k)}$ is defined, the spiral trajectory centered around vertex v can be defined as:

$$S(v,k) = \{ring^{(0)}(v), ring^{(1)}(v), ..., ring^{(k)}(v)\} \tag{2}$$

To be consistent across all vertices, one can pad or truncate S(v,k) to a fixed length L. To fully define the spiral ordering, we have to declare the starting direction and the orientation of a spiral sequence. In the current study, we adopt the settings followed in [6], by selecting the initial vertex of $S(v,k)$ to be in the direction of the shortest geodesic distance between a static reference vertex. Given that all 3D faces share the same topology, spiral ordering $S(v,k)$ will be the same across all meshes and so, their calculation is done only once. With all the above mentioned, *Spiral Convolution* can be defined as:

$$\mathbf{f}_v^* = \sum_{j=0}^{|S(v,k)|-1} \mathbf{f}(S_j(v,k))\mathbf{W}_j \tag{3}$$

where $|S(v,k)|$ amounts to the total length of the spiral trajectory, $\mathbf{f}(S_j(v,k))$ are the d-dimensional input features of the jth vertex of the spiral trajectory, $\mathbf{f}^*$ the respective output, and $\mathbf{W}_j$ are the filter weights.

3.2 Mesh Unpooling Operations

In order to let our graph convolution decoder to generate faces sampled from a latent space, it is essential to use unpooling operations in analogy with transposed convolutions in regular settings. Each graph convolution is followed by an upsampling layer which acts directly on the mesh space, by increasing the number of vertices. We use sampling operations introduced in [35], based on sparse matrix multiplications with upsampling matrices $Q_u \in \{0,1\}^{n\times m}$, where $m > n$. Since upsampling operation changes the topology of the mesh, and in order to retain the face structure, upsampling matrices Q_u are defined on the basis of down-sampling matrices. The barycentric coordinates of the vertices that were discarded during downsampling procedure are stored and used as the new vertex coordinates of the upsampling matrices.

4 Model

The overall architecture of our model is structured by two major components (see Figure 1). The first one contains a temporal encoder, using an LSTM layer that encodes the expected facial motion of the target expression. It takes as input a temporal signal $e \in R^{6\times T}$ with length T, equal to the target facial expression, also equipped with information about the time-stamps that show when the generated facial expression should reach onset, apex and offset modes. Each time-frame of signal e can be characterised as a one-hot encoding of one of the six expressions, with amplitude that indicates the scale of the expression. The second component of our network consists of a frame decoder, with four layers of mesh convolutions, where each one is followed by an upsampling layer. Each upsampling layer increases the number of vertices by five times, and every mesh convolution is followed by a ReLU activation [30]. Finally, the output of the decoder is added to the identity neutral face. Given a time sample from the

latent space the frame decoder network models the expected deformations on the neutral face. Each output time frame can be expressed as:

$$\begin{aligned} \hat{x}_t &= D(z_t) + x_{id}, \\ z_t &= E(e_t) \end{aligned} \tag{4}$$

where $D(\cdot)$ denotes the mesh decoder network, $E(\cdot)$ the LSTM encoder, e_t the facial motion information for time-frame t and x_{id} the neutral face of the identity. The network details can be found in Table 1. We trained our model for 100 epochs with learning rate of 0.001 and a weight decay of 0.99 on every epoch. We used Adam optimizer [22] with a 5e-5 weight decay.

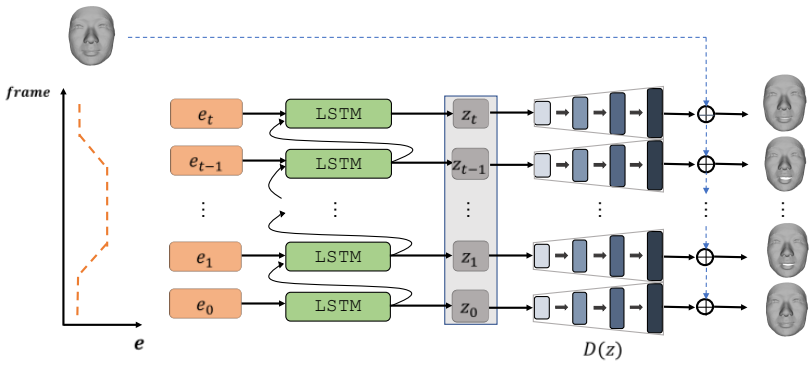

Fig. 1. Network architecture of the proposed method.

