



# 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

Claudio Ferrari
(UNIFI-MICC)

Giuseppe Lisanti
(UNIPV)

Stefano Berretti
(UNIFI-MICC)

Alberto Del Bimbo
(UNIFI-MICC)

# Face Modeling using a 3D Morphable Model

**C. Ferrari**, G. Lisanti, S. Berretti, A. Del Bimbo. "A Dictionary Learning based 3D Morphable Shape Model," ***IEEE Transactions on Multimedia***, 2017

# Our Contribution

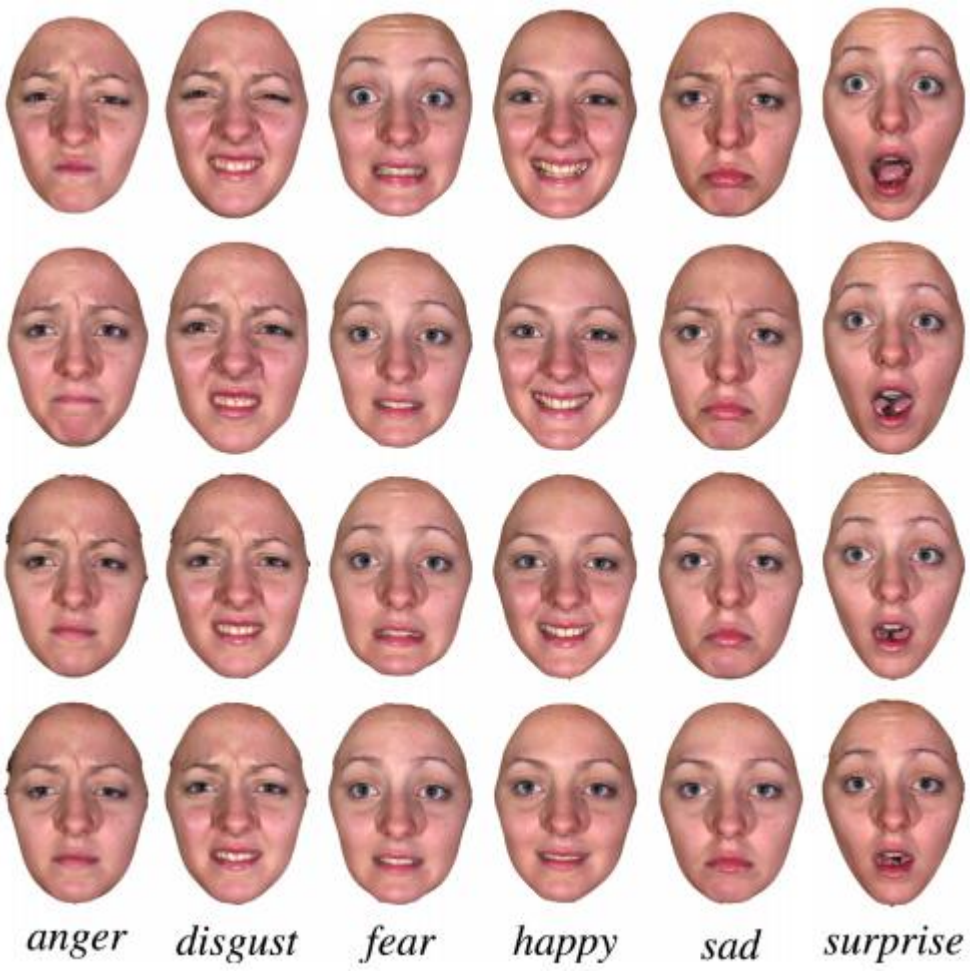
- We propose a new approach to the construction and fitting of a 3DMM
- The proposed 3DMM captures much of the large variability of human faces, thus opening the way to its use in fine grained face analysis
- It is grounded on three distinct contributions
  1. A new method to establish **dense correspondence** between scans even in the case of expressions with topological variations
  2. A new approach to capturing the statistical variability in training data that learns a **dictionary of deformations** from the deviations between each 3D scan and the average model (DL-3DMM)
  3. The application of 3DMM to **Action Unit detection** and **emotion recognition** from 2D images

# 3DMM Pipeline:  Training Data

- Following the 3DMM pipeline, let's first choose the training data…
- Grounding on the observations made, we aim at finding a dataset that include:
  - Balanced proportion of males and females
  - Reasonable range of ages
  - Different ethnicity groups
  - Face variations ( expressions )

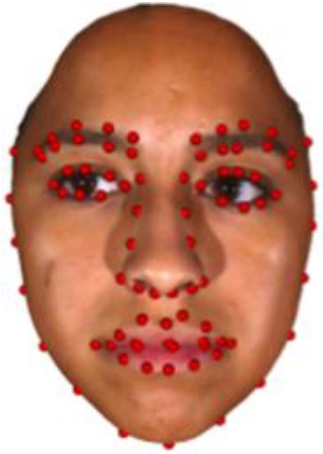
# The Binghamton University 3D Facial Expression (BU-3DFE) dataset

- **BU-3DFE**
  - Scans of **44** females and **56** males with age ranging from **18 to 70 years old**
  - Neutral plus **six expressions**: anger, disgust, fear, happiness, sadness, and surprise
  - **Four levels of expression intensity**, from low to exaggerated (2500 scans in total)
  - Subjects distributed across different ethnic groups, including *White, Black, Indian, East-Asian, Middle East Asian, and Hispanic Latino*



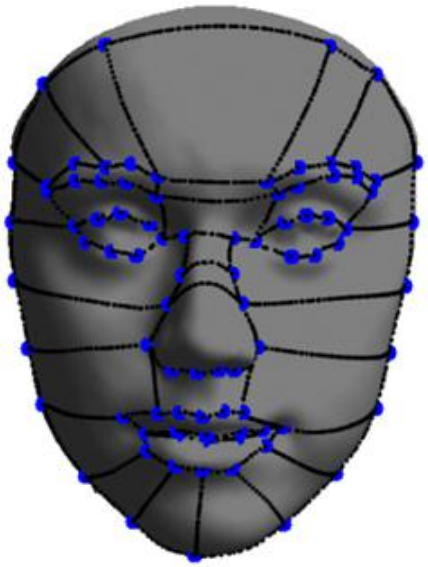
# 3DMM Pipeline: Dense Correspondence

- Now that we have found a suitable dataset, we need to put the scans in dense correspondence
- The task revealed itself to be very challenging mainly because of the presence of strong expressions that lead to significant topological variations of the surfaces
- Standard methods (e.g. non rigid ICP ) did not work well in this case
- **Idea:** Exploit the annotated landmarks

