

Data-Driven Facial Animation (2)

Taku Komura

Deep Fake, Digital Dub

- High interest in animating famous people saying things they don't really say
- Let's look at how this can be done



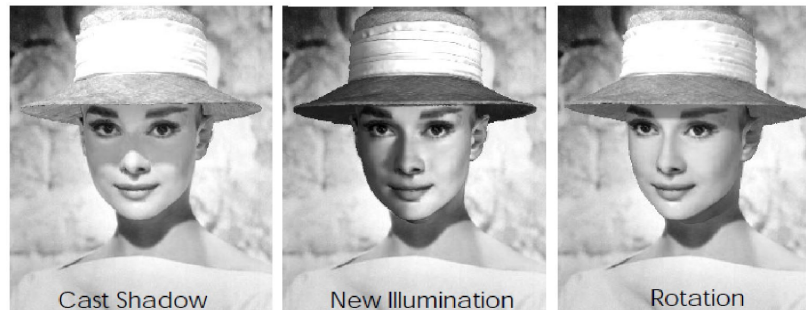
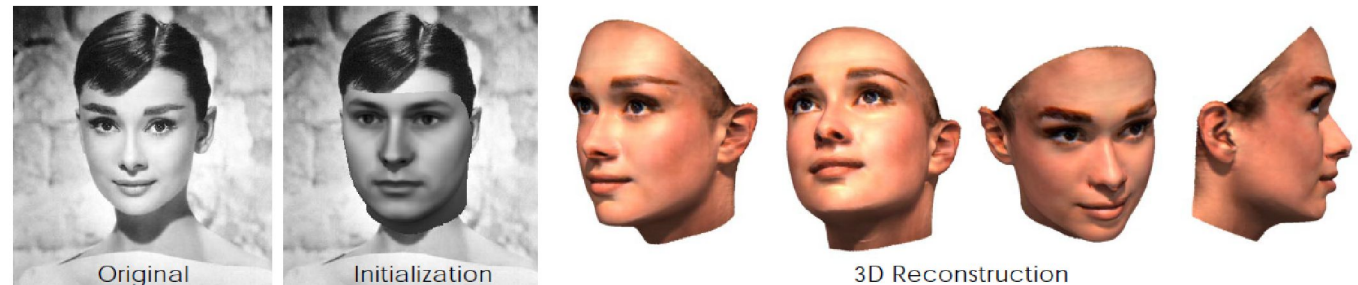
Overview

- Fitting 3D face models into 2D images (Blanz and Vetter '99)
- Controlling the person in the video by your own face (Thies et al. 2016)
- Further improvement by deep learning (Kim et al. 2018)

Let's first think of fitting 3D models into 2D images (Blanz and Vetter '99)

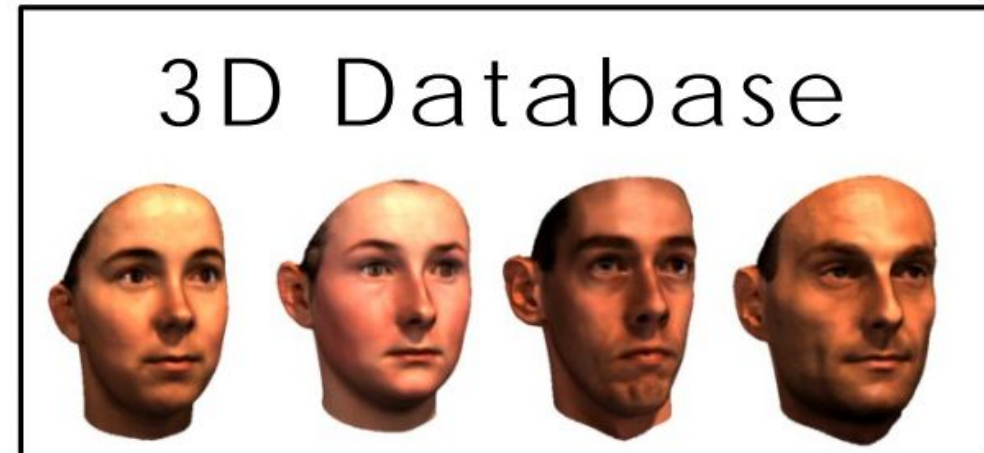
Given a 2D image, if we can fit a 3D model into it, we can

- change the view direction
- change the lighting condition
- edit the face in 3D



Building a space of faces/heads in 3D

- Using a lot of face data of different people, we can construct a space of different faces/head (using face/head of 200 people)
- These are collected using the Cyberware scanning system
- The correspondence of the faces need to be made using a generic face model (as done in deformation transfer, expression cloning)
- Applying Principal Component Analysis to the data