Loss function. The mesh decoder network outputs motion deformation for each time-frame with respect to the expected facial animation. To train our model we minimize both the reconstruction error L_r and the temporal coherence L_c, as proposed in [21]. Specifically, we define our loss function between the generated time frame $\hat{x}_t$ and its ground truth x_t value as:

$$\begin{aligned} L_r(\hat{x}_t, x_t) &= \|\hat{x}_t - x_t\|_1 \\ L_c(\hat{x}_t, x_t) &= \|(\hat{x}_t - \hat{x}_{t-1}) - (x_t - x_{t-1})\|_1 \\ L(\hat{x}_t, x_t) &= L_r(\hat{x}_t, x_t) + L_c(\hat{x}_t, x_t) \end{aligned} \tag{5}$$

Although reconstruction loss L_r term can be sufficient to encourage model to match ground truth vertices at each time step, it does not produce high-quality realistic animation. On the contrary, temporal coherence loss L_c term ensures temporal stability of the generated frames by matching the distances between consecutive frames on ground truth and generated expressions.

Table 1. Mesh Decoder architecture

Layer	Input Dimension	Output Dimension
Fully Connected	64	46x64
Upsampling	46x64	228x64
Convolution	228x64	228x32
Upsampling	228x32	1138x32
Convolution	1138x32	1138x16
Upsampling	1138x16	5687x16
Convolution	5687x16	5687x8
Upsampling	5687x8	28431x8
Convolution	28431x8	28431x3

5 Experiments

5.1 Dynamic 3D face database

To train our expression generative model we use the recently published 4DFAB [11]. 4DFAB contains dynamic 3D meshes of 180 people (60 females, 120 males) with ages between 5 to 75 years. The devised meshes display a variety of complex and exaggerated facial expressions, namely *happy, sad, surprise, angry, disgust* and *fear*. The 4DFAB database displays high variance in terms of ethnicity origins, including subjects from more than 30 different ethnic groups. We split the dataset into 153 subjects for training and 27 for testing. The data were captured with 60fps, thus each expression is sampled every approximately 5 frames in order to allow our model to generate extreme facial deformations. Given the high quality of the data (each mesh is composed by 28K vertices) as well as the relatively big number of subjects, 4DFAB presents a rich and rather challenging choice for training generative models.

Expression Motion Labels In this study, we rely on the assumption that each expression can be characterised by four phases of its evolution (see Figure 2). First, the subject starts from a neutral pose and at a certain point their face starts to deform, in order to express their emotional state. We call this the *onset phase*. After the subject's expression reaches its *apex state*, it will start again its deformation from the peak emotional state until it becomes neutral again. We call this the *offset phase*. Thus, each time frame is assigned a label that reflects its emotional state phase. We consider the emotional state as a value ranging from 0 to 1 assigned to each frame, with 0 representing the neutral phase and 1 the apex phase. Onset and offset phases are represented via a linear interpolation between the apex and neutral phases (see Figure 2). However, expressions may also range in terms of extremeness, i.e. the level of intensity in subject's expression. To let our model learn diverse extremeness levels for each expression, it is essential to scale each expression motion label from $[0, 1]$ to $[0, s_i]$, where $s_i \in (0, 1]$ represents the scaled value of the apex state according to the intensity of the expression. Intuitively, the extremeness of each expression is proportional to the absolute

mean deformation of the expression, we can thus calculate scaling factor s_i as:

$$s_i = \frac{clip(\frac{m_i - \mu_e}{\sigma_e}) + 1}{2} \tag{6}$$

where m_i is a scalar value representing the absolute value of the mean deformation of the sequence from neutral frame and μ_e, σ_e the mean and standart deviation of the deformation of the respective expression. Clip() function is used to clip values to [-1,1].