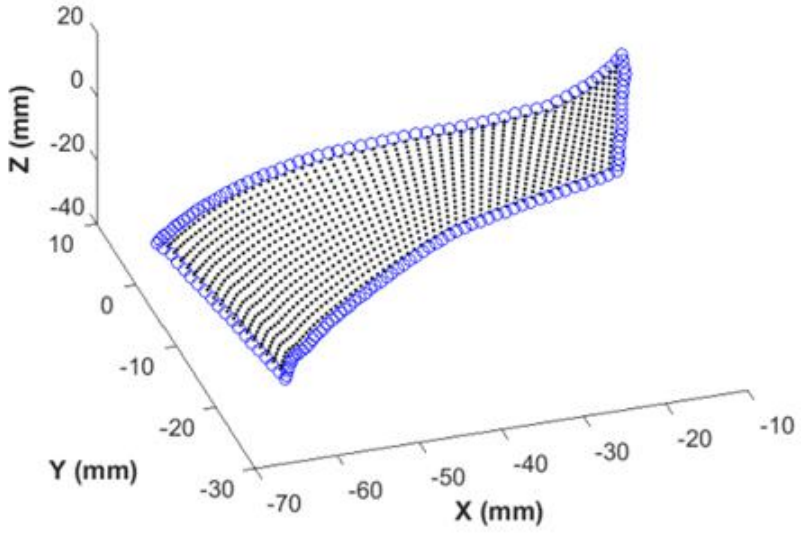


# Dense Correspondence

- The face is partitioned into a set of regions using geodesic paths between facial landmarks



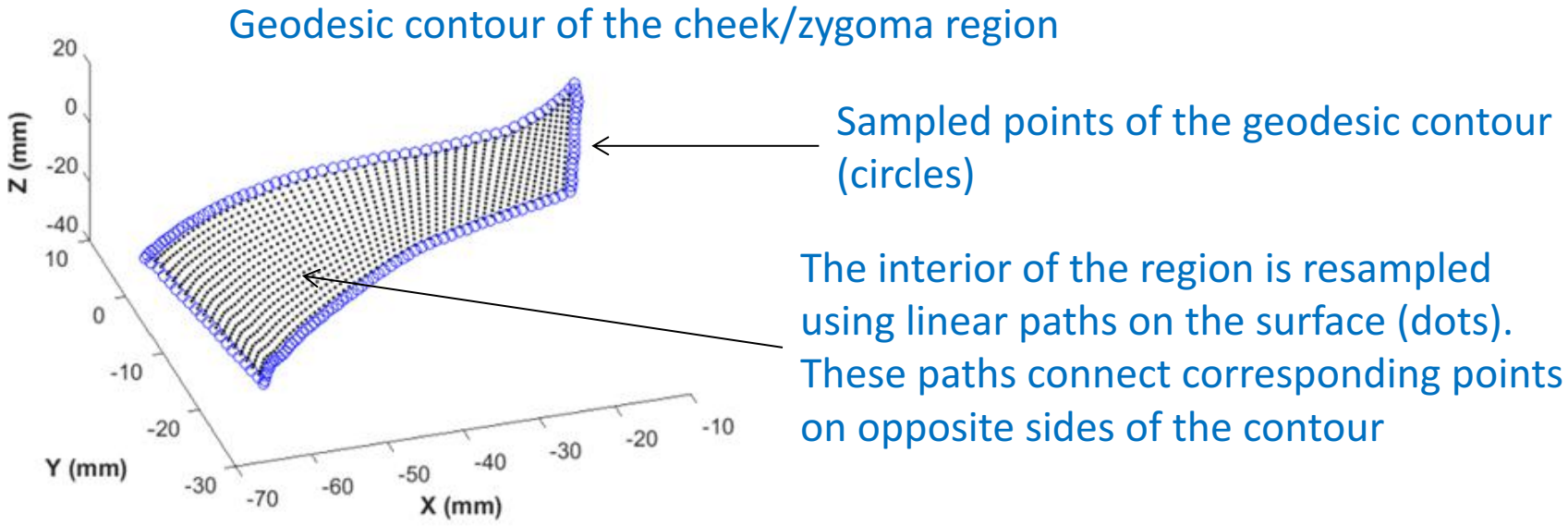
- The geodesic paths are resampled with a predefined number of points posed at equal geodesic distance one from the other



# Dense Correspondence

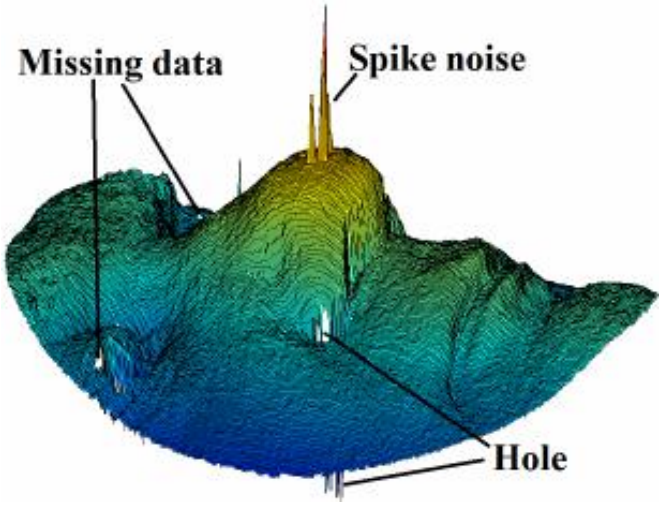
- The geodesic contour of a region guides the dense resampling of its interior surface
  - Pairs of sampling points on opposite side of a geodesic contour are connected with a linear path on the surface
  - This line is then sampled at the desired resolution

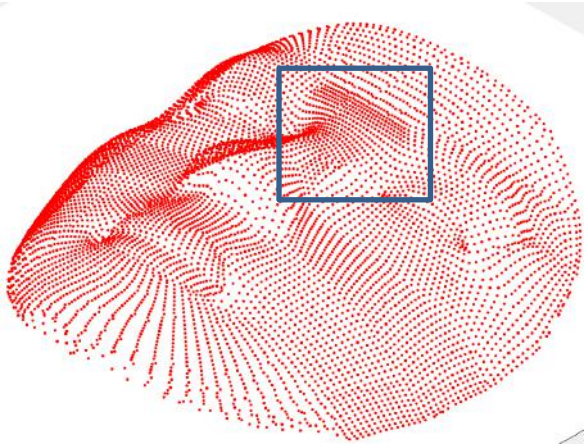
- Being based on the landmarks and their connections, this approach proved to be robust to facial expressions and topological changes