Morphable Models

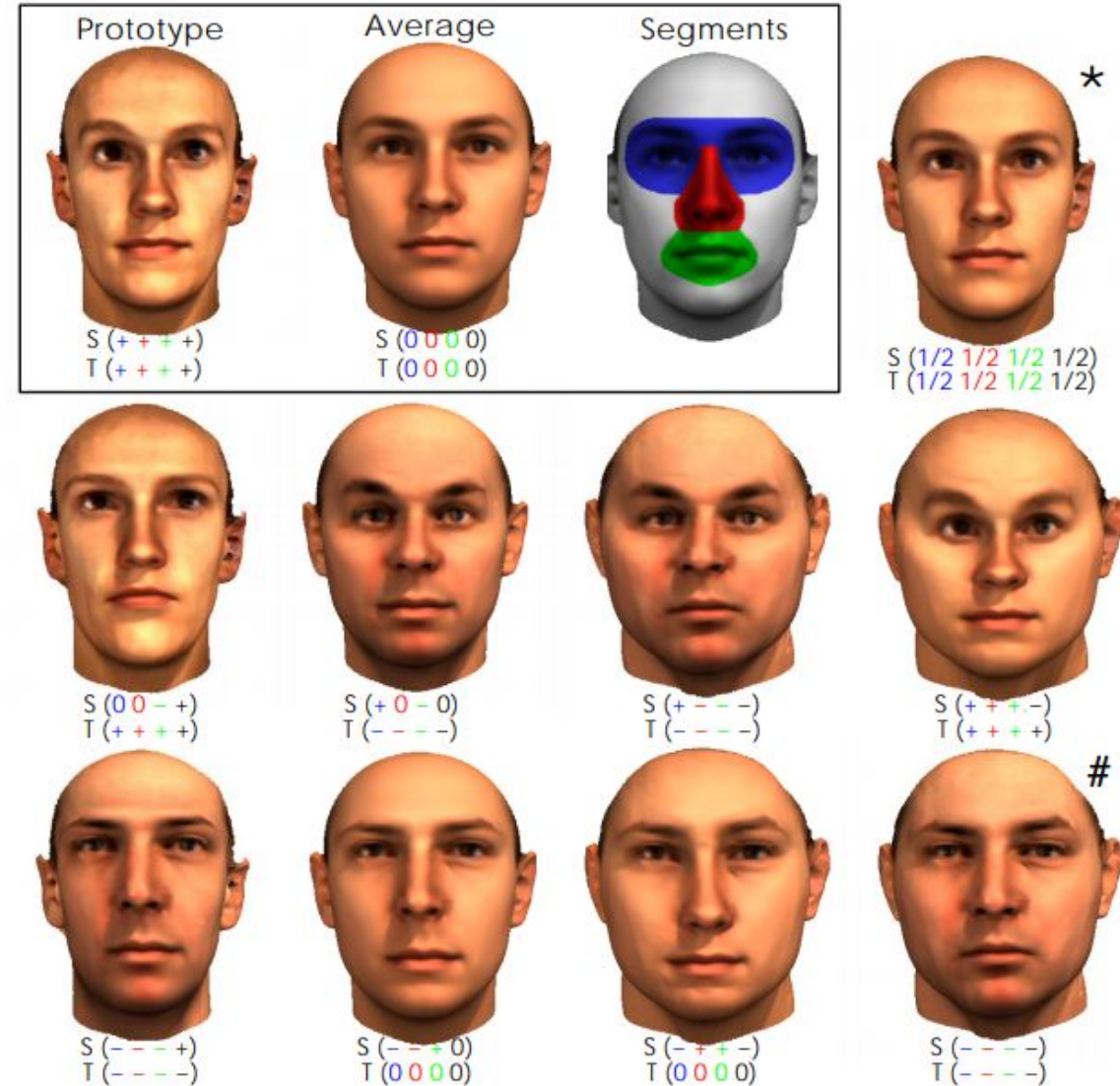
- Principal component analysis is applied to the geometry and texture of the face dataset
- The geometry and texture of a new face can be represented as

$$S_{model} = \bar{S} + \sum_{i=1}^{m-1} \alpha_i s_i, \quad T_{model} = \bar{T} + \sum_{i=1}^{m-1} \beta_i t_i$$

where \bar{S}, \bar{T} are average geometry and texture, s_i, t_i are their principal components, and α_i, β_i are coefficients

Different Faces by Adjusting the Coefficients

- Here we show an example where we produce example faces by adjusting the coefficients of the principal components
- The face is segmented into different parts and the components are adjusted for each part separately
- Local control – this was also done in Spacetime faces



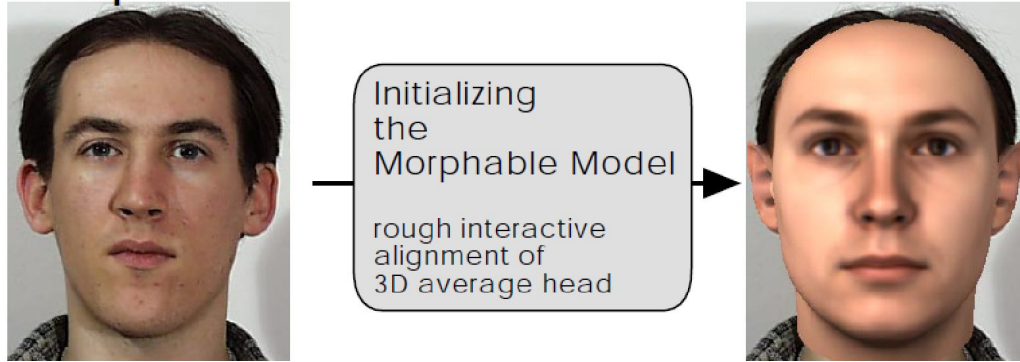
Computing the Morphable Model From Images

- Using the morphable model, we can fit the model to the image
- Given the camera direction, lighting condition, etc, we can compute the image of the morphable model by computer graphics (Phong illumination)
- Minimize the difference of the computer generated image and the original image

$$E_I = \sum_{x,y} \|\mathbf{I}_{input}(x,y) - \mathbf{I}_{model}(x,y)\|^2$$

- We compute $\min_{\alpha,\beta} E_I$ to find the geometry and texture parameters

