FACTS: Facial Animation Creation using the Transfer of Styles

J. R. Saunders & V. P. Namboodiri

Univeristy of Bath, UK

Abstract

The ability to accurately capture and express emotions is a critical aspect of creating believable characters in video games and other forms of entertainment. Traditionally, this animation has been achieved with artistic effort or performance capture, both requiring costs in time and labor. More recently, audio-driven models have seen success, however, these often lack expressiveness in areas not correlated to the audio signal. In this paper, we present a novel approach to facial animation by taking existing animations and allowing for the modification of style characteristics. We maintain the lip-sync of the animations with this method thanks to the use of a novel viseme-preserving loss. We perform quantitative and qualitative experiments to demonstrate the effectiveness of our work.

CCS Concepts

• ***Computing methodologies*** → ***Animation; Machine learning;***

1. Introduction

Digital humans are a crucial component of the entertainment industry, and there is an increasing demand for high-quality facial animations to drive them. Traditional methods of facial animation are time-consuming and require significant artistic skill. As a result, new approaches are needed to generate large quantities of high-quality facial animations at scale. A popular alternative is the use of performance capture. This allows for large volumes of animation but is still a time-consuming process. While some attempts have been made to democratize this process by allowing for performance capture with consumer-grade devices, these methods still have limitations. Another approach to facial animation is the generation of animations from audio, which requires only speech as input and produces animations. As such methods are in their infancy, they often lack quality due to suppressing motion not directly correlated with audio. We propose an alternative solution for generating animations. Our solution, FACTS, involves taking existing animations and altering stylistic characteristics to create new animations. The result is an order-of-magnitude reduction in the amount of animation that needs to be captured. We define two types of style: emotional and idiosyncratic. Emotional style refers to the changes in a character's facial expressions and movements that correspond to different emotional states, while idiosyncratic style refers to the unique mannerisms of a character. While our approach considers only these two forms of style, it is expected to generalize to arbitrary styles, as long as there is sufficient and diverse data available for that particular style.

Our proposed method is a many-to-many style transfer method using a modified StarGAN. [CCK*18]. We employ cycle consistency [ZPIE17] to learn style mappings **without paired data**, resulting in a more efficient and flexible approach to animation production. This is particularly important, as it would be impossible to obtain perfectly synchronized paired data in different styles. Additionally, we ensure the temporal consistency of our animations by treating rig controls as a time series and incorporating temporal information using a recurrent layer.

Related works have attempted to perform style transfer for facial animation, in particular using cycleGAN on a per-frame basis [MD19]. However, without any temporal modelling this leads to instability. More recent work addresses this with an LSTM [KEZ*19]. Nonetheless, such attempts to edit style led to a distinct loss in lip-sync quality. In particular, lip contact during /m/, /b/ and /p/ phonemes is not preserved. Our work, by contrast, introduces a novel viseme preservation loss, improving lip-sync. This enables us to produce high-quality animations that are synchronized with audio. In summary, the contributions of this work are:

(1) A methodology for producing new animations by altering style characteristics. **(2)** An adaptation of the StarGAN that works for animation data and can alter multiple styles with a single network. **(3)** A novel viseme preserving loss that enables lip-sync without constraining the expressiveness of the mouth.

2. Method

Given an animation $\mathbf{x}$, represented as a sequence of MetaHuman rig controls, our proposed method aims to generate a new animation $\mathbf{x}'$ with the desired style s' while preserving the content of $\mathbf{x}$. We adopt the StarGAN framework [CCK*18] as the basis of our method, which consists of a generator network $\mathcal{G}$ and a discriminator network $\mathcal{D}$. The generator network takes both the input animation $\mathbf{x}$ and the target style label s' as input, and outputs the new animation $\mathbf{x}' =$

