micc

UNIVERSITÀ
DEGLI STUDI
FIRENZE

DINFO
Dipartimento di
Ingegneria dell'Informazione

Statistical 3D Face Reconstruction with 3D Morphable Models

Claudio Ferrari, Stefano Berretti, Alberto Del Bimbo

claudio.ferrari@unifi.it

www.micc.unifi.it/3dmm-tutorial/

Department of Information Engineering (DINFO) &
Media Integration and Communication Center (MICC)
University of Florence (UNIFI), Florence, Italy

Tutorial Goal

- The goal of this tutorial is to provide an overview of the 3D Morphable Model technique and its applications

- While being a relatively old technique, it is still widely used as it still plays an important role for many applications

- The tutorial is divided in 4 main parts:

  1. In the first part, we will present the original 3DMM formulation, its main concepts and limitations

  2. In the second part we will present a particular modification of the original 3DMM, so as to show how each step of the process contribute and influence the final model

  3. Hands on! Let’s practice what we have learned

  4. The last part is dedicated to the review of some recent works exploiting the 3DMM in conjunction with deep learning techniques, so as to show that the 3DMM is still quite important (necessary!) for developing deep learning based solutions

Outline

- Part 1: Introduction to the 3D Morhpable Model – **(9:00 – 10:00)**
  - Idea and basic model (Blanz &Vetter)
  - Dense registration
  - Components learning (PCA)
  - 3DMM fitting
  - Limitations
  - Evolutions and applications

- Part 2: A particular case: Dictionary Learning based 3DMM - **(10:00 – 10:45)**

Coffe Break

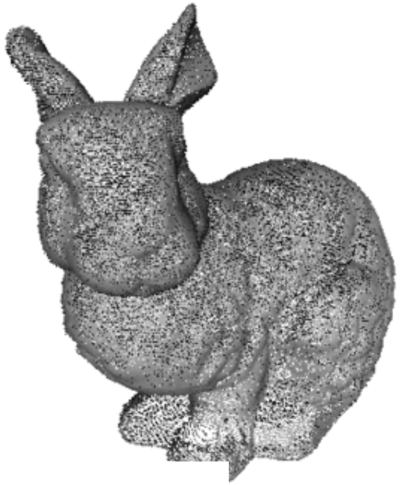
- Part 3: Hands on! Running some examples – **(11:00 – 11:30)**

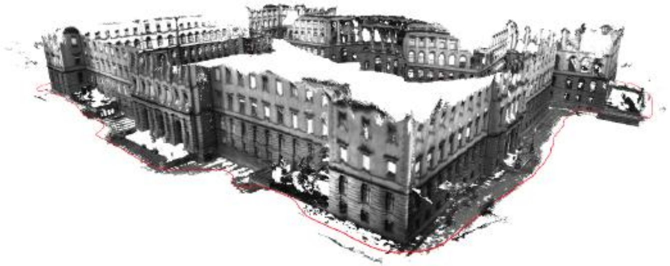
- Part 4: Deep Learning and 3DMM – **(11:30 – 12:15)**

- Part 5: Conclusions – **(12:15 – 12:30)**

Abstract

- The problem of reconstructing the 3D structure of objects or scenes from multiple or single images has been studied for more than three decades;
- Applications are numerous;
- Tons of solutions have been developed so far.





Abstract

- In this tutorial, we will focus on the particular case of reconstructing a 3D model of the human face from a single image by means of the **3D Morphable Model**.

Motivation

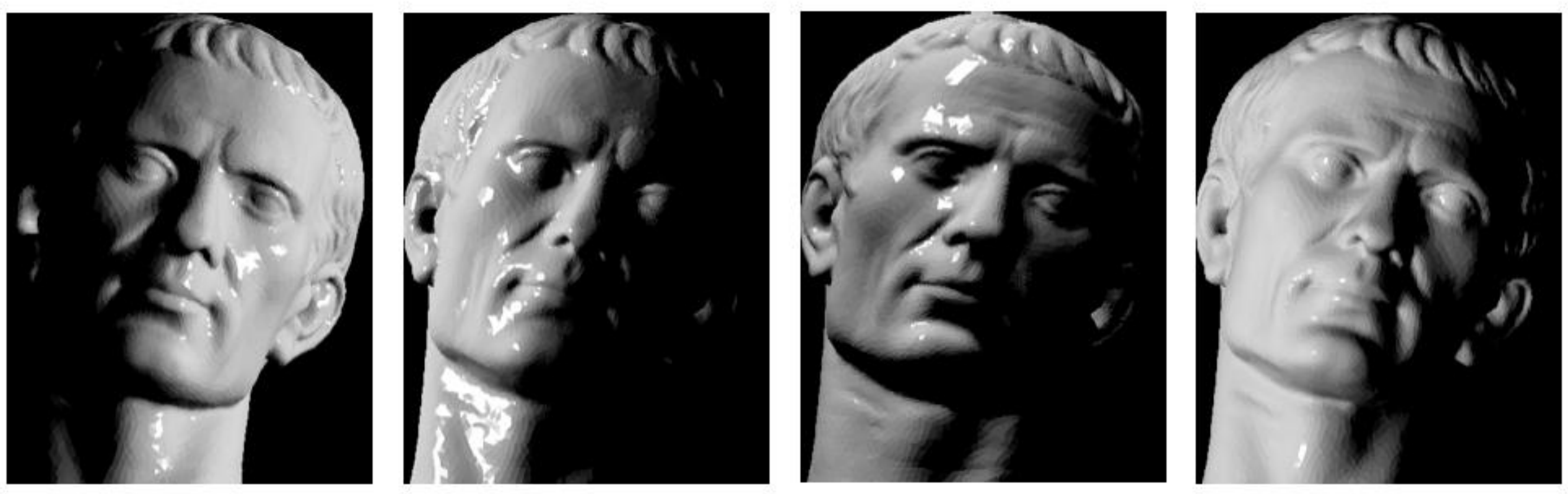
- The idea of deriving 3D information from 2D images using computer vision techniques is a research topic with a quite long tradition that dates back to ’80