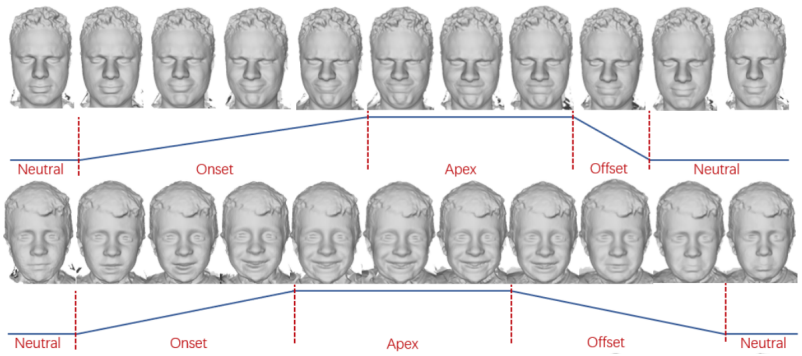

Fig. 2. Sample subjects from the 4DFAB database posing an expression along with expression motion labels.

5.2 Dynamic Facial Expressions

The proposed model for the generation of facial expressions is assessed both qualitatively and quantitatively. The model is trained by feeding the neutral frame of each subject and the manifested motion (i.e. the time-frames where the expression reaches onset, apex and offset modes) of the target expression. We evaluated the performance of the proposed model by its ability to generate expressions of 27 unobserved test subjects. To this end, we calculated the reconstruction loss as the per-vertex Euclidean distance between each generated sample and its corresponding ground truth.

Baseline. As a comparison baseline we implemented an expression blendshape decoder that transforms the latent representation z_t of each time-frame, i.e. the LSTM outputs, to an output mesh. In particular, expression blendshapes were modeled by first subtracting the neutral face of each subject to its corresponding expression, for all the corresponding video frames. With this operation we are able to capture and model just the motion deformation of each expression. Then, we applied Principal Component Analysis (PCA) to the motion deformations to reduce each expression to a latent vector. For a fair comparison, we use the same latent size for both the baseline and the proposed method.

The results presented in Table 2 show that the proposed model outperforms the baseline with regards to all the expressions, as well as on the entire dataset (0.39mm vs 0.44mm).

Table 2. Generalization per-vertex loss over all expressions, along with the total loss.

Model	*Happy*	*Angry*	*Sad*	*Surprise*	*Fear*	*Disgust*	Total
Baseline	0.49	0.37	0.37	0.48	0.43	0.45	0.44
Proposed	**0.37**	**0.35**	**0.36**	**0.43**	**0.42**	**0.42**	**0.39**

Moreover, as can be seen in Figure 3, the proposed method can produce more realistic animation, especially in the mouth region, compared to PCA blendshapes. Error visualizations in Figure 3, show that the blendshape model produces mild transitions between each frame and cannot generate extreme deformations. Note also that the proposed method errors are mostly centered around the mouth and the eyebrows, due to the fact that our model is subject independent and each identity expresses its emotions in different ways and varying extents. In other words, the proposed method models each expression with respect to the motion labels without taking into account identity information, thus the generated expressions can have some variations compared to the ground truth subject-dependent expression.

For a subjective qualitative assessment, Figure 4 shows several expressions that were generated by the proposed method. Due to the fact that several expressions, such as angry and sad expression, mainly relate to eyebrow deformations can not be easily visualised by still images we encourage the reader to check our supplementary material for more qualitative results.