DOI: 10.2312/egs.20241017

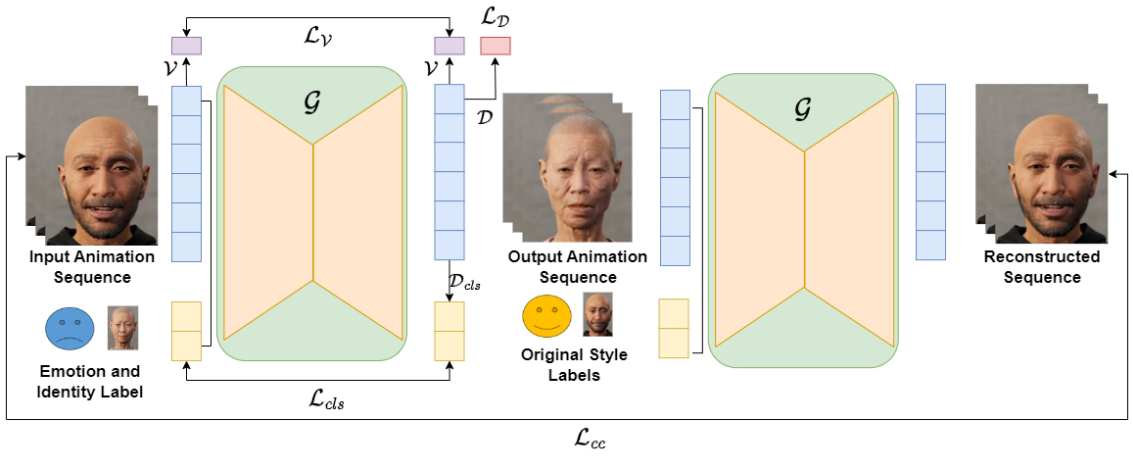

Figure 1: *Overview of the FACTS pipeline, including the losses used in training this model. We take an input animation sequence and target style label and produce novel animations in that style using the generator $\mathcal{G}$. We then recover the original animation by passing the novel animation and original style label to $\mathcal{G}$ giving the cycle consistency loss $\mathcal{L}_{cc}$. The discriminator $\mathcal{D}$ gives adversarial $\mathcal{L}_{\mathcal{D}}$ and classification losses $\mathcal{L}_{cls}$, while the retained viseme classifier $\mathcal{V}$ gives the loss $\mathcal{L}_{\mathcal{V}}$ which helps preserve lip-sync.*

$\mathcal{G}(\mathbf{x}, s')$. The discriminator network has two parts: a critic part $\mathcal{D}_{crt}$ that distinguishes between real and fake data, and a classification part $\mathcal{D}_{cls}$ that predicts the style label of the input data. Specifically, the critic part outputs a scalar value $\hat{y} = \mathcal{D}_{crt}(\mathbf{x})$ that measures the realism of the input data, while the classification part outputs a probability distribution $\hat{s} = \mathcal{D}_{cls}(\mathbf{x})$ over all possible style labels. We represent our data as a time series by concatenating the per-frame vectors of animation controls over a temporal window of length T. This results in sequential data of the form $x \in \mathbb{R}^{T \times N}$, where N is the dimensionality of the per-frame vector. Our method maintains temporal consistency with the use of a recurrent layer, and lip-sync through our novel viseme preserving loss 2.2.1.

2.1. Network Architecture

The generator and discriminator share similar high-level architectures. Each consists of an encoder that maps per-frame rig controls to a higher-dimensional latent space, a recurrent, GRU layer [CvMBB14] to incorporate temporal information, and a decoder to map from the latent space to the desired domain. The encoders and decoders both make use of residual layers and dropout.

The generator takes as input a sequence of rig controls $\mathbf{x} \in \mathbf{T} \times \mathbf{N}$ and a categorical style code $s' \in \{0,1\}^C$, where any present styles are represented as a 1, and outputs a new sequence. This is done by first repeating the style code over the time axis and concatenating it to the sequence (see Figure 1). This concatenated sequence is then fed through the encoder, recurrent and decoder layers. We include a skip connection for better convergence. The discriminator is similar but it does not take a style code and has no skip connection.

2.2. Losses

Cycle Consistency: With facial animations, it is nearly impossible to get paired data, as this would require a frame-to-frame correspondence between animations of different styles. The cycle consistency loss, first introduced for the CycleGAN [ZPIE17] enables us to overcome this barrier by using unpaired data. We can apply a new style s' to an animation $\mathbf{x}$ to obtain $\mathbf{x}'$ and then reapply the original style code s to form a cycle. The cycle consistency loss is then the difference between $\mathbf{x}$ and the cycle. This loss encourages the generator to only change attributes associated with style, and to preserve the content as much as possible. In contrast to the original work, we compute this loss using an ℓ_2 norm rather than ℓ_1 as it is better suited in this domain.