- Now, remaining the 3D acquisition limited to certain constrained domain, the deployment of powerful machine learning tools has pushed forward this research area, with innovative and effective solutions appeared recently

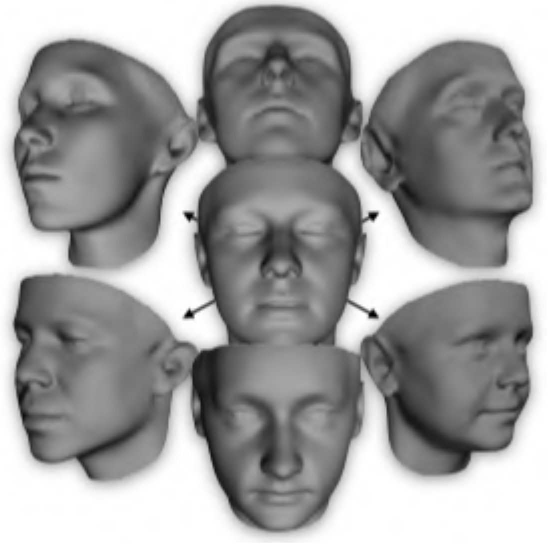
- Estimating the 3D geometry from single or multiple images under the most general conditions, where no a priori knowledge is available about the imaged scene and the capturing conditions is a very challenging task

Motivation

- To make the problem solvable to some extent, priors are usually assumed
- In the case of reconstructing a 3D model of the face, the prior knowledge can be in the form of camera parameters and reflectance properties of the face considering either a single image, as in the **shape from shading** solution (Horn and Brooks, 1989), or multiple images with different illuminations for the **photometric stereo** approach (Woodham, 1980)



Motivation



- Despite differences between subject to subject, faces have 3D shapes with well defined characteristics

- This inspired the idea that faces can be regarded as laying on a (unknown) shape manifold

- Moving on such manifold it is possible to pass from one face to another and generate new ones

- Following a similar intuition, Blanz & Vetter [*] first popularized the idea of capturing the face variability in a training set of 3D scans and constructing a statistical face model (3D Morphable Model, 3DMM)

- Such 3DMM includes an average component and a set of learned principal components of deformation allowing either to generate new face instances, or to deform and fit to 2D or 3D target faces

[*] V. Blanz and T. Vetter. "A morphable model for the synthesis of 3D faces", In *ACM Conf. on Computer Graphics and Interactive Techniques* (SIGGRAPH), 1999

The 3D Morphable Model

- We represent the geometry of a face with a shape-vector S = (X1, Y1, Z1, X2, Y2, Z2,….,Xn, Yn, Zn)T in R^{3n}, that contains the X, Y, Z coordinates of its n vertices (point-cloud)

- For simplicity, we assume that the number of valid texture values in the texture map is equal to the number of vertices

- We therefore represent the texture of a face by a texture-vector T = (R1, G1, B1, R2, G2, B2,….,Rn, Gn, Bn)T in R3n, that contains the R, G, B color values of the n corresponding vertices

- A 3D morphable face model is then constructed using a data set of **m** exemplar faces, each represented by its shape-vector **S**i and texture vector **T**i

The 3D Morphable Model

- In their seminal work, Blanz and Vetter first presented a complete solution to derive a 3DMM by transforming the shape and texture from a training set of 3D face scans into a vector space representation

- The main idea is that arbitrary new shapes **S**$_{mod}$ and textures **T**$_{mod}$ can be generated as a linear combination of some exemplar faces

$$\mathbf{S}_{mod} = \sum_{i=1}^{m} a_i \mathbf{S}_i \,, \quad \mathbf{T}_{mod} = \sum_{i=1}^{m} b_i \mathbf{T}_i \,, \quad \sum_{i=1}^{m} a_i = \sum_{i=1}^{m} b_i = 1.$$

- The 3D morphable model is defined as the set of faces (**S**$_{mod}$(a), **T**$_{mod}$(b)), parameterized by the coefficients **a** = (a1, a2,…,am) and **b** = (b1, b2,…,bm)

- Arbitrary new faces can be generated by varying the parameters **a** and **b** that control shape and texture

The 3D Morphable Model

- For a useful face synthesis system, it is important to be able to quantify the results in terms of their plausibility of being faces

- Therefore, the probability distribution for the coefficients $\mathbf{a}_i$ and $\mathbf{b}_i$ are estimated from the example set of faces

- This distribution enables us to control the likelihood of the coefficients $\mathbf{a}_i$ and $\mathbf{b}_i$ and consequently regulates the likelihood of the appearance of the generated faces

- A multivariate normal distribution is fit to the data set of faces, based on the averages of shape **S** and texture **T** and the covariance matrices $\mathbf{C}_S$ and $\mathbf{C}_T$ computed over the differences $\Delta S_i = S_i - \overline{S}$ and $\Delta T_i = T_i - \overline{T}$

- Principal Component Analysis (PCA) is applied to the scans

The 3D Morphable Model

- PCA performs a basis transformation to an orthogonal coordinate system formed by the eigenvectors **s**$_i$ and **t**$_i$ of the covariance matrices

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

- The probability for coefficients α_i is given by $\boldsymbol{\alpha}, \boldsymbol{\beta} \in R^{m-1}$

$$p(\vec{\alpha}) \sim exp[-\frac{1}{2} \sum_{i=1}^{m-1} (\alpha_i / \sigma_i)^2],$$

with σ_i^2 being the eigenvalues of the shape covariance matrix **C**$_S$

- The probability foe coefficients β_i is computed similarly

The 3D Morphable Model

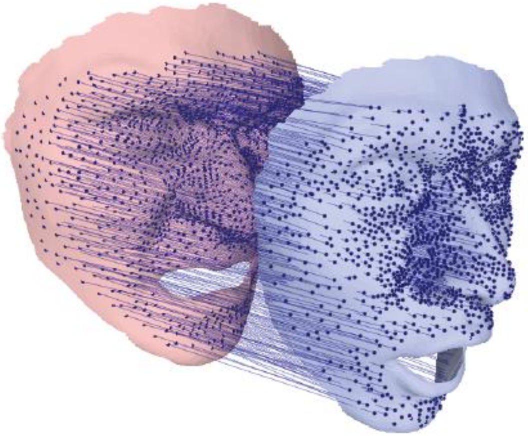
- The idea of using Principal Component Analysis (PCA) as linear dimensionality reduction technique to capture the characterizing dimensions of human faces, and so construct a 3D morphable face model capable of generating plausible faces is the key idea of the 3DMM