# Issues

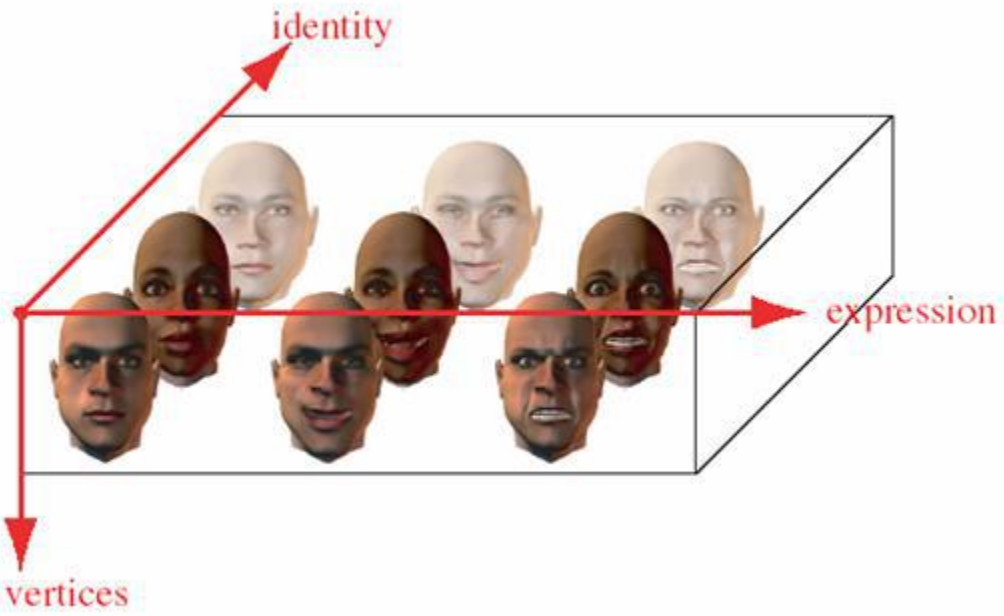
- Each region is re-sampled with a predefined number of points; if a region size changes significantly, we might get an over/under sampled region
- Not straightforward to get a uniformly sampled face model
- Huge contractions of a region might result in noisy resampling
- Large missing regions ( e.g. holes, occlusions ) might impair the resampling;



# 3DMM Pipeline: Components Learning

- We have seen that many solutions exist to learn the components
- We aim at capturing both global shape variations to model identity traits and local variations to model facial expressions
- To address both the tasks, we exploit a *Dictionary Learning* based solution



# Dictionary Learning 3DMM (DL-3DMM)

- We build our DL-3DMM by learning a dictionary of deformation components using *Online Dictionary Learning (DL) for Sparse Coding*

- Let **N** be the set of densely aligned training scans, each with **m** vertices

- Each scan is represented as a column vector $\mathbf{f}_i \in \mathbb{R}^{3m}$ , whose elements are the linearized X, Y , Z coordinates of all the vertices

$$\mathbf{f}_i = [\ X_{i,1}\ Y_{i,1}\ Z_{i,1}\ \ldots\ X_{i,m}\ Y_{i,m}\ Z_{i,m}\ ]^T \in \mathbb{R}^{3m}$$

- The average model m of the training scans is computed as $\mathbf{m} = \frac{1}{|N|} \sum_{i=1}^{|N|} \mathbf{f}_i$

- For each training scan fi, we compute the field of deviations vi with respect to the average model m

$$\mathbf{v}_i \leftarrow \mathbf{f}_i - \mathbf{m}\ , \quad \forall\ \mathbf{f}_i \in N$$

# DL-3DMM Construction

- DL is usually cast as an ***$l_1$*-regularized least squares problem**, but the **sparsity** induced by the $l_1$ **penalty** can lead to directions that deform the average model to a noisy or a discontinuous or punctured one

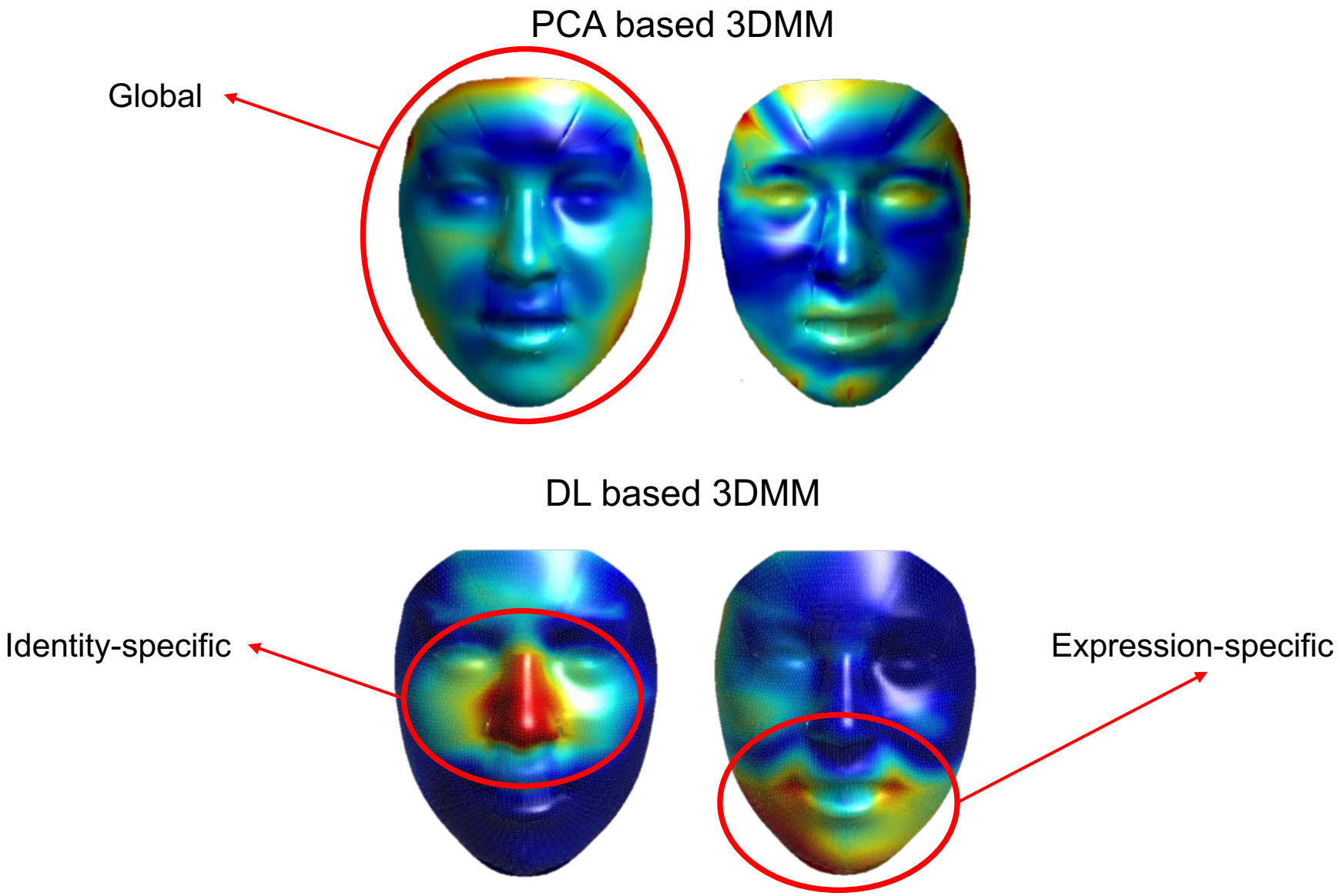
- We formulate the DL as an **Elastic-Net regression** that linearly combines
  - The sparsity-inducing ***$l_1$* penalty**, where $l_1$ norm acts as a **shrinkage operator**, reducing the number of non-zero elements of the dictionary
  - The ***$l_2$* regularization**, where the $l_2$ norm avoids **uncontrolled growth** of the elements magnitude, while forcing **smoothness**