2D Input

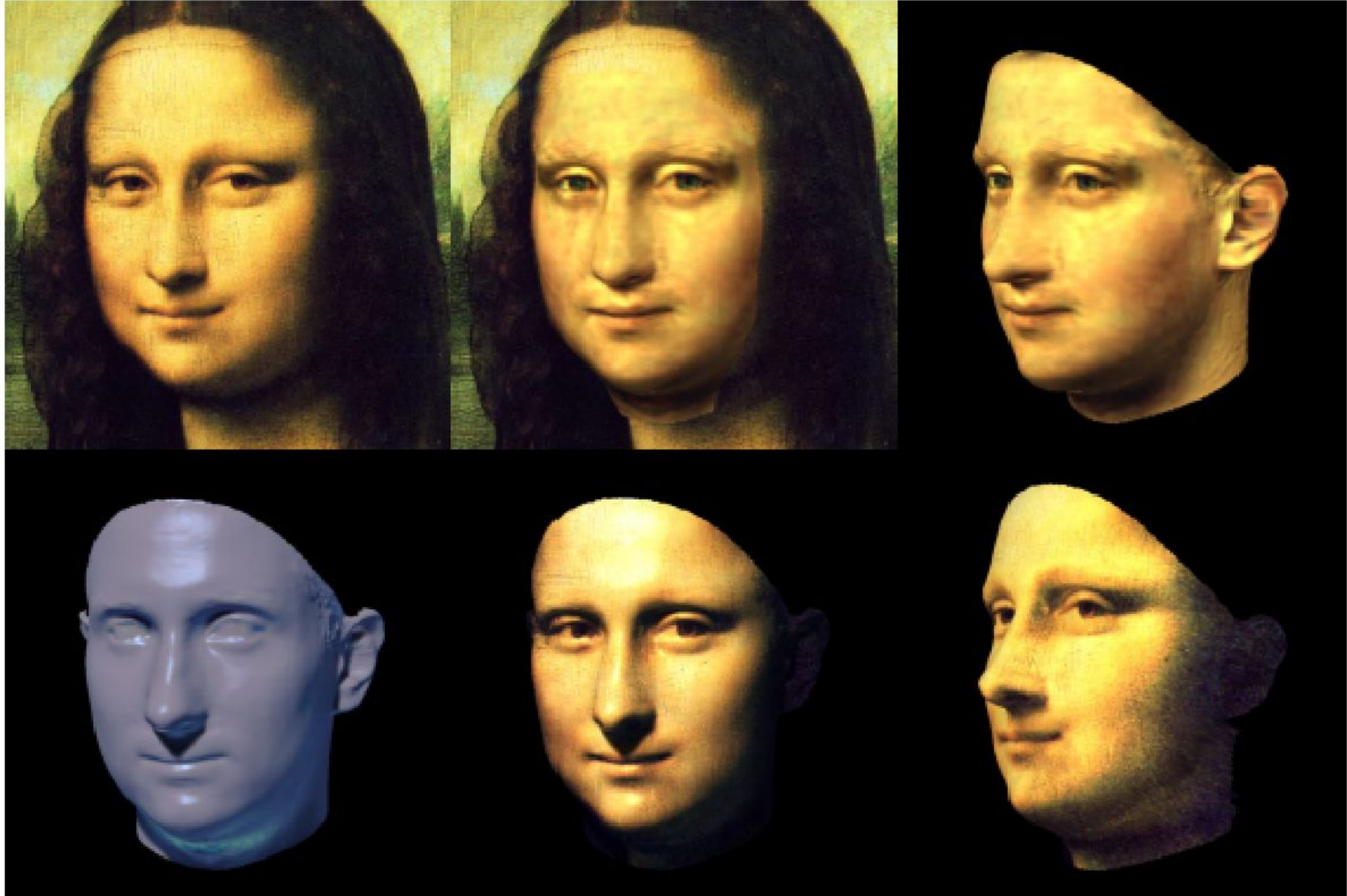


- First a the model is roughly fitted to the image

- The texture and geometry is optimized to match the photo

- Multiple images can be used to further improve the fit





Automated Matching

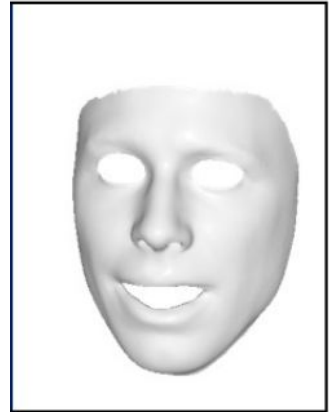


Overview

- Fitting 3D face models into 2D images (Blanz and Vetter '99)
- Controlling the person in the video by your own face (Thies et al. 2016)
- Further improvement by deep learning (Kim et al. 2018)

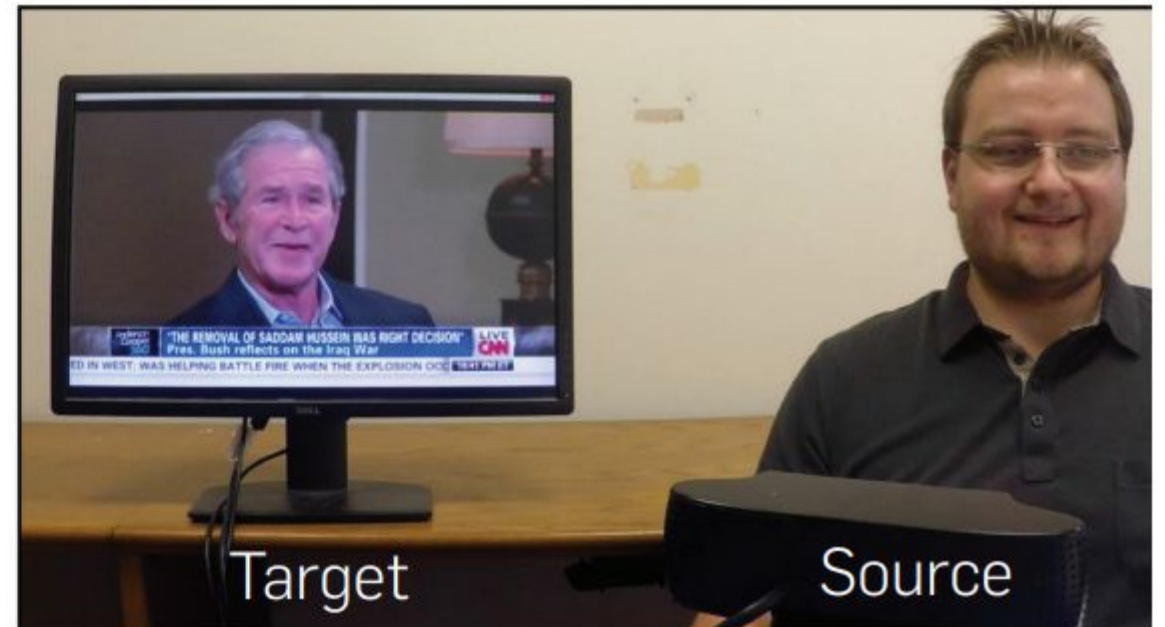
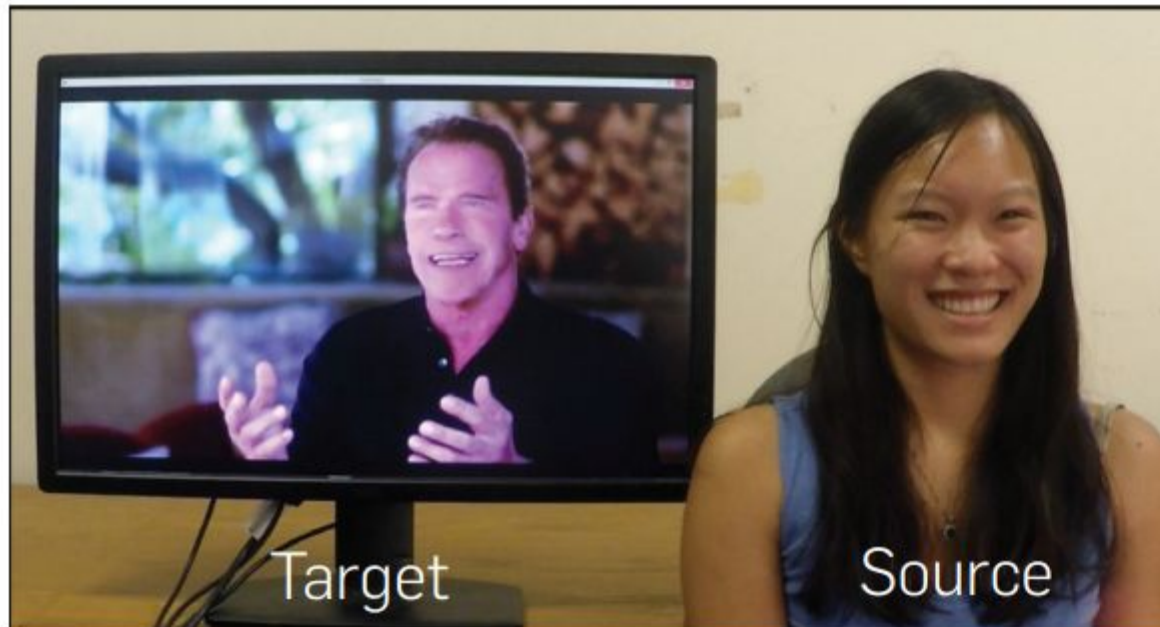
What about doing this with video?

