Paper on Facial Manipulation Techniques

C. Malathy, Mridula Vijendran,Krishna Maneesha Dendukuri

Abstract*: Facial manipulation is the manipulation of pose, identity, albedo, expression, and texture for creative, artistic, and aesthetic manipulation in different styles. It's used in visualizations, videos and image presentations, and digital media, social media, advertising, restoration and preservation of old and damaged photos or images. However, it can also be used for hijacking the facial identity of the target video or image, thus, compromising the integrity of the person in question. This can be accomplished by techniques using supervised convolutional neural networks, unsupervised generative models and physics-based models. These models aim to transfer traits and features that are intuitively comprehensible from the source to the target. Another objective could be to give full control to modify the features in the generated model, in a manner that we can grasp the immediate correlation to the resultant output. This paper considers the various techniques in deep learning and physics simulations to achieve facial manipulation.

Index Terms: 3D parametric face models, Neural Style Transfer, Generative Adversarial Networks(GAN), VariationalAutoencoders(VAE), Physics-based facial manipulation.

I. INTRODUCTION

The popularization of social media put an emphasis on video and image editing of faces for better presentation, expressions of thoughts by compositing faces on different images or by their creatively editing them.

With the advent of deep learning which is growing rapidly, newer technology can be accessed by people for image inpainting in cases of photobombing, selfie beautification for social media and job profiles to improve impressions and seek social acceptance. It can also be used for increasing the resolution and the dynamic range of compressed images to save bandwidth, cleaning and filtering the image to reduce undesirable effects like noise, glare, accidental warping, and artifacts. Also, it's being used for restoration of scratches, dusty, deteriorated, and deformed faces in scanned copies of old photographs and video footage.

But, as these ubiquitous editing techniques are becoming more realistic, intricate, and versatile, these can be used with malicious intent. People intuitively rely on facial cues and body language to identify the other's nature, intentions, behaviors, reactions, and characteristics. These affect the level of familiarity with the target and their own reactions, trust, thoughts, and decisions that are based on emotion. We've developed to rely on faces to an extent that our brain has evolved to have a specialized region, the fusiform face area (FFA) that is responsible for facial recognition. This makes the access to unrestricted editing of faces usable for identity theft, spreading misinformation and manipulation of the author and the production team's intentions.

Facial manipulation is generally achieved by transfer of the source's actions/characteristics onto the target media via convolutional neural networks, where the source can consist of audio[13], video[20,25] and image[21,1,24]. This can be done using a convolutional neural network where the input and output is the source and target image, and generative models like variationalautoencoder[29] and generative adversarial networks[27] which generate the target samples, as the bases. It's also achieved by the control of a 3d face model via biophysical processes[23,5] and snapshot sequences[22] of facial changes in age, appearance or expression.

II. RELATED WORKS

D parametric face models

The parametric face model is a 3d intermediate face model that is a fully controllable mesh that is monocular reconstruction from the source's composites. This is trained with the loss function of the composites to transfer the motion, alignment of parts and restrict them from acquiring extreme and unnatural positions.

The environment's illumination is considered as constant and uniformly smooth around the image (Lambertian surface) and is emulated using spherical harmonics[3,24,19]. The Expression blendshapes[21,24,19,3] and models, maps or loss functions for specific parts like the mouth[24,19] and gaze[1], that are retrieved and tracked from the target, in fine details transfer to the face model. Another method for using the face identifying features as landmarks for alignment and model building in different poses can be generated, by projecting fiducials[21] and PCA low dimensional features[24,19].

The source would transfer and estimate the composited identity, pose, illumination, expression, and skin reflectance or albedo onto the model and projects it onto a target. This uses a supervised learning approach to generate the pairs by reducing the learning problem's complexity by incorporating the face model with upsampling and downsampling. Video-specific problems add an extra dimension of time, which can be handled using edge weights from a fully connected frame graph[20], temporal sliding windows as inputs[25], time slices[3] as a dimension of the parametric model, and temporal loss[19], to reduce variation across frames ensuring a smooth transition.

This model is not very general in nature and can't

Revised Manuscript Received on December 22, 2018.

Dr. C. Malathy,SRM Institute of Science & Technology, malathy.c@ktr.srmuniv.ac.in,
MridulaVijendran,SRM Institute of Science & Technology, mridulavijendran_vi@srmuniv.edu.in,
Krishna Maneesha Dendukuri, ,SRM Institute of Science &Technology, krishnamaneesha_ph@srmuniv.edu.in

handle occlusions of features or uneven video transitions and jerky movements. It can't handle extremities in human facial feature alignment or manipulation.