- Defining $\ell_{1,2}(\mathbf{w}_i) = \lambda_1 \|\mathbf{w}_i\|_1 + \lambda_2 \|\mathbf{w}_i\|_2$ with $l_1$ and $l_2$ the **sparsity** and **regularization** parameters, respectively, we have

$$\min_{\mathbf{w}_i,\, \mathbf{D}} \frac{1}{|N|} \sum_{i=1}^{|N|} \left( \|\mathbf{v}_i - \mathbf{D}\mathbf{w}_i\|_2^2 + \ell_{1,2}(\mathbf{w}_i) \right)$$

- The average model **m**, the dictionary **D** and the diagonal elements of the matrix **W**, namely the vector $\hat{\mathbf{w}} \in \mathbb{R}^k$, constitute our DL-3DMM

# DL-3DMM Construction



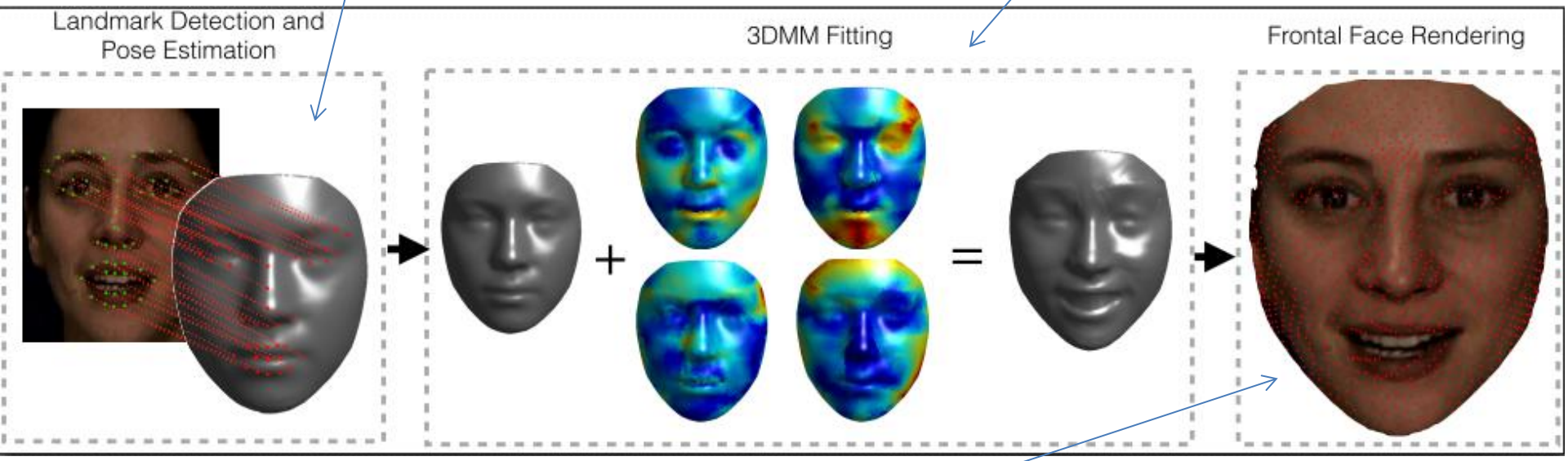
# 3DMM Pipeline: Fitting

- The choice of the fitting algorithm depends heavily on the final purpose, since each algorithm and technique has its own peculiarities
- Our final goal is to use the deformed 3D model to render a frontal view of the input face image and guide the extraction of local image descriptors for face recognition / expression recognition / Action Units detection
- Hence, we are not much interested in a precise and detailed reconstruction but rather, we need our fitting to be fast and robust to large shape deformations
- Indeed, we consider only the shape part, neglecting the texture model
- Our fitting algorithm falls in the “landmark-based” category

# 3DMM Fitting

1. The 3D **head pose** is estimated from the correspondence of 2D and 3D landmarks

2. The average 3D model is deformed using the basis components




3. A frontal face image is rendered showing a subsampling of the mesh vertices back projected onto the frontalized image. As a result of the fitting, **no vertices fall inside the open mouth region**

# 3D Pose Estimation

- In order to estimate the **pose**, we detect a set of 49 facial landmarks $\mathbf{l} \in \mathbb{R}^{2\times 49}$ on the 2D face image using a state of the art detector [*]

- An equivalent set of vertices $\mathbf{L} = \hat{\mathbf{m}}(\mathbf{I}_v) \in \mathbb{R}^{3\times 49}$ is manually annotated on the average 3D model, being $\mathbf{I}_v$ the set of indices of the vertices corresponding to the landmark locations

- Under an affine camera model, the relation between $\mathbf{L}$ and $\mathbf{l}$ is

$$\mathbf{l} = \mathbf{A} \cdot \mathbf{L} + \mathbf{T}$$

where $\mathbf{A} \in \mathbb{R}^{2\times 3}$ contains the affine camera parameters, and $\mathbf{T} \in \mathbb{R}^{2\times 49}$ is the translation on the image

[*] V. Kazemi and J. Sullivan. "One millisecond face alignment with an ensemble of regression trees". In *IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR), 2014

# 3D Pose Estimation

- Firstly, we subtract the mean from each set of points and recover the affine matrix $\mathbf{A}$ solving the least squares problem

$$\arg\min_{\mathbf{A}} \|\mathbf{l} - \mathbf{A} \cdot \mathbf{L}\|_2^2$$

with solution given by $\mathbf{A} = \mathbf{l} \cdot \mathbf{L}^+$, where $\mathbf{L}^+$ is the pseudo-inverse matrix of $\mathbf{L}$

- Direct estimation via least squares solution is possible since, by construction, facial landmark detectors assume a **consistent structure** of the 3D face parts so they do not permit outliers or unreasonable arrangement of the face

- The 2D translation is estimated as $\mathbf{T} = \mathbf{l} - \mathbf{A} \cdot \mathbf{L}$

- The estimated pose $\mathbf{P}$ is represented as $[\mathbf{A}, \mathbf{T}]$ and used to map each vertex of the 3DMM onto the image

# Efficiently Fitting the DL-3DMM

- Using the learned dictionary $\hat{\mathbf{D}} = [\hat{\mathbf{d}}_1, \ldots, \hat{\mathbf{d}}_k]$ , we find the coding that non-rigidly transforms the average model $\hat{\mathbf{m}}$ such that the projection minimizes the error in correspondence to the landmarks