- For video, more things to do
- Expression changing, mouth opening and closing
 - A morphable model that can change the facial expression is needed
- Also, the head is moving around
 - We need to predict the head orientation, translation too



Face2Face (Thies et al. 2016)

- Controlling another person in a video with your own face



Extra Basis for Expression

- Another set of basis are computed from different expressions of subjects and added

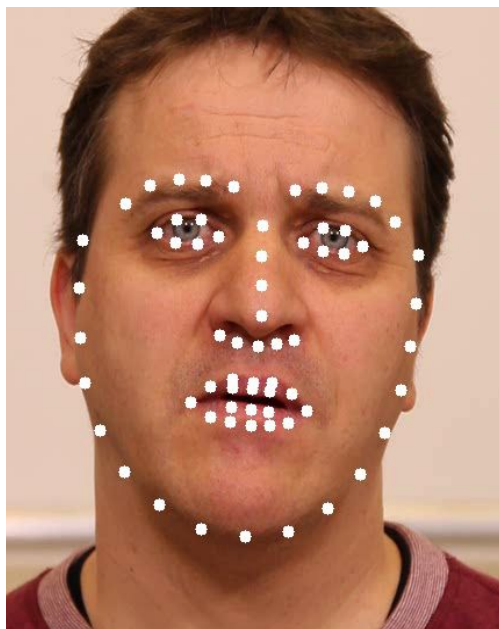
$$\mathbf{S}_{\text{model}} = \bar{\mathbf{S}} + \sum_{i=1}^{m-1} \alpha_i \mathbf{s}_i + \sum_{k=1}^N \delta_k \mathbf{b}_k^{\text{exp}}$$

where $\mathbf{b}_k^{\text{exp}}$, δ_k are the extra basis and coefficients to produce different expressions, computed from people with different expressions



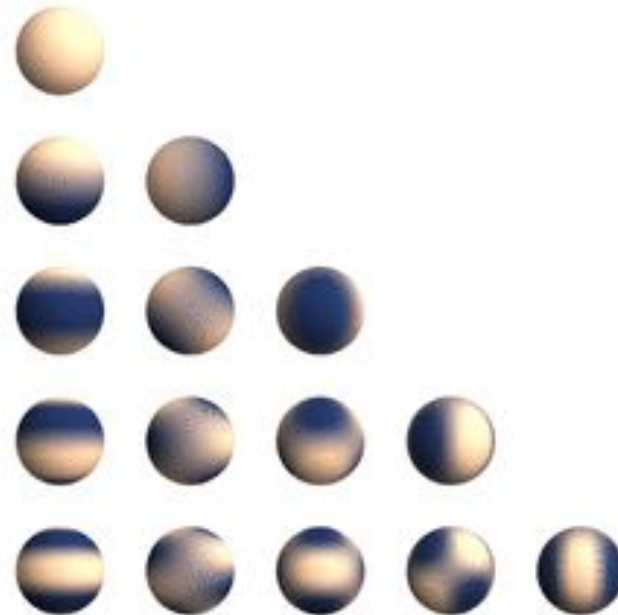
Aligning the Face Model with the Face Image

- We need to further predict the translation and rotation of the face
- Also need to make sure the important feature points (eyes, mouth, nose, chin etc) align well
- Using feature points that are detected by computer vision techniques is very useful



Lighting with spherical harmonics

- Representing the lights by spherical harmonics representation
- Something like Fourier basis of a function on a sphere domain
- Can represent natural lighting condition with a smaller number of coefficients



Monocular 3D Face Reconstruction

- Parametric 3D face model

$$p = (\text{Pose}, \text{Expression}, \text{Identity}, \text{Lighting}) \in \mathbb{R}^{257}$$

Diagram illustrating the components of the parametric 3D face model p :

- Pose**: A 3D coordinate system with axes labeled +X, +Y, and +Z.
- Expression**: A 3D face model showing an open mouth.
- Identity**: A 3D face model showing a neutral expression.
- Lighting**: A color gradient sphere representing lighting parameters.