- The PCA computation and the morphing between faces requires full (i.e., dense) **correspondence** across all of the faces in the training set

Dense Correspondence: Why

- This step is fundamental to make the learned components meaningful
- This can be seen as a sort of mesh re-parametrization, where corresponding points must have the same anatomical reference across all the scans
- We need to make sure that the each vertex has the same semantic meaning across all the scans e.g. nose tip, lip corner
- All the scans also need to share the same number of vertices
- If these not hold, the principal components will not make any sense!

Optical Flow for Dense Correspondence



- A reference 3D face is chosen
- A modification of the gradient-based optical flow algorithm is used to establish correspondence between each scan and the reference face
- Each scan is represented by its RGB texture values and 3D coordinates
- Based on these correspondence, each mesh is re-parametrized with respect to the reference face so that corresponding points have the same index in the vector
- Can fail in case of “unusual” faces

The 3D Morphable Model: Summary

1. A dense point-to-point correspondence between the vertices of a set of 3D training faces is established

2. An average model is computed from the aligned scans

3. PCA is computed on the aligned scans to retrieve the principal components

4. The average model and the principal components constitute the 3DMM

5. New textured models are generated deforming the average model with a linear combination of the principal components

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

Segmented 3D Morphable Model

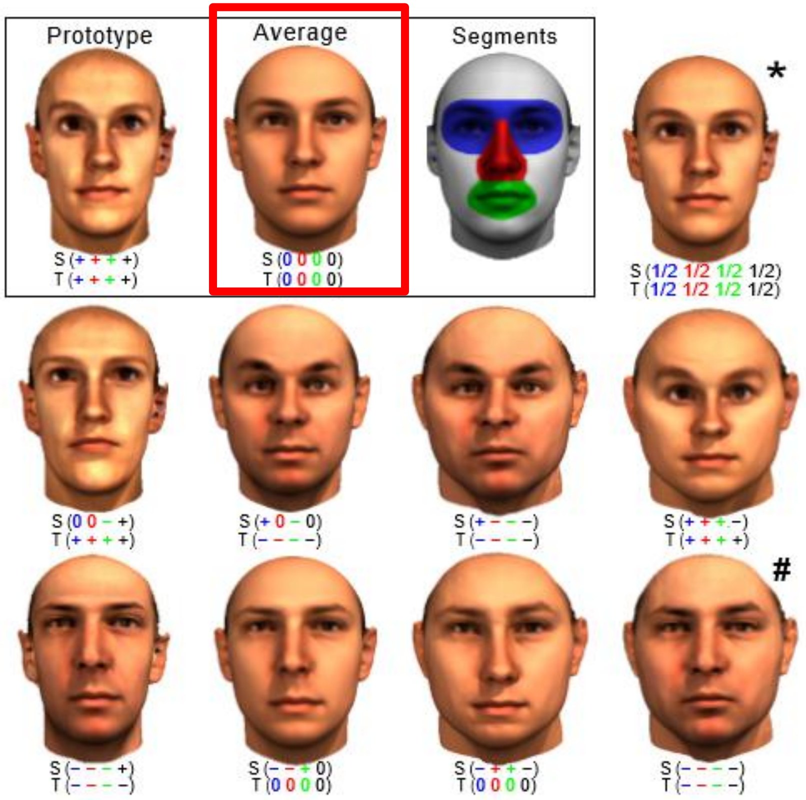
- To increase the versatility of the model, it is segmented to isolate specific areas i.e. eyes, nose, mouth

- Each area is deformed separately and an image blending technique is used to make smooth transitions between areas

- The deviation of a prototype from the average is added (+) or subtracted (-)

- A standard morph (*) is located halfway between average and the prototype.

- Adding and subtracting deviations independently for shape (S) and texture (T) on each of four segments produces a number of distinct faces

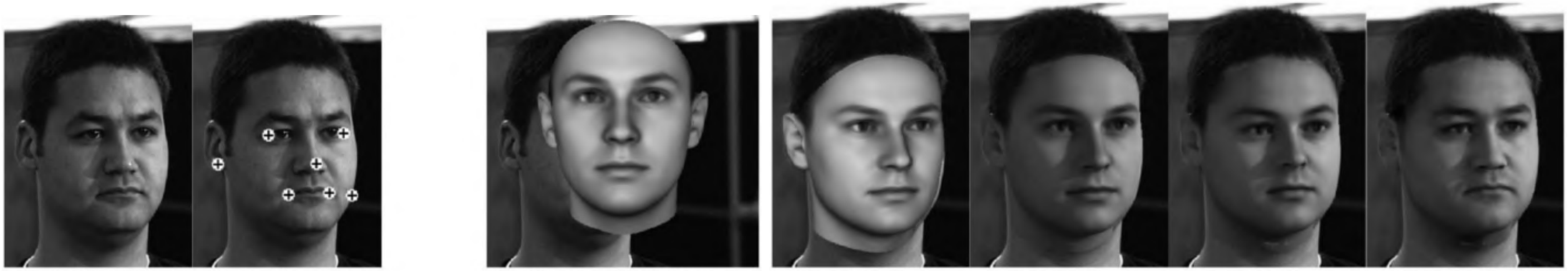


Matching the 3DMM to Images

- 3DMM is used either to generate arbitrary faces or to reconstruct a face image

- To reconstruct a face given an image the coefficients of the 3DMM are optimized along with a set of rendering parameters such that they produce an image as close as possible to the input