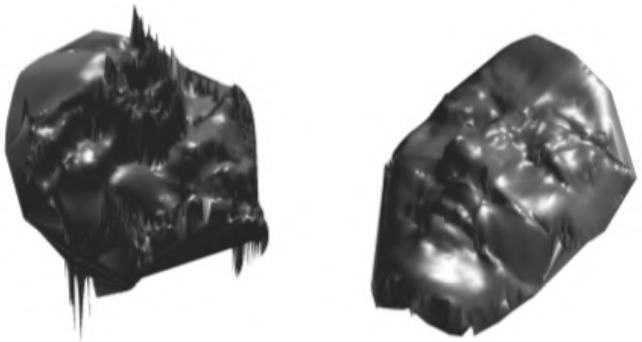
- The coding is formulated as a **regularized Ridge-Regression** problem (cost)

$$\arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v) - \sum_{i=1}^{k} \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)\alpha_i \right\|_2^2 + \lambda \left\| \boldsymbol{\alpha} \circ \hat{\mathbf{w}}^{-1} \right\|_2$$

regularization term

where $\circ$ is the Hadamard product

- The term $\hat{\mathbf{w}}^{-1}$ is used to associate a **reduced cost** to the deformation induced by the **most relevant components**



Deformed models when the **regularization term is removed**. The uncontrolled growth of the deformation coefficients $\alpha$ leads to excessive deformations

# Efficiently Fitting the DL-3DMM

- Since the pose $\mathbf{P}$, the basis components $\hat{\mathbf{d}}_i$, the landmarks $\mathbf{l}$, and $\hat{\mathbf{m}}(\mathbf{I}_v)$ are known, we can define $\hat{\mathbf{X}} = \mathbf{l} - \mathbf{P}\hat{\mathbf{m}}(\mathbf{I}_v)$ and $\hat{\mathbf{y}}_i = \mathbf{P}\hat{\mathbf{d}}_i(\mathbf{I}_v)$

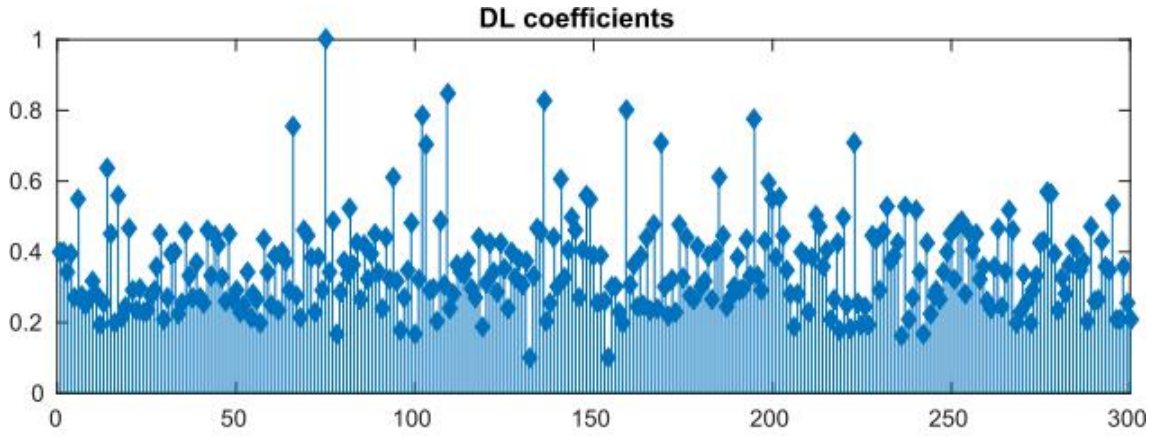
- By considering their linearized versions $\mathbf{X} \in \mathbb{R}^{98}$ and $\mathbf{y}_i \in \mathbb{R}^{98}$ with $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]$, we can estimate the non-rigid coefficients which minimize the cost of the previous regularized Ridge-Regression, in closed form

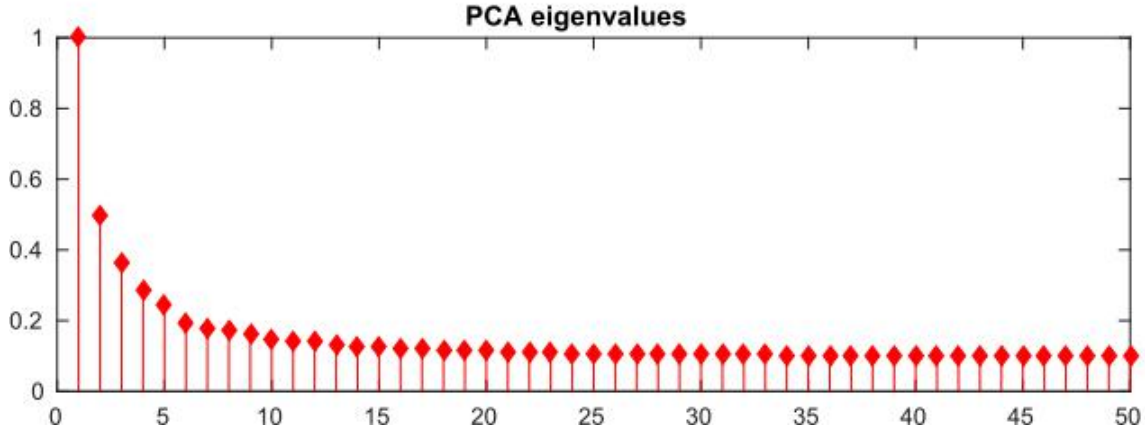
$$\boldsymbol{\alpha} = \left(\mathbf{Y}^T\mathbf{Y} + \lambda \cdot \mathrm{diag}(\hat{\mathbf{w}}^{-1})\right)^{-1}\mathbf{Y}^T\mathbf{X}$$

$\mathrm{diag}(\hat{\mathbf{w}}^{-1})$ denotes the diagonal matrix with vector $\hat{\mathbf{w}}^{-1}$ on its diagonal

- The pose estimation and fitting steps are alternated; better reconstructions are obtained by repeating the process while keeping a high regularization value

# Comparison Between DL coefficients and PCA eigenvalues




The DL-3DMM coefficients contain the energies used by the dictionary atoms to reconstruct the training signals; though all the atoms contribute to the reconstruction, the actual contribution of an atom is quantified by the related coefficient. In this sense, the weighting $\hat{\mathbf{w}}^{-1}$ privileges the more contributing atoms

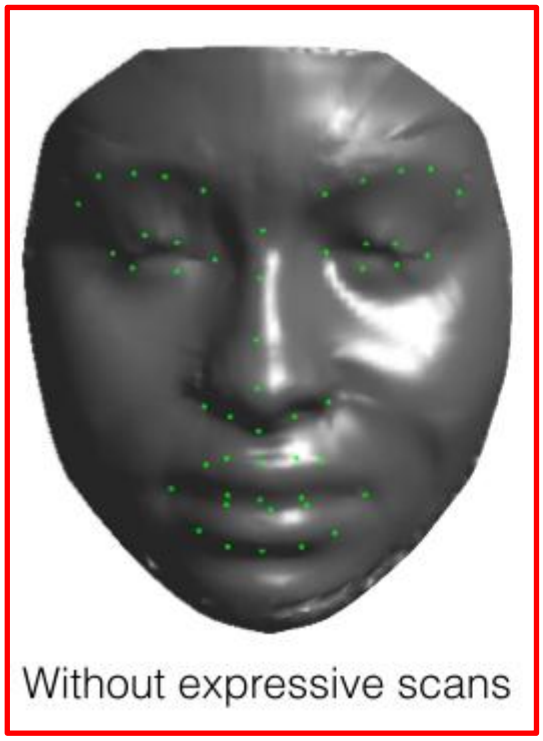
# Deformation of Single Dictionary Atoms