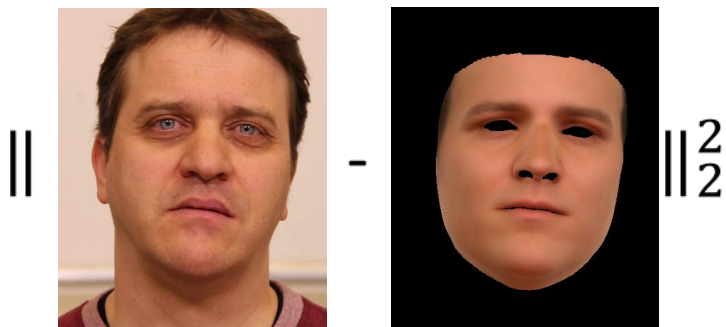
$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

Monocular 3D Face Reconstruction

- Parametric 3D face model

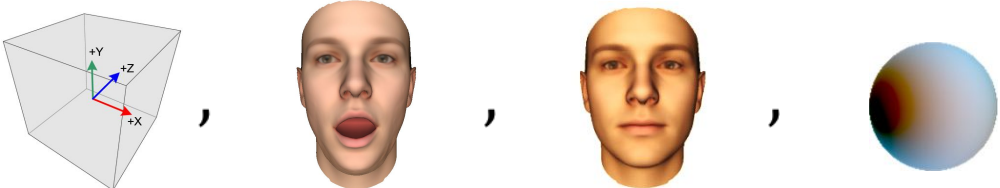
$$p = (\text{Pose}, \text{Expression}, \text{Identity}, \text{Lighting}) \in \mathbb{R}^{257}$$

$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$



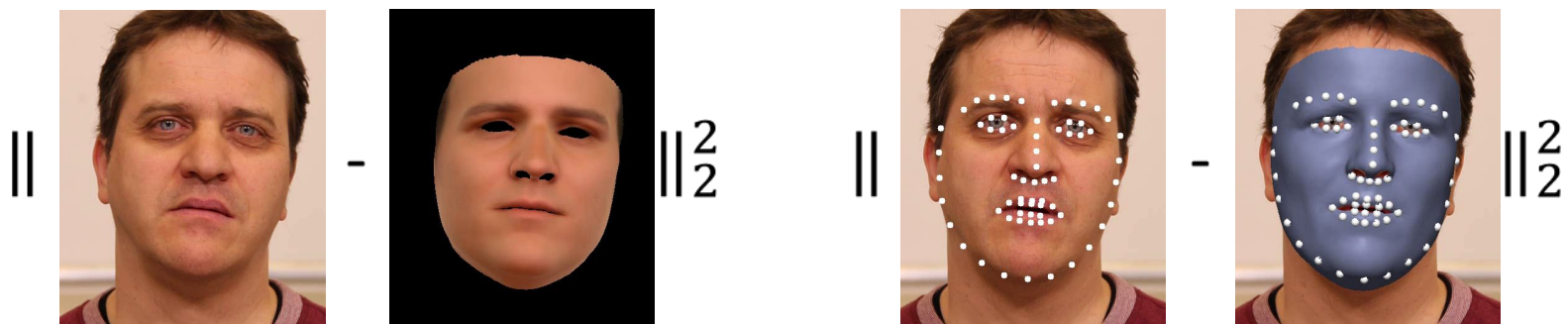
Monocular 3D Face Reconstruction

- Parametric 3D face model

$$p = (\text{Pose}, \text{Expression}, \text{Identity}, \text{Lighting}) \in \mathbb{R}^{257}$$


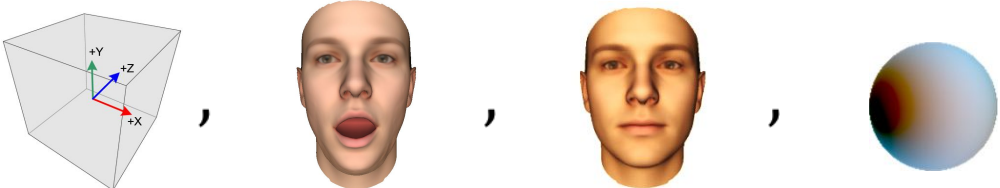
Pose Expression Identity Lighting

$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$



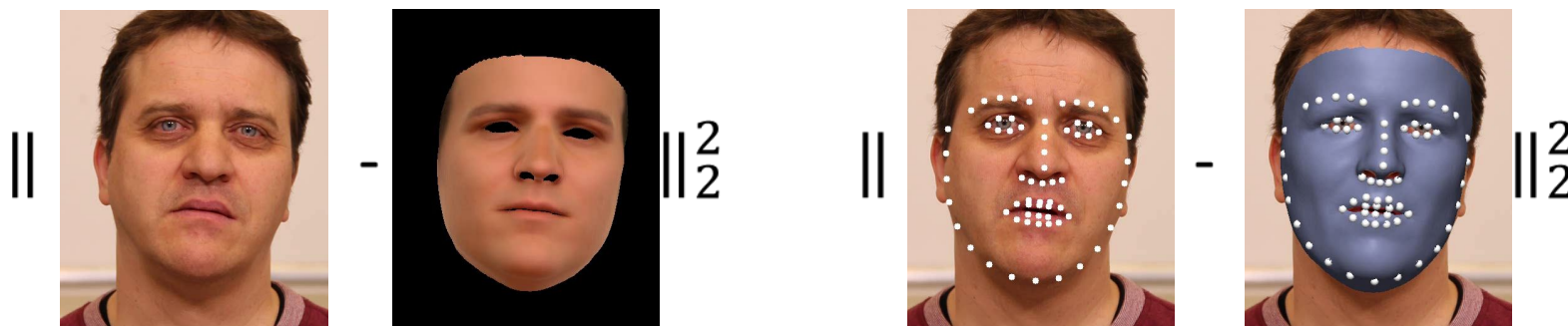
Monocular 3D Face Reconstruction

- Parametric 3D face model

$$p = (\text{Pose}, \text{Expression}, \text{Identity}, \text{Lighting}) \in \mathbb{R}^{257}$$