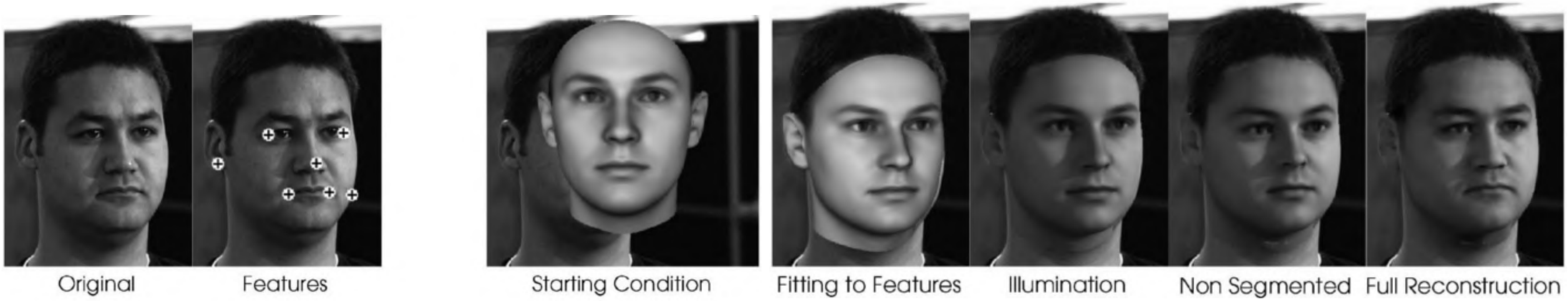
- In an analysis-by-synthesis loop, the algorithm creates a texture mapped 3D face from the current model parameters, renders an image, and updates the parameters according to the residual difference



Original Features Starting Condition Fitting to Features Illumination Non Segmented Full Reconstruction

Matching the 3DMM to Images

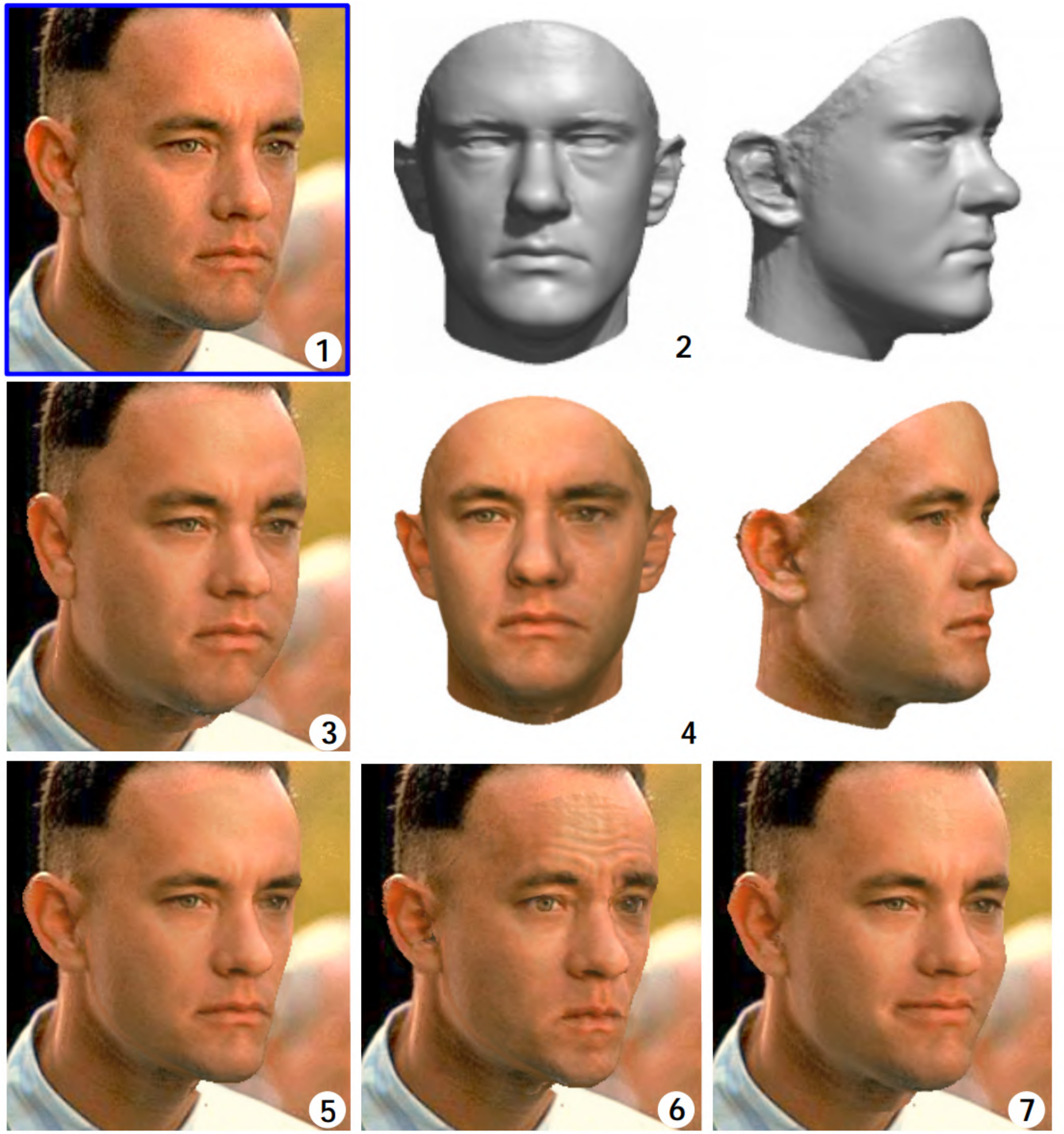
- 3DMM matching requires both 3DMM parameters and rendering parameters
- Rendering parameters contain the camera position, object scale, image plane rotation and translation, ambient light, directed light, color contrast, camera distance, light direction ecc.
- The 3DMM parameters are restricted to the vector space spanned by the training data, thus non-face-like surfaces are avoided
- Optimizing with respect to all those parameters is complex and costly