3.2 Neural Style Transfer

This uses a trained supervised convolutional neural network to extract features representing low-level features and high-level features from the lower and higher layers respectively [28], given an input source and target image. They're used to incorporate the style of the target image onto the source image while preserving its content and structure. Traditional neural style transfer techniques minimize style and perceptual losses on images using the Frobenius norm while optimizing the slow adversarial loss to find a balance between the style and content manipulation since they both together can't be faithful to the target's style and the source's content.

$$\mathcal{L}_{content}(\vec{p},\vec{x},l) = \frac{1}{2}\sum_{i,j}\left(F_{ij}^{l} - P_{ij}^{l}\right)^{2} . \quad (1)$$

The content loss(1) is given by a loss, in this case, l2 loss, between original and generated image as P and F respectively.

$$G_{ij}^{l} = \sum_{k} F_{ik}^{l} F_{jk}^{l}. \quad (2),$$

$$E_{l} = \frac{1}{4N_{l}^{2}M_{l}^{2}}\sum_{i,j}\left(G_{ij}^{l} - A_{ij}^{l}\right)^{2} \quad (3),$$

$$\mathcal{L}_{style}(\vec{a},\vec{x}) = \sum_{l=0}^{L} w_{l} E_{l} \quad (4)$$

The feature correlations that represent the style, are given by the Gram matrix(2) which is calculated over the feature maps for the original and generated image as A and G respectively in a layer. Then, a normalized Frobenius norm is applied with an L2 loss between G and A(3) and finally, a weighted l1 norm is used to get the style loss(4) between the original and generated image.

They're not stable across video frames and are more faithful to the style image and overflow into the source's content, thus being not very photo realistic. They're also not very generalized since it's effectiveness depends on the supervised model, its domain and input distribution. It won't be good when dealing with sparse samples and it can be rigid when considering that other layer combination for the style and perceptual losses can be better for subsets of a class.

$$A : A' :: B : B' \quad (5)$$

To make the style transfer more faithful to the style images, while preserving the source's content, deep correspondence[9] can be considered, where A is the input and the source, A' and B are latent images and B' is the output image or the target(5).

$$\phi_{a\to b}^{L}(p) = \arg\min_{q} \sum_{x\in N(p), y\in N(q)} (||\overline{F}_{A}^{L}(x) - \overline{F}_{B}^{L}(y)||^{2} + ||\overline{F}_{A'}^{L}(x) - \overline{F}_{B'}^{L}(y)||^{2}), \quad (6)$$

This concept is implemented by learning the latent representations using sequential intermediate layers from coarse, higher layers to detailed lower layers using fast randomly sampled nearest neighbor fields and calculating the l2 loss between the images A, A', B and B'(6).

The flaws in the deep correspondence based model are that, they need images that are similar in scene and content, as in, they should both have the same objects that are recognizable to the supervised models.

3.3 Generative Adversarial Networks

General Adversarial Networks(GAN)[27] as generative models that consist of a generator that generates samples dependent on the distribution of the discriminator. The discriminator is a neural network that's used to distinguish between generated fake data and real data. GANs are used to generate high-resolution samples using unsupervised deep learning and convolutional neural networks, by sampling from the generator or interpolating between the generated samples. People have used it for data augmentation[15], image to image translation[8,10] for recoloring, 2D to 3D projection and vice versa, time lapse, style transfer, simplified characteristics mapping (like semantic maps), super-resolution and image inpainting.

The generator learns a mapping from random vectors to the input distribution and the discriminator learns to distinguish between the fake and real samples(binary classification problem).

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}$$

(7)

This model can be trained using a min-max function(7) which acts as the adversarial loss between the generator that wants to minimize the discriminator's loss difference between that of the generated sample and the input, and the discriminator which wants to maximize the loss difference between that of the fake and the real data.

The main disadvantage in GANs is that they can be hard to train and can fall into mode collapse, where they generate output images with very less variance.

To make a cross-domain transformation of the input data from the source distribution to the output data in the target distribution, conditional GANs or cGANs[10] can be used. These need pairwise data that represent the image before and after transformation.

$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_{z}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}$$

(8)

The model uses a class vector which helps in conditional generation of data for the particular class used in the objective function(8) which

uses the GAN objective function as its base.

These can't generate samples that actually belong to the target distribution, but doesn't match the generated data due to the conditional constraint.