5.3 Classification of generated 4D expressions

To further assess the quality of the generated expressions we implemented a classifier trained to identify expressions. The architecture of the classifier is based on a projection on a PCA space followed by three fully connected layers. The sequence classification is performed in two steps. In particular, first we computed 64-PCA coefficients to represent all expression and deformation variations of the training set and then we used them as a frame encoder to map the unseen test data to a latent space. Following that, the latent representations of each frame are concatenated and processed by fully-connected layers in order to predict the expression of the given sequence. The network was trained on the same training set that was originally used for the generation of expressions, with Adam optimizer and 5e-3 weight decay for 13 epochs. Table 3 shows the achieved classification performance for the ground truth test data and the generated data from the proposed model and from the baseline model, respectively.

As can be seen in Table 3, the generated data from the proposed model achieve similar classification performance with ground truth data across almost

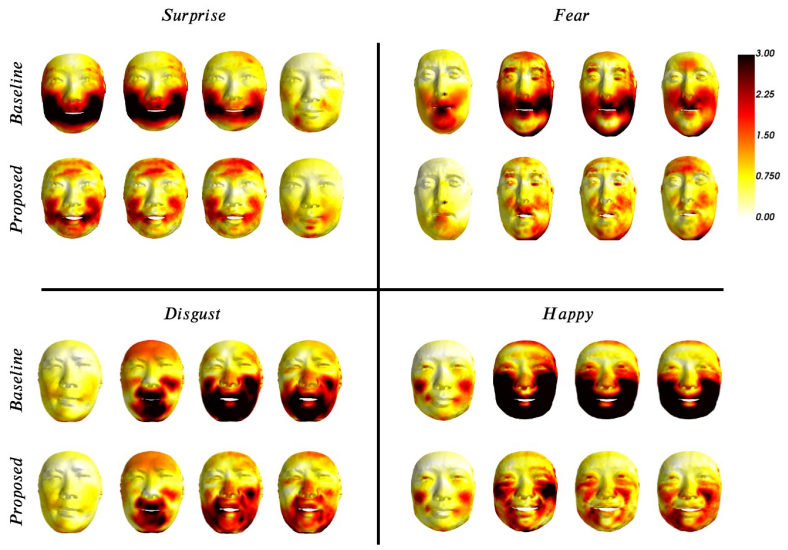

Fig. 3. Color heatmap visualization of error metric of both baseline (top rows) and proposed (bottom rows) model against the ground truth test data for four different expressions.

Table 3. Constructing classification performance between ground truth (test) data, generated (by the proposed method) data, and the blendshape baseline.

	Baseline			**Proposed**			Ground Truth		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Surprise	0.69	0.62	0.65	**0.74**	**0.90**	**0.81**	0.69	0.83	0.75
Angry	0.67	0.55	0.60	0.75	0.52	**0.61**	**0.81**	**0.59**	0.60
Disgust	0.51	0.72	0.60	**0.65**	**0.83**	**0.73**	0.58	0.72	0.65
Fear	0.59	0.36	0.44	**0.67**	**0.43**	**0.52**	0.64	0.32	0.43
Happy	0.63	0.59	0.60	0.71	0.86	0.78	**0.73**	**0.93**	**0.82**
Sad	0.43	0.57	0.49	0.62	0.60	0.61	**0.74**	**0.77**	**0.75**
Total	0.58	0.57	0.57	0.69	0.69	0.68	**0.70**	**0.70**	**0.68**

every expression. In particular, the generated *surprise, disgust* and *fear* expressions from the proposed model can be even easier classified compared to the ground truth test data. Note also that both ground truth data and the data generated by the proposed model achieve 0.68 F1-score.