Facial Attributes Transfer

Original

Transferred attributes



Limitations of the standard approach

- 3DMM and its variants have been used in pose robust and 3D face recognition
- However, there are no convincing examples of 3DMMs applied to face analysis applications where facial expressions are involved
  - Difficulty with coping with noise, local deformations and topology variations

- Limitations derive from the methods used for the 3DMM construction
  - Training data: dataset of 200 middle aged Caucasians with limited face variations
  - Statistical tools applied to the data: if the correspondence is not accurate, most principal components include noise

Evolutions of the standard approach

- The techniques presented so far constitute the original 3DMM as first presented

- However, each step of the pipeline can be modified

- The main steps are:
  - Training scans collection
  - Dense registration
  - Components learning
  - Matching (fitting) to images

- In the following, we will analyze some relevant methods and applications

Training Data

- The data and their characteristics that are used to learn the components are of fundamental importance for the expressive power of the 3DMM

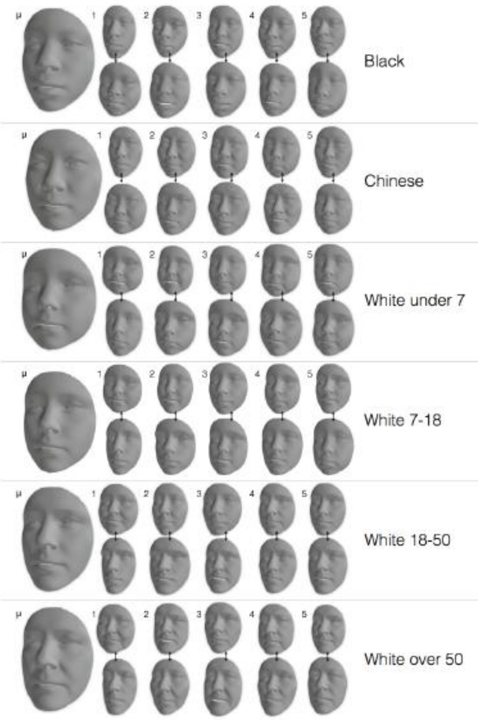
- For instance, if expressive scans are not included in the dataset, the 3DMM will not be able to model facial expressions

- The same applies for the texture, for instance if the RGB images are taken under controlled conditions e.g. inside a lab with constant illumination, then it will be hard to model more complex textures

- Desirable qualities of a 3D face dataset are:
  - Large variabilities in term of age, ethnicity, gender
  - Presence of expressive scans
  - High resolution of the scans
  - “In the wild” textures

Training Data

- To address this issue, methods to collect large scale 3D face datasets are lately being developed
- Booth et al.[1] developed an automatic method and collected a dataset of ~10K face scans so as to capture a larger spectrum of facial shapes
  - This model produces better reconstructions with respect to the original approach
- An “in the wild” texture model [2] was proposed to enhance and simplify the texture fitting: it is computed from 2D “in the wild” face images
- Instead of directly using the RGB value, the model is built using a dense feature representation of the image
- Complex rendering parameters (illumination, light ecc) are not necessary



[1] Booth, James, et al. "A 3d morphable model learnt from 10,000 faces." *CVPR*. 2016.
[2] Booth, James, et al. "3D face morphable models “in-the-wild”." *CVPR*. 2017.

Scans Dense Registration

- A large number of complex scans are beneficial for learning effective components and model the statistical variability of human faces

- On the other hand, it adds complexity to the dense registration problem

- The optical flow approach is prone to failures if the reference scan and the source scan are relatively diverse

- Moreover, the human face is a non rigid object that can significantly change its shape and the properties of the surface

ICP for Dense Correspondence

- **Iterative Closest Point**: Given a source set of points, compute for each point the nearest one in the target set

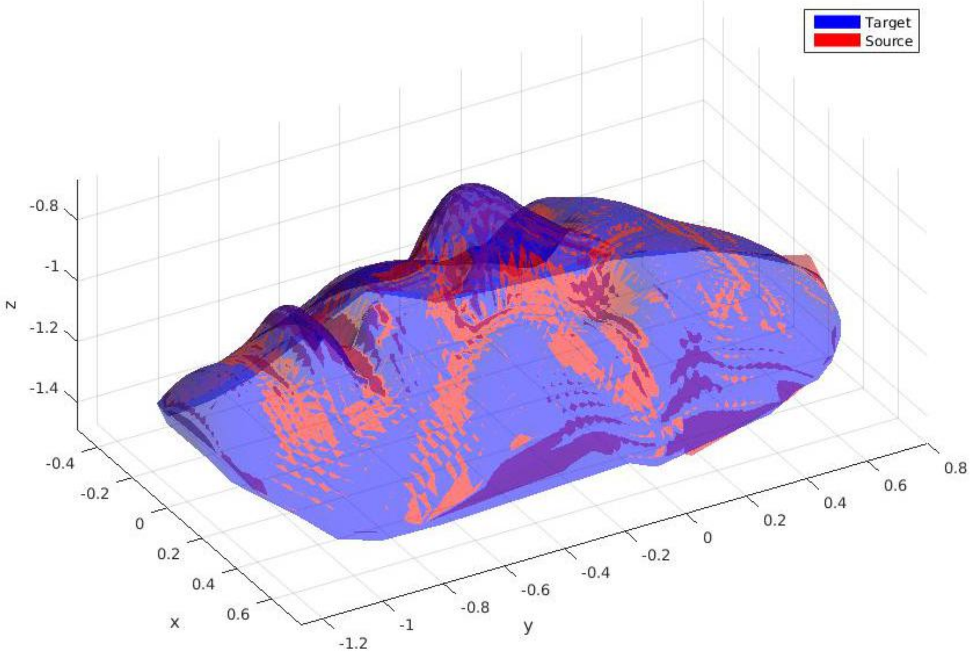
- Estimate a rigid transformation to bring the source points to the target