Classification Loss: Given an animation sequence $\mathbf{x}$ and style label s' we want the generator to produce a new animation $\mathbf{x}'$ that has the style characteristics of s'. We are able to do this using the classification branch of $\mathcal{D}$. This classifier should be capable of correctly identifying the style code s'. Simultaneously, the generator should be encouraged to produce sequences that the classifier identifies as having the desired style. This means we must decompose the classification loss $\mathcal{L}_{cls}$ into two components. The first of these components is used to train $\mathcal{D}_{cls}$ to correctly label the styles of the real training data. The second component of this loss is for training $\mathcal{G}$ to produce sequences that are labeled correctly. Both losses are computed using cross entropy.

Adversarial Loss: In addition to creating animations with a desired style, the generator needs to produce animations that appear realistic. This is achieved through an adversarial loss. We do not use the adversarial loss first defined for GANs, but a Wasserstein loss with gradient penalty [ACB17, GAA*17]. Such a loss has been shown to improve the stability of the GAN during training.

2.2.1. Preserving Lip-Sync

Any animation style transfer method must preserve the motion of the lips in such a way that they remain synchronised to audio. This is

a challenging task as even small errors can lead to the lips appearing out of sync. Previous work has [KEZ*19] addressed this by using a cosine mouth loss to preserve the general shape of the mouth. However, such a method will not work in our case as we require the shape of the mouth to vary with style. For example, applying a happy style should be expected to pull the corners of the mouth into a smile. Ideally, we would use the audio to encourage lip sync. However, we do not alter the style of the audio, so any naive attempt to match animation to audio would work against the style transfer. Specifically, consider an expert discriminator trained with contrastive learning, as is common in the 2D domain [PMNJ20]. If we were to train such a model using stylised data, all in-sync pairs would have the same style. Such a discriminator would then only recognise data as in-sync if it had the same style. Say, for example, we then wanted to convert happy to sad and used this expert discriminator, the expert loss would encourage the model to output happy animation to maintain its notion of sync. To overcome this limitation, we need a style-independent representation of speech. For this we use visemes, the visual counterpart to phonemes. The goal is to produce a network that predicts visemes from animation which can be used as an additional loss for the generator. To do this, we first use a pre-trained phoneme classifier [GNZ*21] to predict phonemes from audio. This model is based on Wav2Vec2 [BZMA20] and finetuned to predict phonemes. This classifier takes as input raw waveforms and outputs unnormalised log probability for 392 phoneme tokens over temporal windows of $0.02s$. We then resample the outputs to 60Hz using linear interpolation. This gives us phoneme probabilities for each frame in our animations. Multiple phonemes are expressed with the same mouth shape, so we use a phoneme-to-viseme map [Pho] to obtain per-frame predictions of 16 visemes from audio.

Next we train a neural network classifier $\mathcal{V} : \mathbb{R}^{T\times N} \rightarrow \mathbb{R}^{T\times 16}$ to predict the visemes from the animation curves alone. Such a network has a few advantages. The classifier takes in only animation data and does not require audio, therefore it can be applied to animations created by the generator which do not have corresponding audio. It also predicts visemes **independent** of speaker identity and emotion, giving a fixed measure of lip shape that can be preserved by the generator. The network has a similar architecture to the discriminator network, with an encoder, GRU layer, and decoder. The only difference is that the final linear layer has different output dimensions to match the number of visemes. We can then use this network to derive the *viseme-preserving loss*, which is simply the cross-entropy between the viseme classifications of the real sequences $\mathbf{x}$ and the viseme classifications of the generated sequences $\mathcal{G}(\mathbf{x}, s')$. The loss encourages the generator to produce sequences with the same mouth shapes as the input sequence, without constraining the style.

3. Results

3.1. Data

Our dataset consists of a total of 30 minutes of animations across MetaHuman rigs. The animations are obtained using head-mounted cameras with a custom solver designed by Cubic Motion [Ree22]. Each actor is asked to record a series of seven phonetic pangrams, covering all phonemes. These consist of spoken sentences in the desired emotion. Three emotions are chosen: happy, neutral, and sad. For each of these, we record $\approx$ 5 minutes of animation per actor.