Cycle GAN[8] is used to introduce a cycle consistency via F(G(X)) ~X, where F and G are generators and this model is used for an unpaired image to image translation using 2 distributions of source and target images, that are best similar in scene and content, using unsupervised learning.

$$\mathcal{L} = \mathcal{L}_{reconstruct} + \lambda_1 \mathcal{L}_{prior} + \lambda_2 \mathcal{L}_{flow}$$

$$\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] + \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))] \quad (9)$$

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]. \quad (10)$$

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F), \quad (11)$$

The objective function(11), where G maps source X to target Y, F maps the reverse, Dx and Dy are the discriminators for distributions of X and Y. The GAN loss is taken between the D of an image and the generator of the image in another class(9) and the cyclic loss is an l2 loss between the generated image as a composition from 2 generators, to ensure symmetric and sureness about the concept learned, and the actual image(10).

This network is not good at translation from content diverse images and is susceptible to gradient vanishing problem.

Domain Transfer Networks[15] utilize the cycle GAN while changing the losses from the source and the target losses from a log loss to MSE (Mean Squared Error) to combat the gradient vanishing problem. It's good for domain transfer from diverse, but abstractly similar data pairs and it makes classification task better, by populating the sparse data manifold. But the generated data isn't as visually appealing as its manually generated sample and the technique hasn't been extensively tested with other classification tasks.

3.4 VariationalAutoencoders

VariationalAutoencoders[29] are an unsupervised deep learning technique that is mainly used to learn the latent space representation of the images allowing the system to learn the reconstruction of the image from the most significant features that represent it(included in the latent space). As the VAEs got better at reconstruction and generation of the new images, it's applications have also been extended to more realistic and efficient manipulation of images.

$$\mathbf{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}^{(l)} \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbf{I}) \quad (12)$$

(12) is used to make the stochastic problem of sampling from a Gaussian distribution to model an input distribution, deterministic by only sampling ε from the Gaussian distribution. μ and σ are derived from the model's parameters is the expected lower bound optimization that's used for reducing the distance between the posterior distribution that the model learns and the prior distribution that the input distribution represents that is the Gaussian distribution, using the KL divergence. The second term is the reconstruction term that's used to improve the model's distribution to the ideal posterior distribution.

The facial attributes and the expressions can be effectively controlled by changing the values of the corresponding vector in the latent space. For instance, the expression such as a smile can be converted into a neutral expression by controlling the values of the vector corresponding to mouth in particular from the latent space.

But due to the fact that the reconstruction is solely based on the major features of the face, there is a fair probability of losing the complex details and minute attributes of the face and thus getting blur results in few cases. This is due to the network's optimization function, which is commonly an L2 loss function, which is inappropriate to accurately model the input distribution. To overcome the same, there have been various addendums that were incorporated on the autoencoders such as flow variationalautoencoders(FVAE)[16] that utilizes the flow field between the source and the target images to generate transitory states from the learned posterior distribution. In the case of FVAEs, the latent space from the encoder network is mapped to the flow space instead of the pixel space.

where λ1 and λ2 are hyperparameters to be tuned. The 3 loss terms in (12) correspond to the level of detail and accuracy of the generated image, the modeling of the latent vector distribution as a multivariate Gaussian, and generation of the flow field latent vector.

Adversarial autoencoders is another version that's utilized to incorporate the semantic-based modification of the input data of the VAE along with the high-resolution results from adversarial networks in GAN. A model that's based on this is the conditional adversarial autoencoders(CAAE)[17] which uses two discriminators that take care of the smooth transition between the input to desired output, while the other evaluates the level of detail. CAAEs are similar to the AAEs which use the best of GANs and VAEs.

$$\min_{E,G} \max_{D_z, D_{img}} \lambda \mathcal{L}(x, G(E(x), l)) + \gamma TV(G(E(x), l)) + \mathbb{E}_{z^* \sim p(\mathbf{z})}[\log D_z(z^*)] + \mathbb{E}_{x \sim p_{data}(\mathbf{x})}[\log(1 - D_z(E(x)))] + \mathbb{E}_{x,l \sim p_{data}(\mathbf{x},\mathbf{l})}[\log D_{img}(x, l)] + \mathbb{E}_{x,l \sim p_{data}(\mathbf{x},\mathbf{l})}[\log(1 - D_{img}(G(E(x), l)))] \quad (15)$$

In equation (15), λ is used for smoothness regularization and σ is used for resolution regularization. The first term refers to the reconstruction loss between the input image

and the generated image, whereas the second term refers to the resolution of the face generated via controlling the total variance of the image. The discriminator on z is trained by the third and the fourth terms, and that on the image is trained using the fifth and the sixth terms. These two use min-max optimization to facilitate adversarial training with their corresponding generators.