5.4 Loss per frame

Since the overall loss is calculated for all frames of the generated expression, we cannot assess the ability of the proposed model to generate with low error-rate the onset, apex and offset of each expression. To evaluate the performance of the

model on each expression phase we calculated the average L_1 distance between the generated and the corresponding ground truth frame for each of the frames of the evolved expression. Figure 5 shows that apex phase, which is usually taking place between time-frames 30-80, has an increased L_1 error for both models. However, the proposed method exhibits a stable loss around 0.45mm across the

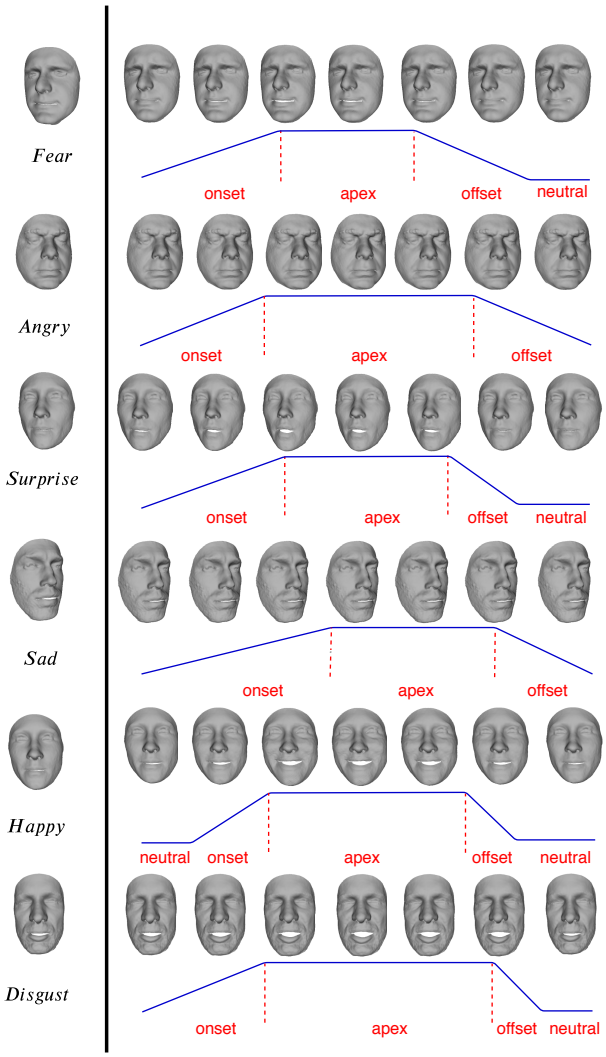

Fig. 4. Frames of generated expressions along with their expected motion labels: Fear, Angry, Surprise, Sad, Happy, Disgust (from top to bottom).

apex phase, compared to the blendshape baseline that struggles to model the extreme deformations that take place and characterize the apex phase of the expression.

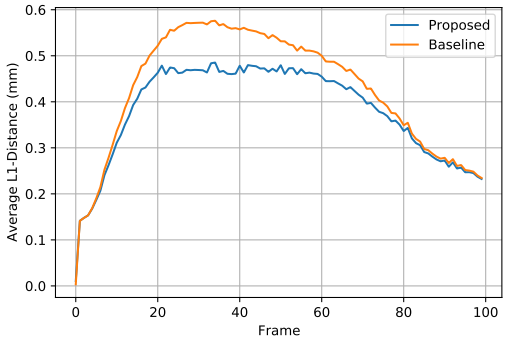

Fig. 5. Average per-frame L_1 error between the proposed method and the PCA-based blendshape baseline.

5.5 Interpolation on the latent space

To qualitatively evaluate the representation power of the proposed LSTM encoder we applied linear interpolation to the expression latent space. Specifically, we choose two different apex expression labels from our test set, we encode them using our LSTM encoder to two latent variables z_0 and z_1, each one of size 64. We then produce all intermediate encodings by linearly interpolating the line between them, i.e. $z_\alpha = \alpha z_1 + (1-\alpha)z_0$, $\alpha \in (0,1)$. The latent samples z_α are then fed to our mesh decoder network. We visualize the interpolations between different expressions in Figure 6.