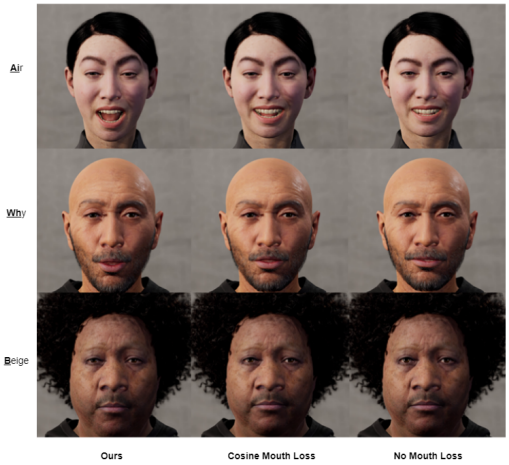

Figure 2: *The presence of the Viseme Loss in our work significantly improves the lip sync when compared to both our method without this loss, and the work of [KEZ*19]*

Method	LSE-D ↓	LSE-C ↑	Emo-P ↑
Ours w/o Viseme Loss	9.501	1.781	0.421
Cosine Mouth Loss [KEZ*19]	8.699	2.667	0.417
Ours Full	**7.33**	**4.667**	**0.443**
Captured Data	7.617	4.502	0.385

Table 1: *A table comparing our work with and without the Viseme Loss to that of Neural Style Preserving Visual Dubbing [KEZ*19] across metrics for lip sync (LSE-C and LSE-D) and for emotional clarity (EMO-P).*

To ensure we have captured a wide enough range of facial motion, we make use of seven phonetic pangrams. This gives us a total of 6 different forms of style, which is sufficient to demonstrate that our method can indeed perform many-to-many style transfer.

3.2. Implementation Details

For all networks, we use a hidden dimension of 256. Dropout is applied after each layer, except the GRU, with probability 0.4. We use a batch size of 32 and using the Adam optimiser with learning rate of 10^{-4}. During training, the sequence length is fixed to 30 frames. During inference, the sequence length can be arbitrary. We train for 100 epochs taking approximately 12 hours on an P6000.

3.3. Metrics

To justify the improvement made by our work, we quantitatively and qualitatively evaluate the results. For lip sync, a focus of our work, quantitative metrics exist. We use a pre-trained syncnet [CZ16] to measure lip sync by passing renderings of the animations together with the corresponding audio. We report the LSE-C and LSE-D metrics [PMNJ20]. The LSE-D metric quantifies the level of lip-sync, while the LSE-C metric measures the confidence, where low scores for LSE-C mean that the audio and video are only weakly correlated. In our experiments, we perform two forms of style transfer.

Statement	Lip-sync (95% CI)	Naturalness (95% CI)	Emotion (95% CI)
Ours > Cosine Mouth Loss	**66 (52-82)**	**66 (51-81)**	**73 (59-86)**
Ours > No Mouth Loss	**66 (52-81)**	**67 (51-81)**	63 (50-79)

Table 2: *The results of our user study (N = 20), showing the percentage of users that agree with the provided statement. We include 95% confidence intervals for each comparison in brackets. Where the lower bound is > 50% (e.g. we are 95% confident our method is indeed preferred) we show the result in bold.*

For emotional style transfer, we can define a quantitative metric. For this purpose, we use a pre-trained emoNet [TKB*21]. This network predicts the per-frame emotion of a given video without considering the audio, outputting logits related to each emotion. We restrict the output of this network to just the three emotions we are considering: happy neutral, and sad, and take the softmax to give pseudo-probabilities for each emotion. We propose using this as a metric. We denote this metric as Emo-P.