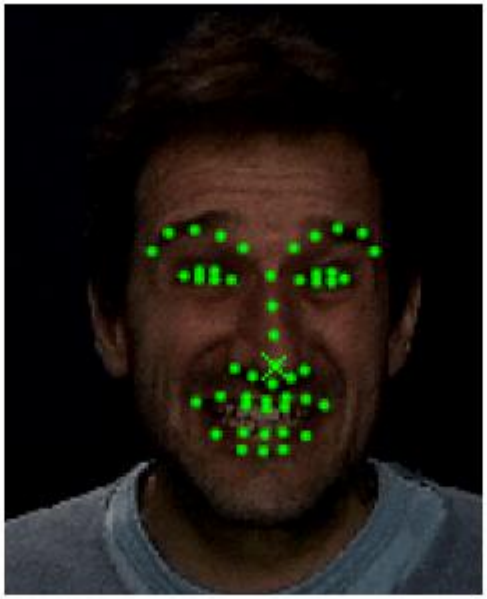

Deformation heat-maps

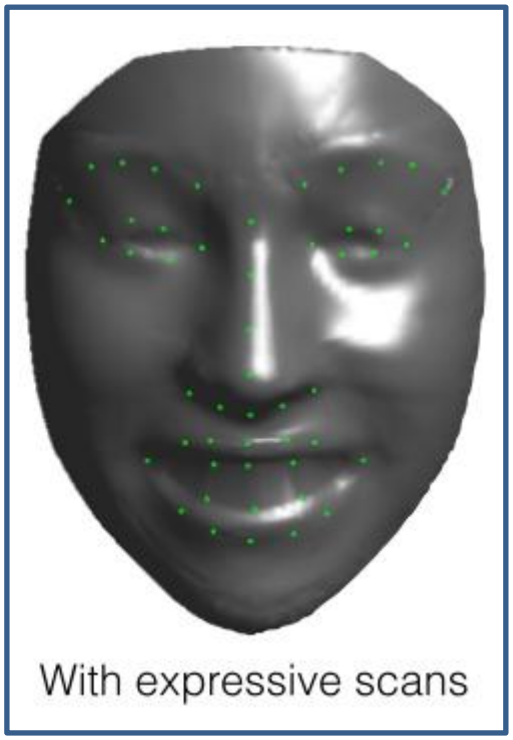
Models generated by applying different deformation magnitudes

# Importance of Expressive Training Scans

Fitting an expressive face with a 3DMM: a 3DMM built without expressive scans fails in fitting the expressive face