- Re-parametrize the source points

- Iterate until convergence

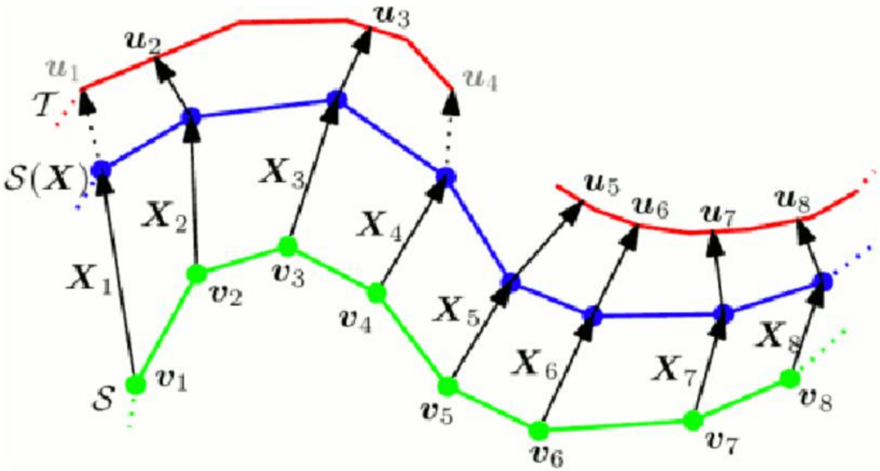
- Cannot handle holes or topological variations

- Not suitable for expressive scans



Non Rigid ICP for Dense Correspondence

- Overcomes the problem of holes and missing regions
- Estimate an (initial) rigid transformation to bring the source points to the target
- Apply local affine transformations weighed by a stiffness parameter to the points so as to match the target shape and eliminate non matching points
- Iterate until convergence
- Large topological differences still give troubles



Amberg, Brian, et al. "Optimal step nonrigid ICP algorithms for surface registration." CVPR, 2007.

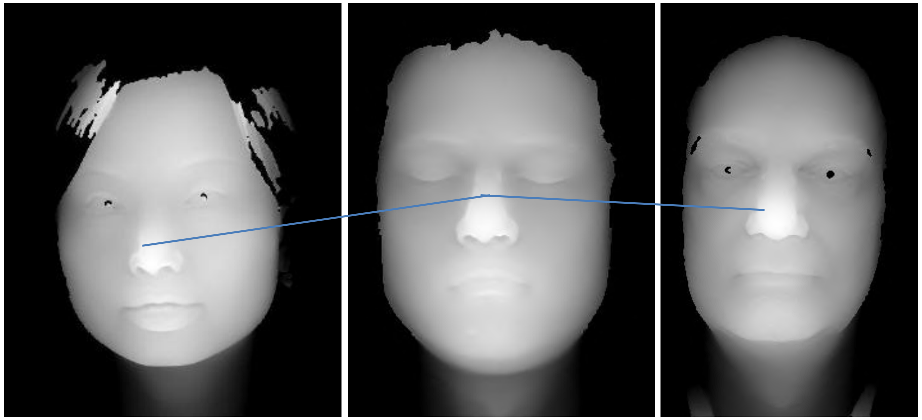
Other Methods for Dense Correspondence

- Many other methods exist which take into account several aspects of a point cloud (or mesh)

- Sparse points (landmarks) can be used to guide the initial estimation

- Some ICP variants account for local surface properties like normal vectors or curvatures

- Range images (or depth images) are used as well to solve the problem by exploiting image features



Components Learning: Mixture of Gaussians

- Now that we have a training set of registered 3D scans, we might want to apply a statistical tool to learn the deformation components

- In [1] it is shown that if the training data spans a large spectrum of variabilities, modeling the underlying distribution with a mixture of Gaussians is more suitable than PCA

- The distributions of the subpopulations might have very different means

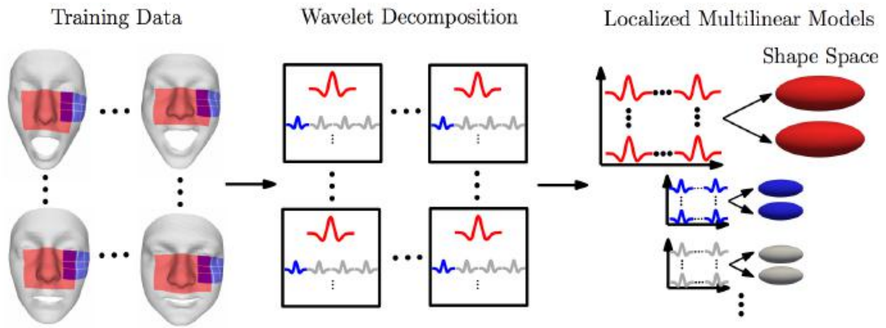
[1] Koppen, Paul, et al. "Gaussian mixture 3D morphable face model." *Pattern Recognition* (2018)

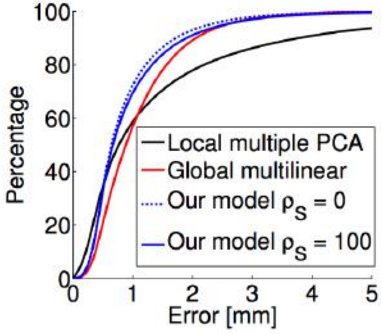
Components Learning: Multilinear Models

- In [2], the face is decomposed using a wavelet transform and a set of localized multilinear models are learnt to account for identity and expression separately

- The de-correllated wavelet coefficients allow learning many independent low-dimensional models

- This method achieves better results with respect to the classic PCA (or local-multiple PCA)





[2] Brunton, et al. "Multilinear wavelets: A statistical shape space for human faces ECCV, 2014.