3.5 Physics-based facial manipulation

This technique is used to simulate human skin using their biophysical properties on exposure to time, environmental interactions and physical and natural forces. It considers the skin anatomy, reflectance (optical properties), it's layered structure their reflections, scattering and absorption properties to realistically mimic human skin and transform an existing model[23] for simulations in the 3D digital model for aging and lighting effects.

Another model uses the sequencing of expression changes to automatically generate the physics-based 3D model[5], after considering the muscle activation model which accounts for tissue material, stiffness and volume, face geometry, muscle behavior. This considers blendshapes to create a template of expressions that the face can take, bone and joint kinematics and collision handling along with their interaction with the overlying skin.

III. CONCLUSION

There has been a drastic increase in the efficiency of the image manipulation techniques with the spread of the Deep Learning techniques like neural style transfer and generative models like the Generative Adversarial Networks and the VariationalAutoencoders. However, there are challenges faced by each of these methods which need still need to be addressed.

In every method and technique presented in this paper, the manipulation is commonly done by either a single source's attributes being mapped to a single target's attributes or by self-reenactment in case of the expression manipulation. Manipulation of multiple target's with respect to a single source hasn't been achieved yet.

Also, the performance of these techniques is questionable when the person occupies only a small part of the image, giving rise to the image localization problems. These issues, if addressed, will open up a lot more possibilities for useful applications of these techniques

REFERENCES

1. Cloud Security Alliance, Top Threat to Cloud Computing V1.0, March 2010.
2. S. Muqtyar Ahmed, P. Namratha, C. Nagesh. Prevention Of Malicious Insider In The Cloud Using
3. Decoy Documents
4. Ajey Singh, Dr. Maneesh Shrivastava Overview of Attacks on Cloud Computing
5. D.Jamil and H. Zaki. Security Issues in Cloud Computing and Countermeasures, International Journal of Engineering Science and Technology, Vol. 3 No. 4, pp. 2672-2676, April 2011.
6. K. Zunnurhain and S. Vrbsky. Security Attacks and Solutions in Clouds, 2nd IEEE International Conference on Cloud Computing Technology and Science, Indianapolis, December 2010.
7. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, Creating evolving user behavior profiles automatically, IEEE Trans. on Knowl. and Data Eng., vol. 24, no. 5, pp. 854867, May 2012.
8. F. Rocha and M. Correia, Lucy in the sky without diamonds: Stealing confidential data in the cloud, in Proceedings of the 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops, ser. DSNW 11. Washington, DC, USA: IEEE Computer Society, 2011.
9. M. B. Salem and S. J. Stolfo, Modeling user search behavior for masquerade detection, in Proceedings of the 14th international conference on Recent Advances in Intrusion Detection, ser. RAID11. Berlin, Heidelberg: SpringerVerlag, 2011, pp. 181-200.
10. S. et al, Decoy document deployment for effective masquerade attack detection, in Proceedings of the 8th international conference on Detection of intrusions and malware, and vulnerability assessment, ser. DIMVA11. Berlin, Heidelberg: Springer-Verlag, 201
11. Rajesh, M., and J. M. Gnanasekar. "Path Observation Based Physical Routing Protocol for Wireless Ad Hoc Networks." Wireless Personal Communications 97.1 (2017): 1267-1289.
12. Rajesh, M., and J. M. Gnanasekar. "Sector Routing Protocol (SRP) in Ad-hoc Networks." Control Network and Complex Systems 5.7 (2015): 1-4.
13. Rajesh, M. "A Review on Excellence Analysis of Relationship Spur Advance in Wireless Ad Hoc Networks." International Journal of Pure and Applied Mathematics 118.9 (2018): 407-412.
14. Rajesh, M., et al. "SENSITIVE DATA SECURITY IN CLOUD COMPUTING AID OF DIFFERENT ENCRYPTION TECHNIQUES." Journal of Advanced Research in Dynamical and Control Systems 18.
15. Rajesh, M. "A signature based information security system for vitality proficient information accumulation in wireless sensor systems." International Journal of Pure and Applied Mathematics 118.9 (2018): 367-387.