5.6 Expression generation in-the-wild

Since expression generation is an essential task in graphics and film industries, we propose a real-world application of our 3D facial expression generator. In particular, we have collected several image pairs with neutral and various expressions of the same identity and we have attempted to realistically synthesize the 4D animation of the target expression solely relying on our proposed approach. In order to acquire a neutral 3D template for our animation purposes, we applied a fitting methodology in the neutral image as proposed in [18]. By utilizing the fitted neutral mesh, we applied our proposed model in order to generate several target expressions as shown in Figure 7. Our method is able

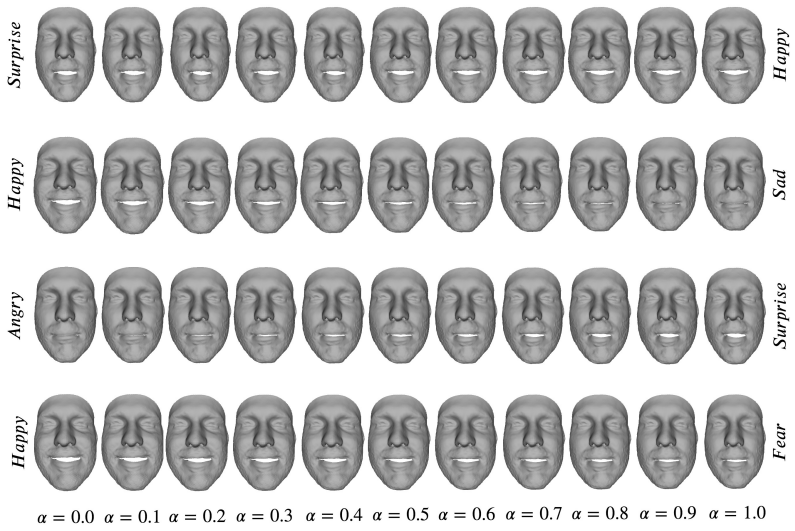

Fig. 6. Interpolation on the latent space between different expressions.

to synthesize/generate a series of realistic 3D facial expressions which demonstrate the ability of our framework to animate a template mesh given a desired expression. We can qualitatively evaluate the similarity of the generated expression with the target expression of the same identity by comparing the generated mesh with the fitted 3D face.

6 Limitations and Future Work

Although the proposed framework is able to model and generate a wide range of realistic expressions, it cannot model thoroughly extremeness variations. As mentioned in section 5.1, we attempted to adapt the extremeness variations of each subject into the training procedure by using an intuitive scaling trick. However, it's not certain that the mean absolute deformation of each mesh always represents the extremeness of the conducted expression. In addition, in order to synthesize realistic facial animation, it is essential to model along with shape deformations also facial wrinkles of each expression. Thus, we will attempt to generalize facial expression animation to both shape and texture, by extrapolating our model to texture prediction.

7 Conclusion

In this paper, we propose the first generative model to synthesize 3D dynamic facial expressions from a still neutral 3D mesh. Our model captures both local

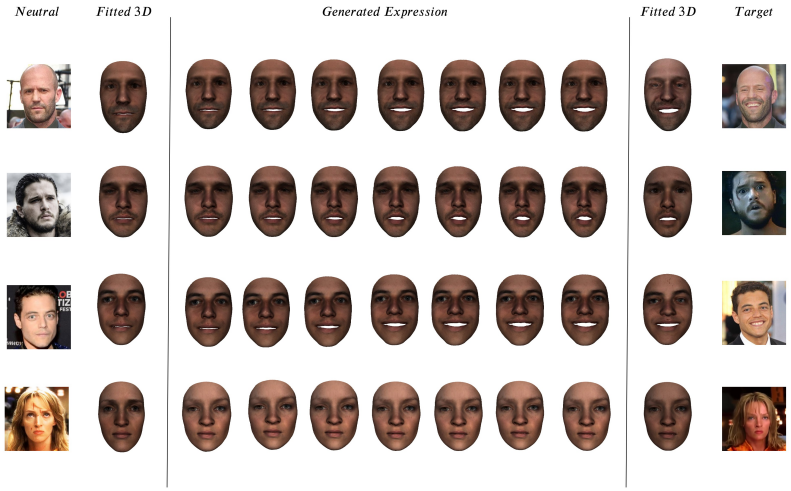

Fig. 7. Generation of expressions in-the-wild form 2D images.