Pose Expression Identity Lighting

$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$



Statistical and temporal regularization

Garrido et al., ToG 2016

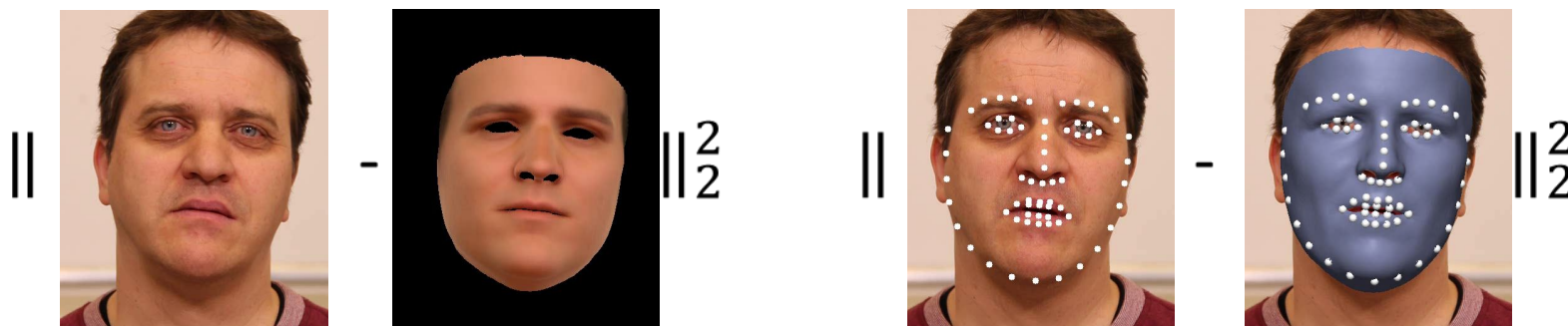
Monocular 3D Face Reconstruction

- Parametric 3D face model

$$p = (\text{Pose}, \text{Expression}, \text{Identity}, \text{Lighting}) \in \mathbb{R}^{257}$$

Pose Expression Identity Lighting

$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$



Statistical and temporal regularization

Garrido et al., ToG 2016

Monocular 3D Face Reconstruction

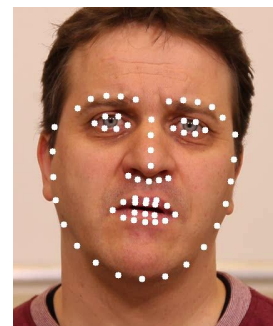
- Parametric 3D face model

$$p = (\text{Pose}, \text{Expression}, \text{Identity}, \text{Lighting}) \in \mathbb{R}^{257}$$

$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

- Eye model

$$e = (\text{Eye Model}) \in \mathbb{R}^4$$



Saragih et al.,
FG 2011

Expression Transfer

- Now we can fit a 3D model to the source video
- Let's think of changing the contents of the person in the target video
- Changing the head motion, the expression of the person, the mouth movements




Producing the target model

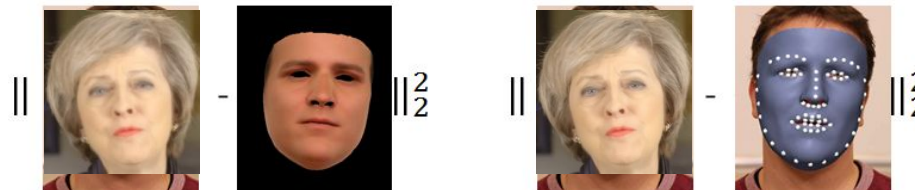
- The previous step is repeated for the target video
- We can then obtain the identity and expression basis for the target

Monocular 3D Face Reconstruction

- Parametric 3D face model

$$p = (\text{Pose}, \text{Expression}, \text{Identity}, \text{Lighting}) \in \mathbb{R}^{257}$$


$$\min_p E(p) = E_{\text{photo}}(p) + E_{\text{land}}(p) + E_{\text{reg}}(p)$$

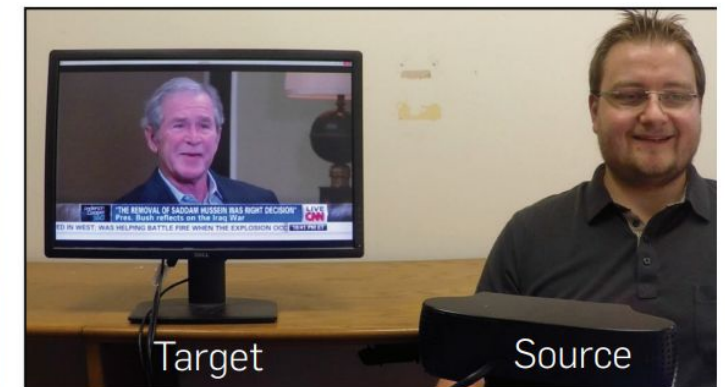
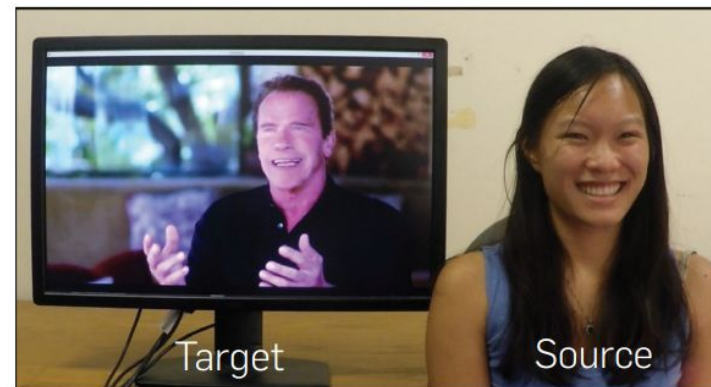


Statistical and temporal regularization

[Garrido et al., ToG 2016](#)

Expression Transfer (2)

- Transferring the expression changes from the source to the target actor is done by deformation transfer
- Assuming source identity α^S and target identity α^T fixed, transfer takes as input the neutral δ_N^S , deformed source δ^S , and the neutral target δ_N^T expressions.
- Output is the transferred facial expression δ^T directly in the reduced sub-space of the parametric prior.