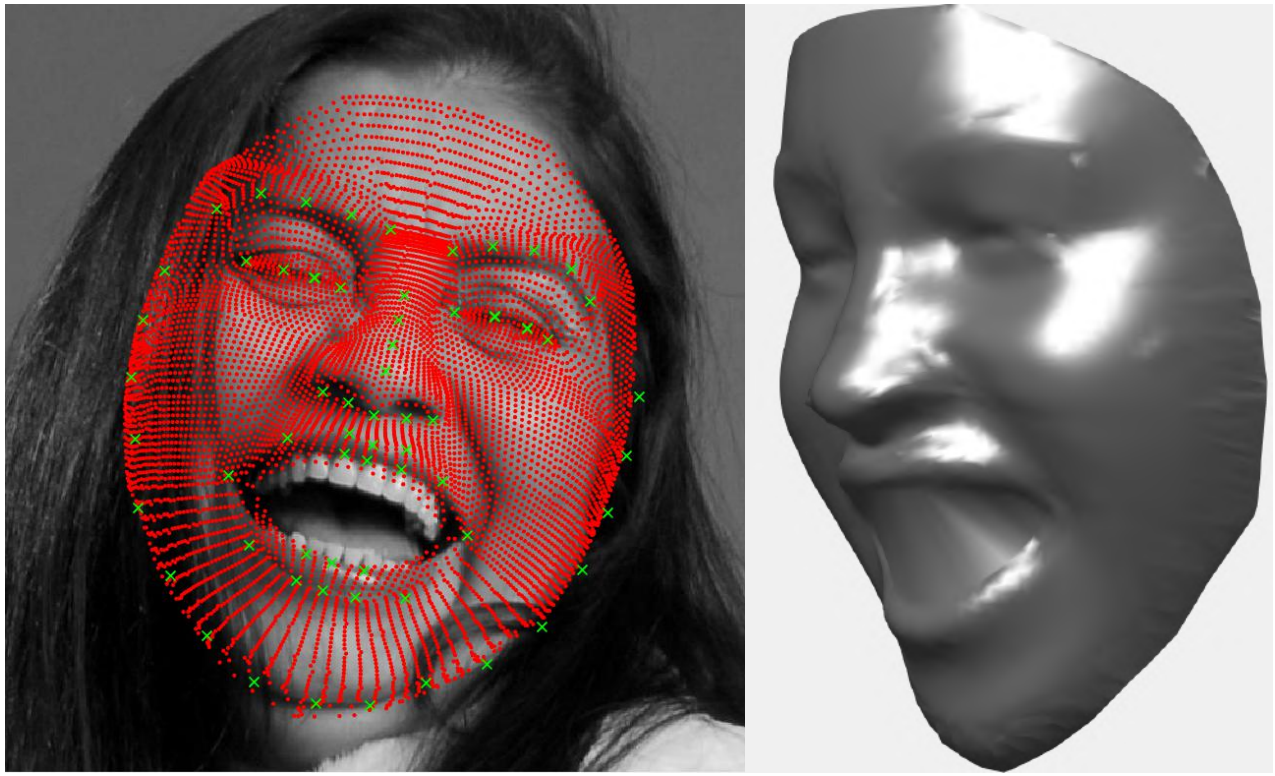


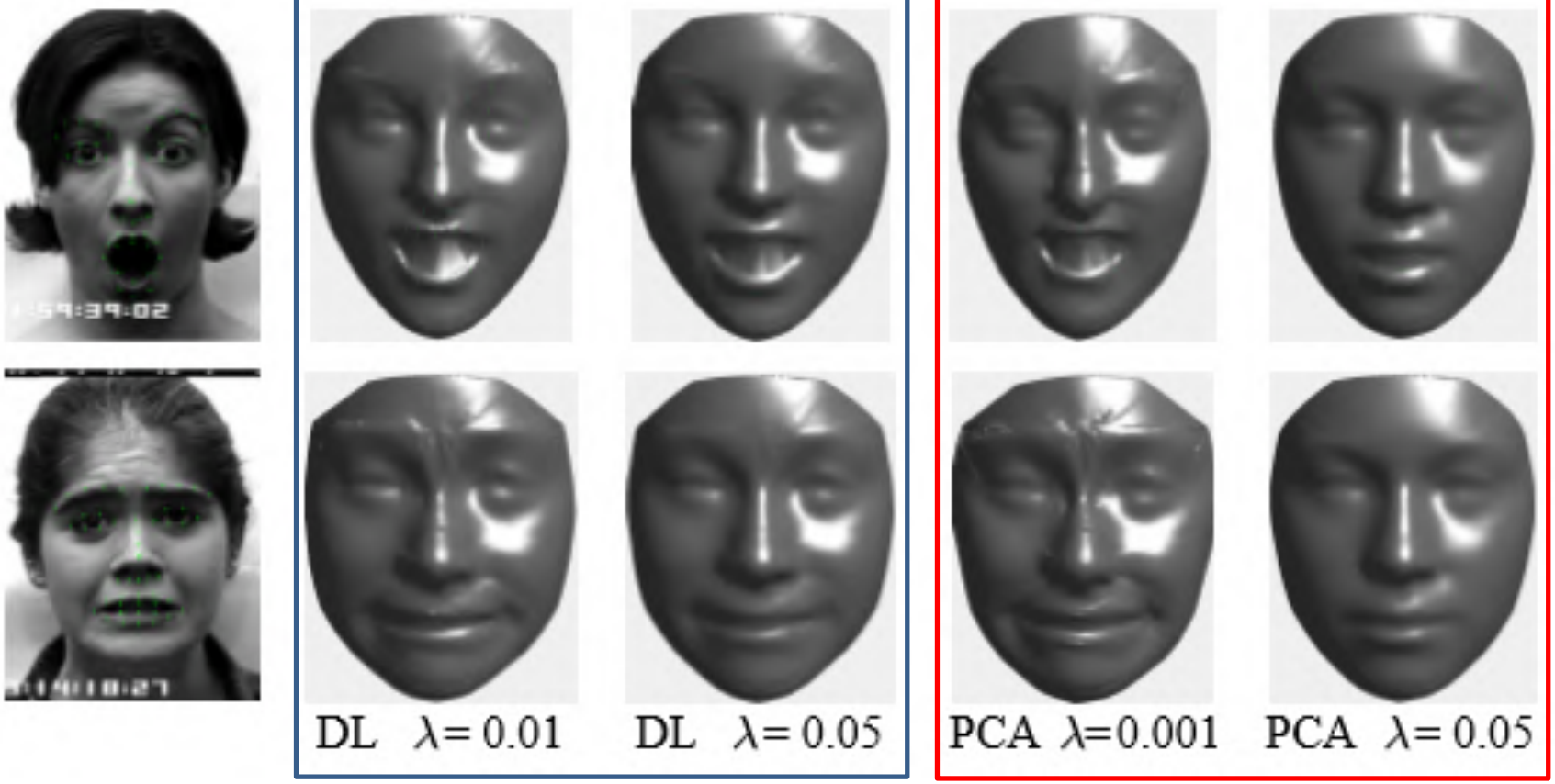


# Importance of Expressive Training Scans

Fitting of face images with strong expressions

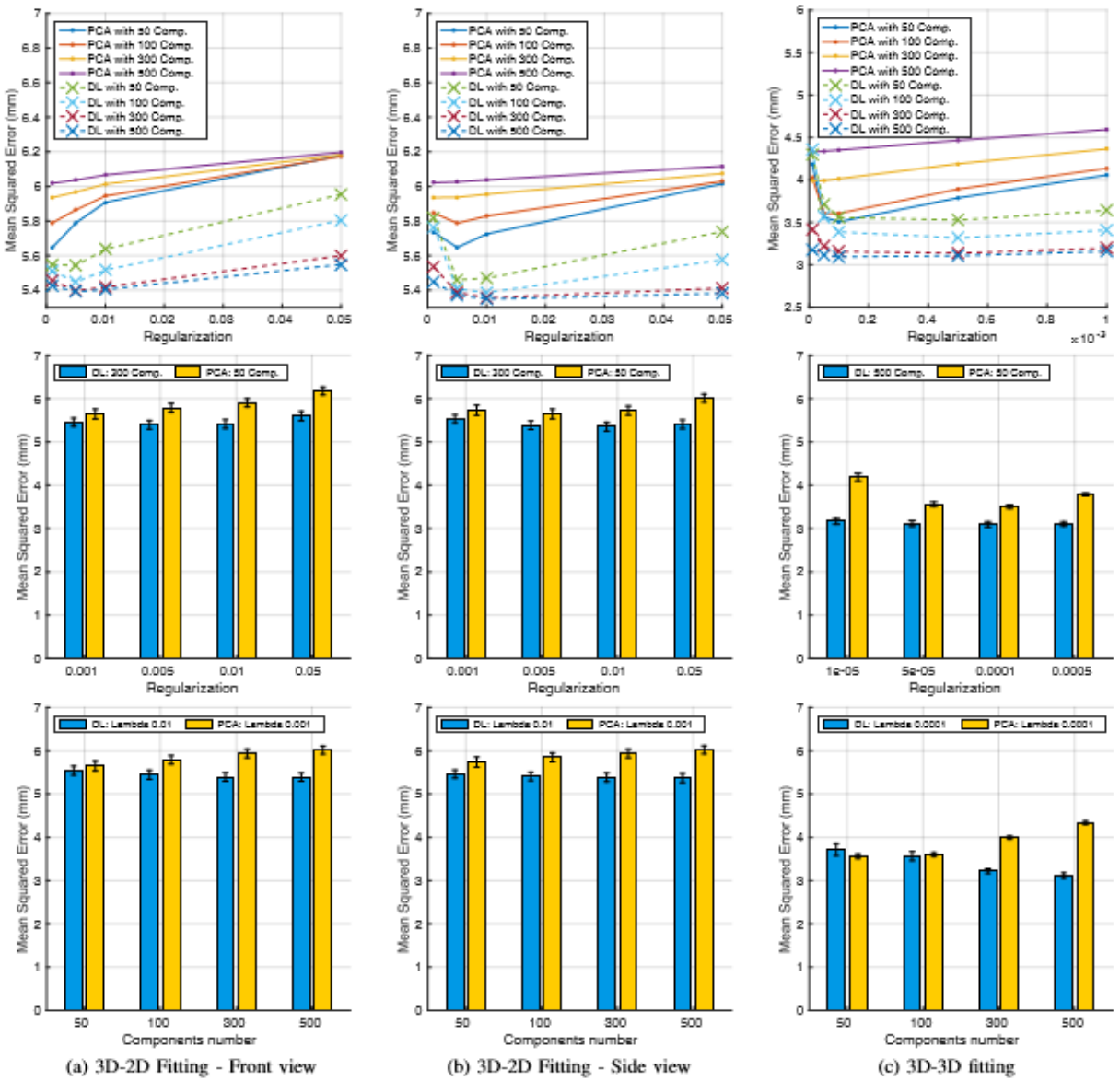


# DL-3DMM vs PCA-3DMM




DL- and PCA-based 3DMM for optimal or high regularization values: DL-3DMM both introduces less noise in the 3D models and retains its modeling ability even for high regularization values

# Reconstruction Error on BU-3DFE




(a) 3D-2D Fitting - Front view (b) 3D-2D Fitting - Side view (c) 3D-3D fitting

**First row:** errors for both DL- and PCA-based 3DMM as a function of the regularization parameter $\lambda$ and for different number of components.