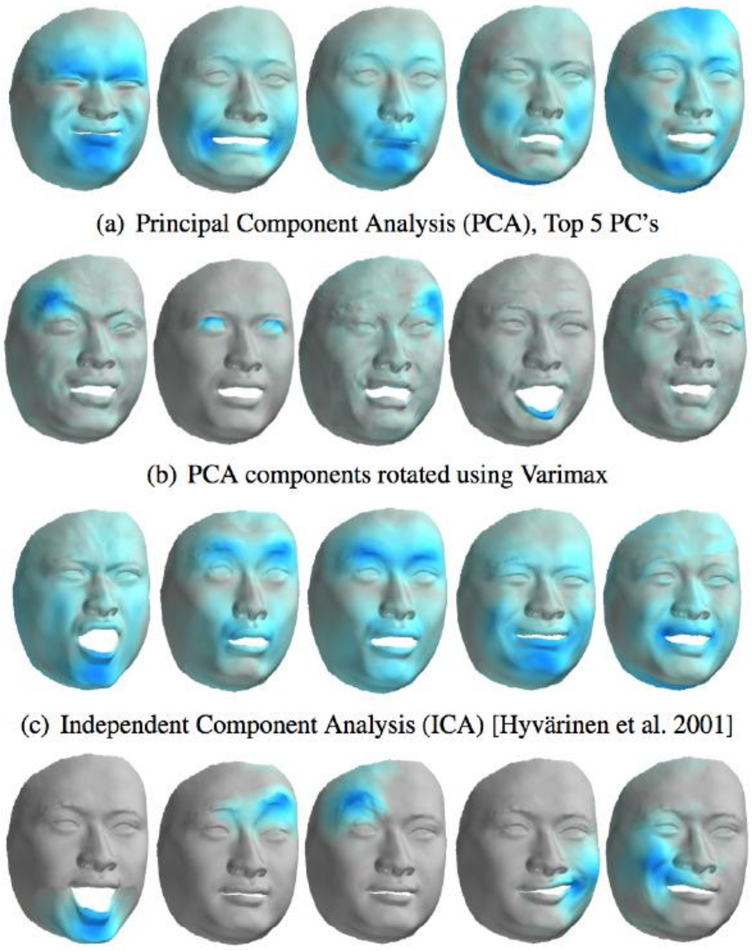
Components Learning: Sparse PCA

- If we want to model local deformations a global model might introduce noise in surrounding areas

- In [3], the components are learned by applying a sparse implementation of the PCA on mesh sequences

- This results in sparse components that are used to model local deformations

- Modeling differences between individuals is somewhat harder



(a) Principal Component Analysis (PCA), Top 5 PC's

(b) PCA components rotated using Varimax

(c) Independent Component Analysis (ICA) [Hyvärinen et al. 2001]

(d) Proposed Sparse Localized Deformation Components

[3] Neumann, Thomas, et al. "Sparse localized deformation components." *ACM Transactions on Graphics (TOG)* 32.6 (2013): 179.

3DMM Fitting

- While the generation of arbitrary faces starting from the underlying statistical model does not change i.e. linear combination of coefficients, many approaches to fit the 3DMM to face images are possible

- These are divided roughly in:

  - **Analysis-by-synthesis**: minimize the difference between the input image and a rendered image of the 3DMM, updating the parameters iteratively

  - **Geometric**: the 3DMM parameters are chosen so as to minimize the euclidean re-projection error between a set of landmarks (similar to AAM) or curves

  - **Feature Based**: similar to the landmark based but the optimization is intended to either minimize the difference in a feature space (HOG, SIFT …) or learn a regressor to map the

3DMM Fitting: Analysis by synthesis

- This kind of approaches aim at finding the fitting parameters that generate a synthetic image as similar as possible to the input one in an iterative manner

- Require initialization

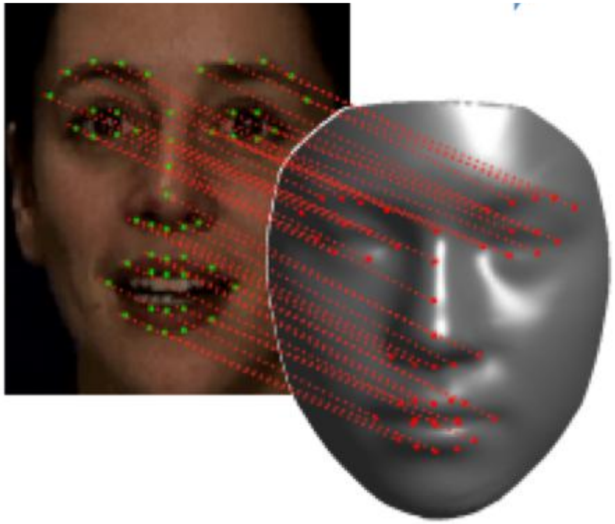
- They are accurate but complex: a lot of parameters to define and optimize
  - Camera model
  - Illumination model
  - 3DMM shape and texture
  - Optimization algorithm

- Usually slow

- Get stuck in local minima

3DMM Fitting: Geometric

- These approaches deform the 3DMM so as to match some geometric elements extracted from the face image
- Usually, a corresponding set of landmarks (in 2D and 3D) are used to estimate the projection matrix to project the 3DMM onto the image plane
- The 3DMM is deformed so as to minimize some geometric elements:
  - Landmarks position
  - Curves
  - Contours
  - …
- Faster and not computationally expensive
- Less accurate



3DMM Fitting: Feature Based

- Sort of ”mix” of the previous methods
- Deform the 3DMM so as to render an image as similar as possible to the input one in a **feature space** (HOG, SIFT …)
- Many rendering parameters (e.g. illumination model) are not needed since these are already encoded in the features
- Features can be extracted on a dense or scattered set of points
- Can be used in conjunction with geometric methods
- Regressors can be trained to map the feature differences to parameters update [4]

[4] Zhu, Xiangyu, et al. "Discriminative 3D morphable model fitting." Face and Gestures, 2015.