and global expression deformations using graph based upsampling and convolution operators. Given a neutral expression mesh of a subject and a time signal that conditions the expected expression motion, the proposed model generates dynamic facial expressions for the same subject that respects the time conditions and the anticipated expression. The proposed method models the animation of each expression and deforms the neutral face of each subject according to a desired motion. Both expression and motion can be fully defined by the user. Results show that the proposed method outperforms expression blendshapes and creates motion-consistent deformations, validated both qualitatively and quantitatively. In addition, we assessed whether the generated expressions can be correctly classified and identified by a classifier trained on the same dataset. Classification results endorse our qualitative results showing that the generated data can be similarly classified, compared to the ones created by blendshapes model. In summary, the proposed model is the first that attempts and manages to synthesize realistic and high-quality facial expressions from a single neutral face input.

Acknowledgements

S. Zafeiriou and J. Zheng acknowledge funding from the EPSRC Fellowship DEFORM: Large Scale Shape Analysis of Deformable Models of Humans (EP/S010203/1). R.A. Potamias and G. Bouritsas were funded by the Imperial College London, Department of Computing, PhD scholarship.

References

1. Alashkar, T., Ben Amor, B., Daoudi, M., Berretti, S.: A 3D Dynamic Database for Unconstrained Face Recognition. In: 5th International Conference and Exhibition on 3D Body Scanning Technologies. Lugano, Switzerland (Oct 2014)
2. Amberg, B., Knothe, R., Vetter, T.: Expression invariant 3d face recognition with a morphable model. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition. pp. 1–6 (2008)
3. Atwood, J., Towsley, D.: Diffusion-convolutional neural networks. In: Advances in neural information processing systems. pp. 1993–2001 (2016)
4. Boscaini, D., Masci, J., Rodolà, E., Bronstein, M.: Learning shape correspondence with anisotropic convolutional neural networks. In: Advances in neural information processing systems. pp. 3189–3197 (2016)
5. Bouaziz, S., Wang, Y., Pauly, M.: Online modeling for realtime facial animation. ACM Transactions on Graphics (ToG) **32**(4), 1–10 (2013)
6. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
7. Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine pp. 18–42 (2017)
8. Bruna, J., Zaremba, W., Szlam, A., Lecun, Y.: Spectral networks and locally connected networks on graphs. In: International Conference on Learning Representations (ICLR2014), CBLS, April 2014 (2014)
9. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Transactions on graphics (TOG) pp. 1–10 (2014)
10. Chang, Y., Vieira, M., Turk, M., Velho, L.: Automatic 3d facial expression analysis in videos. In: International Workshop on Analysis and Modeling of Faces and Gestures. pp. 293–307. Springer (2005)
11. Cheng, S., Kotsia, I., Pantic, M., Zafeiriou, S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5117–5126 (2018)
12. Cosker, D., Krumhuber, E., Hilton, A.: A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In: 2011 International Conference on Computer Vision. pp. 2296–2303. IEEE (2011)
13. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3D speaking styles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10101–10111 (2019)
14. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in neural information processing systems. pp. 3844–3852 (2016)
15. Fan, L., Huang, W., Gan, C., Huang, J., Gong, B.: Controllable image-to-video translation: A case study on facial expression generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3510–3517 (2019)
16. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-d audio-visual corpus of affective communication. IEEE Transactions on Multimedia **12**(6), 591–598 (2010)