3.4. Comparisons

Quantitative: To the best of our knowledge, the only work to attempt style transfer for facial animation is the work of Neural Style-Preserving Visual Dubbing (NSPVD) [KEZ*19]. Our work varies from theirs in two key ways. The first is that we use a starGAN whereas they use a CycleGAN. These networks perform different tasks and it is therefore difficult to compare the two. The second is our use of the viseme preserving loss in place of a cosine mouth loss. We compare the effects of this novel loss in Table 1. We also perform a simple ablation study to demonstrate that the inclusion of the viseme preserving loss does improve the lip-sync. These results are also in Table 1. Our method outperforms the mouth cosine loss from NSPVD [KEZ*19] across all metrics, which in turn is an improvement over not using any additional mouth loss. **Qualitative:** We are also able to show the effectiveness of our method qualitatively. Figure 2 shows example frames of animation. It can be seen that our method produces the expected lip shapes for the highlighted section of the given words better than NSPVD [KEZ*19] and the baseline method. For example, the start of the word “why" has more clearly pursed lips. For further qualitative results, we refer to the supplementary material. **User Study:** We also run a two-alternative forced choice test to compare our method with the cosine mouth loss [KEZ*19] and with no mouth loss. To prevent bias, we display the methods side-by-side (without labels) in random order and simply ask users to select their preferences. The results are shown in Table 2 and show that our method is preferred across all metrics. Under a binomial assumption, all results are significant to a confidence of 95% except for comparing to no lip loss on emotion. Of particular note is how emotion is much clearer in our method compared with the cosine mouth loss, validating our hypothesis that the mouth is over-constrained.

4. Conclusions and limitations

In this work, we proposed a novel method for style transfer in facial animation using the StarGAN framework and a viseme preserving loss. Our method is able to generate new facial animations with a desired style label while preserving the content of the original animation. We also demonstrated that the inclusion of the viseme preserving loss improves lip-sync in the generated animations. Compared to previous work on facial animation style transfer, our method achieves superior results in terms of lip-sync accuracy and style transfer quality. We believe that our method has potential applications in areas such as animation and film production. Our work does have limitations, in particular, we have only tested a small range of styles. Future work could focus on improving the generalization of the method to different datasets and exploring the use of additional losses to further improve the quality of the generated animations.

References

[ACB17] ARJOVSKY M., CHINTALA S., BOTTOU L.: Wasserstein generative adversarial networks. In *International conference on machine learning* (2017), PMLR, pp. 214–223. 2

[BZMA20] BAEVSKI A., ZHOU Y., MOHAMED A., AULI M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems 33* (2020), 12449–12460. 3

[CCK*18] CHOI Y., CHOI M., KIM M., HA J.-W., KIM S., CHOO J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018). 1

[CvMBB14] CHO K., VAN MERRIËNBOER B., BAHDANAU D., BENGIO Y.: On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (Doha, Qatar, Oct. 2014), Association for Computational Linguistics, pp. 103–111. 2

[CZ16] CHUNG J. S., ZISSERMAN A.: Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV* (2016). 3

[GAA*17] GULRAJANI I., AHMED F., ARJOVSKY M., DUMOULIN V., COURVILLE A. C.: Improved training of wasserstein gans. *Advances in neural information processing systems 30* (2017). 2

[GNZ*21] GAO H., NI J., ZHANG Y., QIAN K., CHANG S., HASEGAWA-JOHNSON M. A.: Zero-shot cross-lingual phonetic recognition with external language embedding. *Interspeech 2021* (2021). 3

[KEZ*19] KIM H., ELGHARIB M., ZOLLÖFER MICHAEL SEIDEL H.-P., BEELER T., RICHARDT C., THEOBALT C.: Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 178:1–13. 1, 3, 4

[MD19] MA L., DENG Z.: Real-time facial expression transformation for monocular rgb video. *Computer Graphics Forum 38*, 1 (2019), 470–481. 1

[Pho] English, british (en-gb). https://docs.aws.amazon.com/polly/latest/dg/ph-table-english-uk.html. Accessed: 22-02-2022. 3

[PMNJ20] PRAJWAL K. R., MUKHOPADHYAY R., NAMBOODIRI V. P., JAWAHAR C.: A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia* (New York, NY, USA, 2020), MM '20, Association for Computing Machinery, p. 484–492. 3

[Ree22] REED K.: *An Analysis of Example*. Phd thesis, University of Bath, Bath, UK, May 2022. Available at https://researchportal.bath.ac.uk/files/243071931/ThesisCorrectionsFinal.pdf. 3

[TKB*21] TOISOUL A., KOSSAIFI J., BULAT A., TZIMIROPOULOS G., PANTIC M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* (2021). 4

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on* (2017). 1, 2