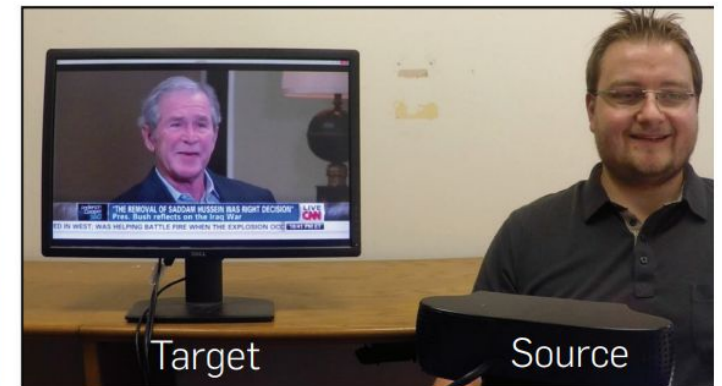
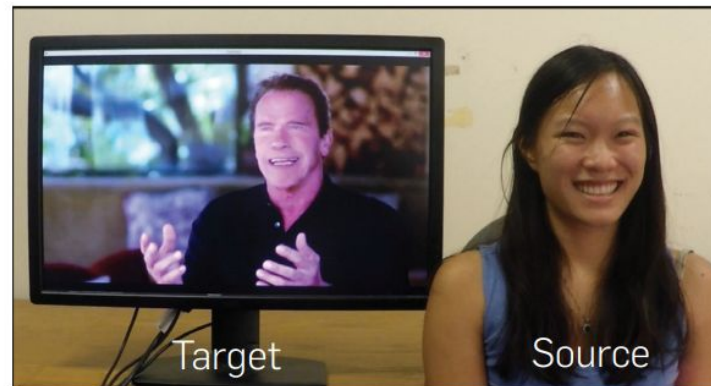


Expression Transfer (3)

- We first compute the source deformation gradients $\mathbf{A}_i \in \mathbf{R}^{3 \times 3}$ that transform the source triangles from neutral to deformed.
- The deformed target $\hat{\mathbf{V}}$ is then found based by solving a linear least-squares problem

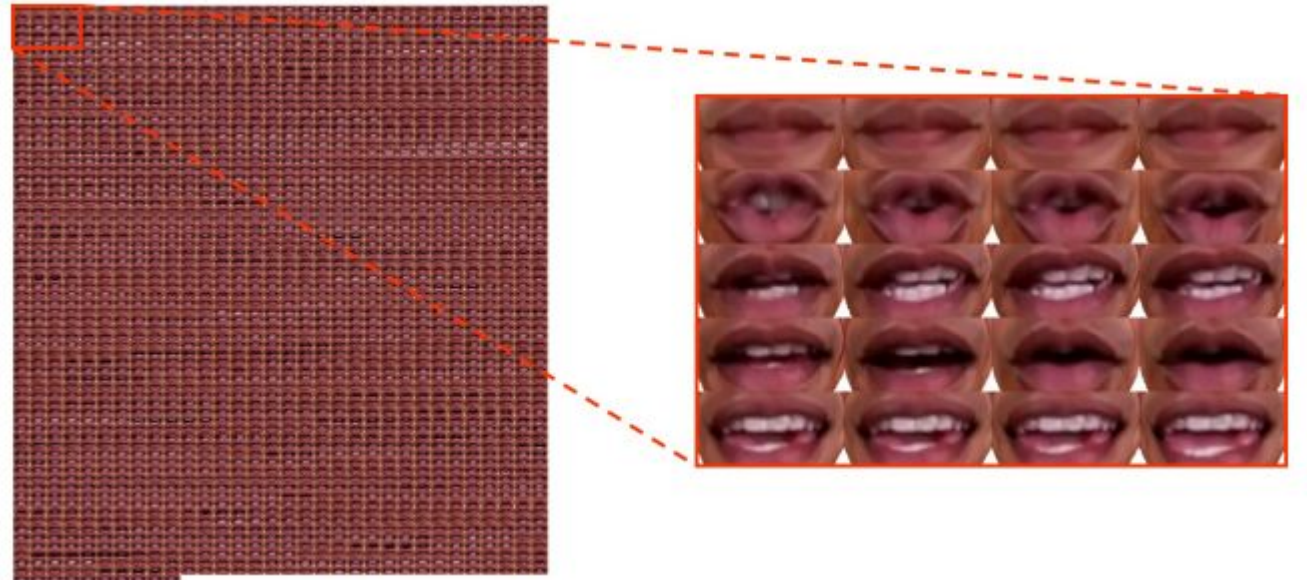
$$E(\delta^T) = \sum_{i=1}^{|\mathcal{F}|} \left\| \mathbf{A}_i \mathbf{V}_i - \hat{\mathbf{V}}_i \right\|_F^2$$

where neutral state \mathbf{V} is the neutral state of the triangle



Mouth Retrieval

- For a given transferred facial expression, it is necessary to synthesize a realistic target mouth region.
- To this end, the best matching mouth image from the target actor sequence is retrieved and fitted to the deformed target model