17. Fey, M., Eric Lenssen, J., Weichert, F., Müller, H.: Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
18. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1155–1164 (2019)
19. Gupta, S., Markey, M.K., Bovik, A.C.: Anthropometric 3d face recognition. International journal of computer vision pp. 331–349 (2010)
20. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. ArXiv (2014)
21. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG) p. 94 (2017)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ArXiv (2014)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR) (2017)
24. Levie, R., Monti, F., Bresson, X., Bronstein, M.M.: Cayleynets: Graph convolutional neural networks with complex rational spectral filters. IEEE Transactions on Signal Processing pp. 97–109 (2018)
25. Li, Y., Min, M.R., Shen, D., Carlson, D., Carin, L.: Video Generation From Text. ArXiv (2017)
26. Lim, I., Dielen, A., Campen, M., Kobbelt, L.: A simple approach to intrinsic correspondence learning on unstructured 3d meshes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 349–362 (2018)
27. Masci, J., Boscaini, D., Bronstein, M., Vandergheynst, P.: Geodesic convolutional neural networks on riemannian manifolds. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 37–45 (2015)
28. Monti, F., Boscaini, D., Masci, J., Rodola, E., Svoboda, J., Bronstein, M.M.: Geometric deep learning on graphs and manifolds using mixture model cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5115–5124 (2017)
29. Moreno, A.: Gavabdb: a 3d face database. In: Proc. 2nd COST275 Workshop on Biometrics on the Internet, 2004. pp. 75–80 (2004)
30. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 807–814 (2010)
31. Otberdout, N., Daoudi, M., Kacem, A., Ballihi, L., Berretti, S.: Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. ArXiv (2019)
32. Pham, H.X., Cheung, S., Pavlovic, V.: Speech-Driven 3D Facial Animation with Implicit Emotional Awareness: A Deep Learning Approach. In: 30th IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2017. pp. 2328–2336. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE Computer Society, United States (2017). https://doi.org/https://doi.org/10.1109/CVPRW.2017.287
33. Ploumpis, S., Ververas, E., Sullivan, E.O., Moschoglou, S., Wang, H., Pears, N., Smith, W.A., Gecer, B., Zafeiriou, S.: Towards a complete 3d morphable model of the human head. ArXiv (2019)

34. Ploumpis, S., Wang, H., Pears, N., Smith, W.A., Zafeiriou, S.: Combining 3d morphable models: A large scale face-and-head model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10934–10943 (2019)
35. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 704–720 (2018)
36. Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3d face analysis. In: European Workshop on Biometrics and Identity Management. pp. 47–56. Springer (2008)
37. Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE signal processing magazine pp. 83–98 (2013)
38. Stratou, G., Ghosh, A., Debevec, P., Morency, L.P.: Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In: Face and Gesture 2011. pp. 611–618. IEEE (2011)
39. Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. **34**(6), 183–1 (2015)
40. Tian, G., Yuan, Y., Liu, Y.: Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 366–371. IEEE (2019)
41. Tzirakis, P., Papaioannou, A., Lattas, A., Tarasiou, M., Schuller, B.W., Zafeiriou, S.: Synthesising 3D Facial Motion from "In-the-Wild" Speech. CoRR (2019)
42. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. International Conference on Learning Representations (2018)
43. Verma, N., Boyer, E., Verbeek, J.: Feastnet: Feature-steered graph convolutions for 3d shape analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2598–2606 (2018)
44. Yang, C.K., Chiang, W.T.: An interactive facial expression generation system. Multimedia Tools and Applications pp. 41–60 (2008)
45. Yang, F., Wang, J., Shechtman, E., Bourdev, L., Metaxas, D.: Expression flow for 3d-aware face component transfer. In: ACM SIGGRAPH 2011 Papers. SIGGRAPH '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1964921.1964955, https://doi.org/10.1145/1964921.1964955
46. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition. pp. 1–6 (2008)
47. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P.: A high-resolution spontaneous 3d dynamic facial expression database. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2013)
48. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3438–3446 (2016)
49. Zhong, C., Sun, Z., Tan, T.: Robust 3d face recognition using learned visual codebook. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–6. IEEE (2007)

50. Zhou, Y., Shi, B.E.: Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 370–376. IEEE (2017)