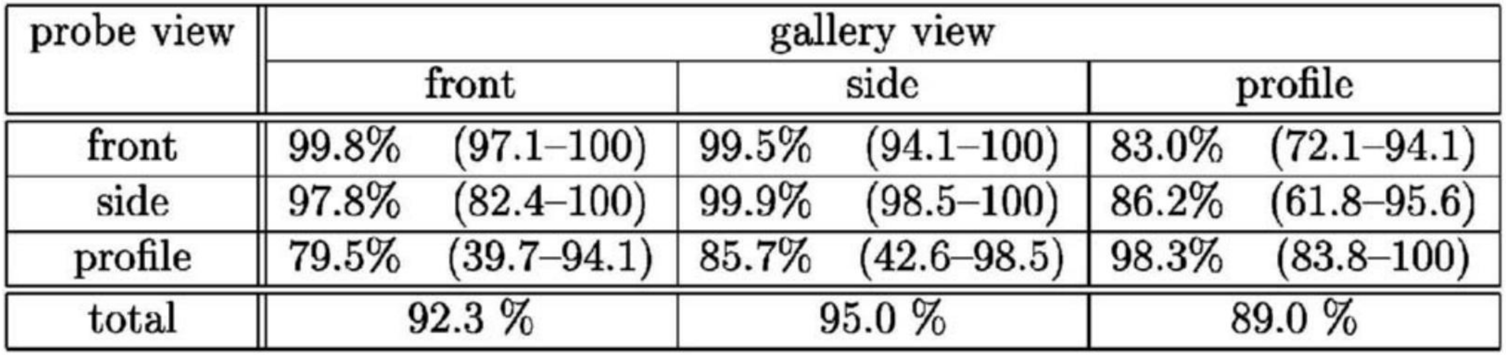
Applications

- The 3DMM has many different applications in image face analysis, from computer graphics for face inverse lighting and reanimation, craniofacial surgery to 3D shape estimation from 2D image face data, 3D face recognition, pose robust face recognition etc.

- Its first application was pose robust face recognition (2D) [5]

- The 3DMM is fit to face images and recognition is performed by matching the shape and texture coefficients of the 3DMM
  - Same subjects (even in different poses) should have a similar set of coefficients

CMU-PIE dataset

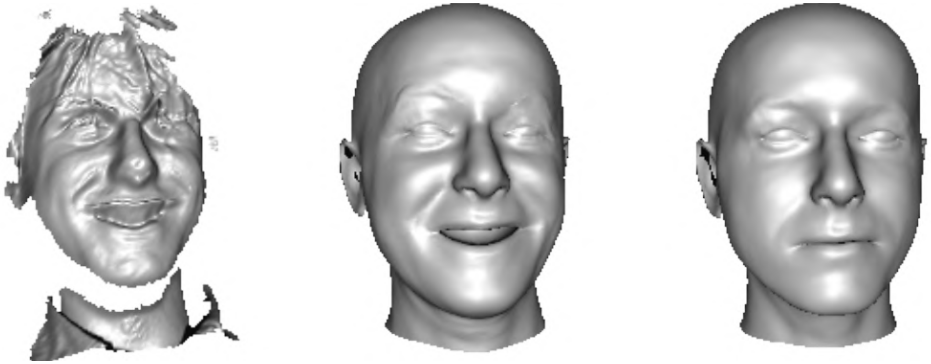


probe view	gallery view		
	front	side	profile
front	99.8% (97.1–100)	99.5% (94.1–100)	83.0% (72.1–94.1)
side	97.8% (82.4–100)	99.9% (98.5–100)	86.2% (61.8–95.6)
profile	79.5% (39.7–94.1)	85.7% (42.6–98.5)	98.3% (83.8–100)
total	92.3 %	95.0 %	89.0 %

[5] Blanz, Volker, and Thomas Vetter. "Face recognition based on fitting a 3D morphable model." IEEE TPAMI (2003)

Applications

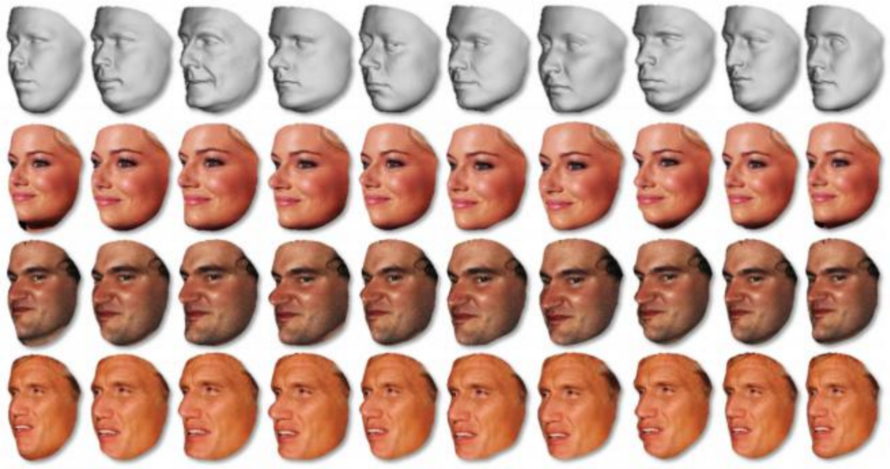
- Similarly, it has been applied for 3D expression invariant face recognition [6]
- Two separate PCA based models: *identity* and *expressions*
- To build the expression model, PCA is applied to the offset vector between an expressive scan and the neutral scan of the same subject
- The expression model is used to normalize the reconstructed model so as to remove the expression
- Recognition is performed comparing the coefficients vectors



[6] Amberg, Brian et al. "Expression invariant 3D face recognition with a morphable model." FG'08.

Other applications

- Being the 3DMM a generative model, another important (we will see that later…) application of the 3DMM is the generation of synthetic face images



- Or synthetic expressions

- 3D pose compensation or synthesis

- And so forth…

A Particular Case

- In the following, we will present a particular implementation of the 3DMM

- The goal was to build a model able to accurately fit expressive face images quickly so as to compensate the 3D head pose and render a frontal view of the face image

- We will revise the whole pipeline development so as to show how each step can be modified so as to reach a particular goal