Demo video

-



Overview

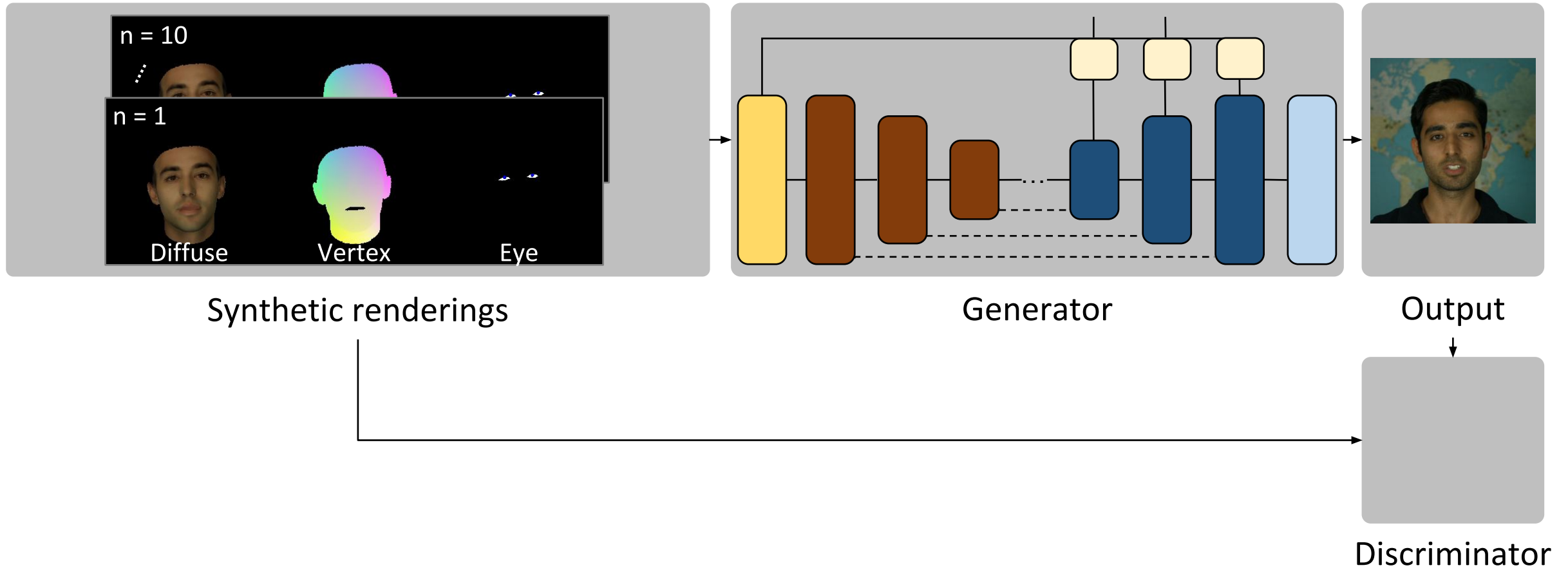
- Fitting 3D face models into 2D images (Blanz and Vetter '99)
- Controlling the person in the video by your own face (Thies et al. 2016)
- Further improvement by deep learning (Kim et al. 2018)

Deep Video Portrait (Kim et al. 2018)

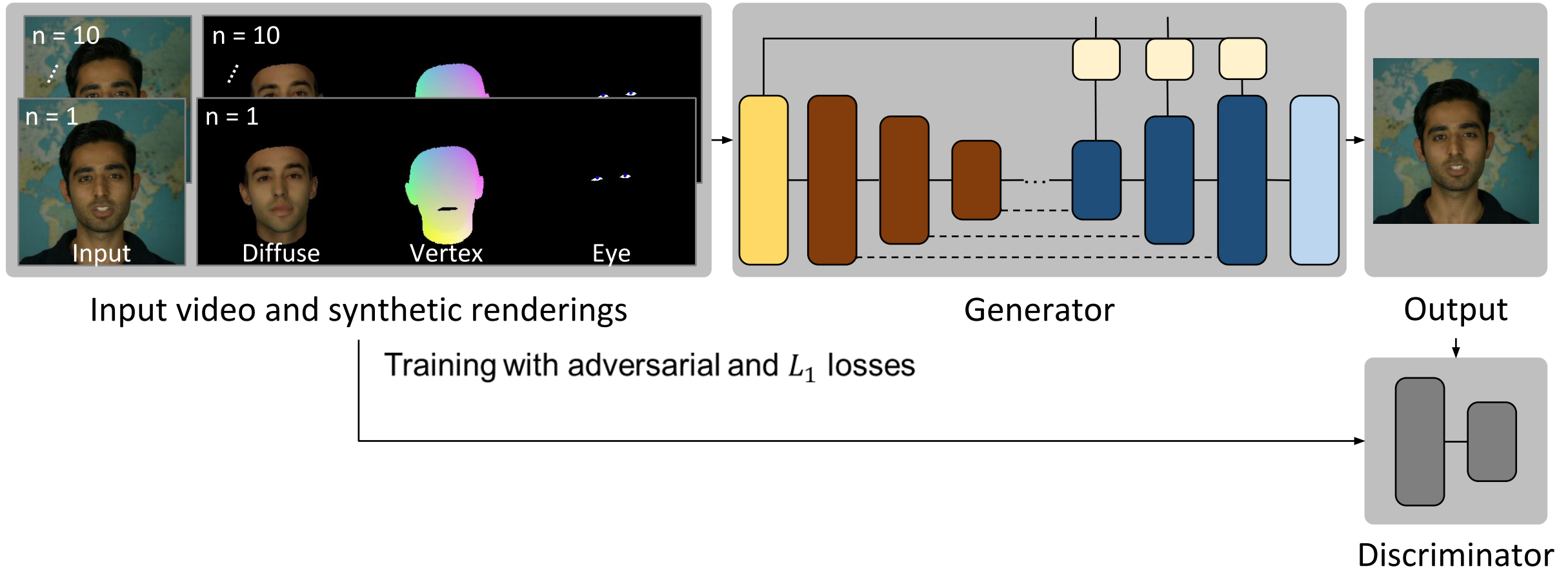
- A further improvement of the Face2Face by making use of deep neural networks
- Allowing the head pose to be edited too (changing the orientation, translation)
- Need to produce realistic background, shadows, hair motion



Rendering-to-Video Translation Network



Rendering-to-Video Translation Network



Interactive Editing



2x Speed

Summary

- Very realistic fake videos can be produced nowadays
- Digital dub are frameworks to convert the input source video to a target person, whose images/videos are available
- PCA models are used to represent the faces – the parameters are computed by fitting the face model to the target video. (Face2Face)
- The results are further passed to a generator network that is trained by adversarial training (Deep Video Portrait)

References

- Volker Blanz and Thomas Vetter , A Morphable Model For The Synthesis Of 3D Faces, SIGGRAPH 1999
- Thies et al. Face2Face: Real-time Face Capture and Reenactment of RGB Videos, CVPR 2016
- Kim et al. Deep Video Portrait, SIGGRAPH 2018