**Second row:** effect of varying $\lambda$ for the best number of components.

**Third row**: effect of varying the number of components for the best value of $\lambda$

# Frontal Face Rendering

- We have now a complete 3D Morphable ( shape ) Model that is able to accurately fit expressive face images
- Rendering a canonical frontal view of the face exploits the knowledge of the 3D face shape to compute a pixel-wise inverse transformation, which associates to each pixel a 3D location in the coordinate system of the 3D model
- Once the 3D model is fit and projected onto the image, for each 3D vertex we know the coordinates $v_j = (X_j, Y_j, Z_j)$ of the pixel $(x_j, y_j)$ corresponding to the projection of the vertex on the 2D image plane
- Conversely, many pixels of the image have not a direct map in 3D, since they do not correspond to the projection of any 3D vertex

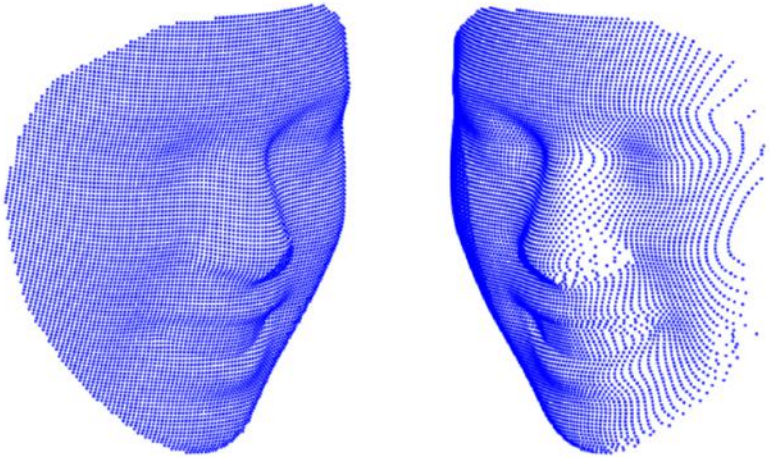
# Frontal Face Rendering

- The 3D locations of these pixels can be estimated by fitting a function $h(x,y)$ across all the scattered pixels for which the 3D to 2D mapping is known

- Defining $\Omega$ as the convex hull of the projected 3DMM, the 3D position $g_{u,v}$ of each pixel $(u, v) \in \Omega$ is estimated as

$$g_{u,v} = h(u, v) \ , \ \ \forall \ (u, v) \in \Omega$$

Original image

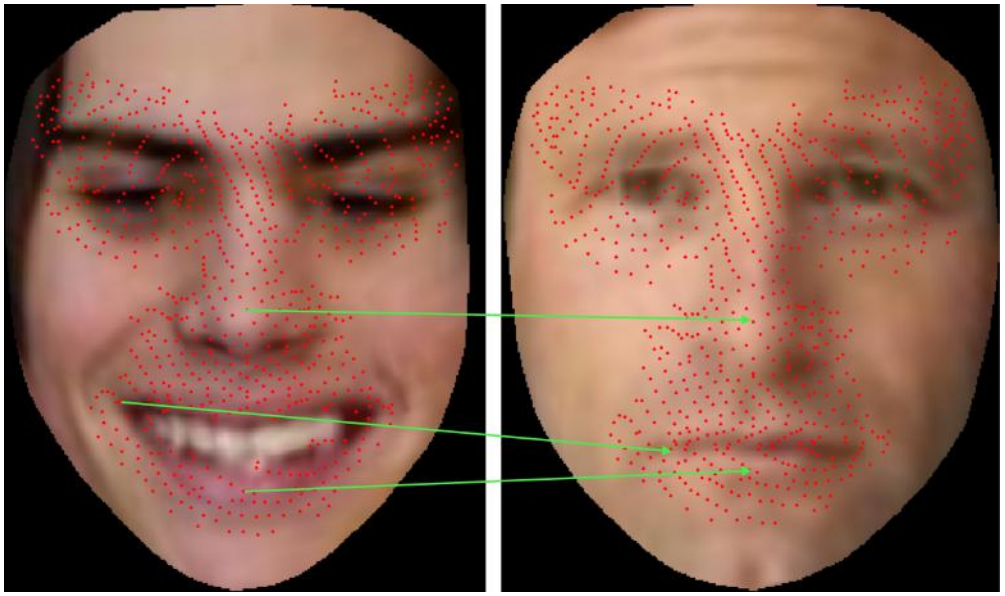


The resampled 3DMM is parametrized by the image

Frontal rendering ( artifact free )

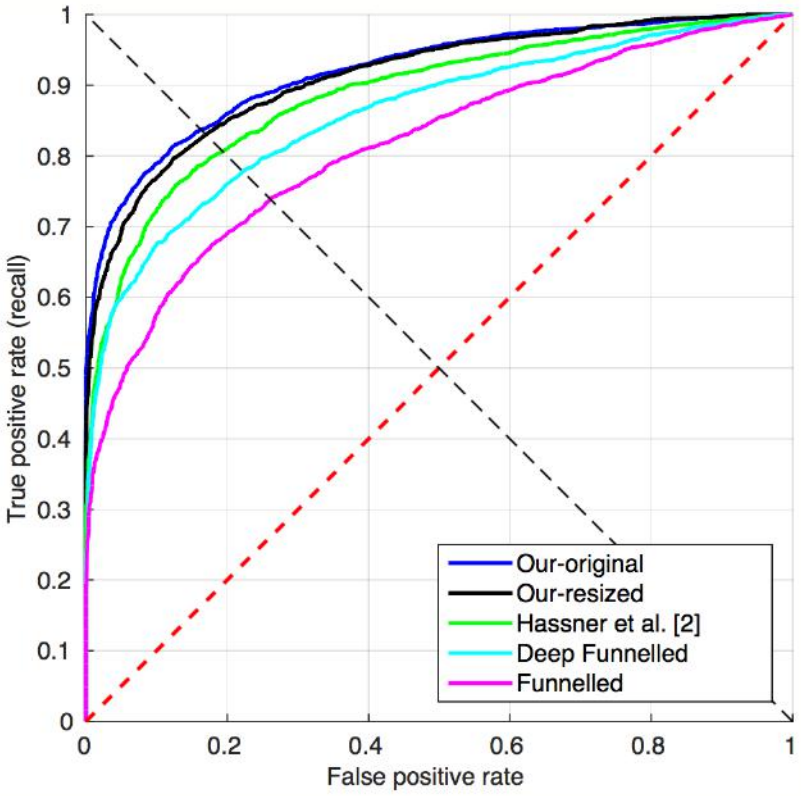
# Features Localization

- On the frontal images, the deformed vertices of the 3DMM are used to localize and match descriptors
- Match regions with the same semantic meaning e.g. mouth corner
- LBP descriptors are used as face descriptors

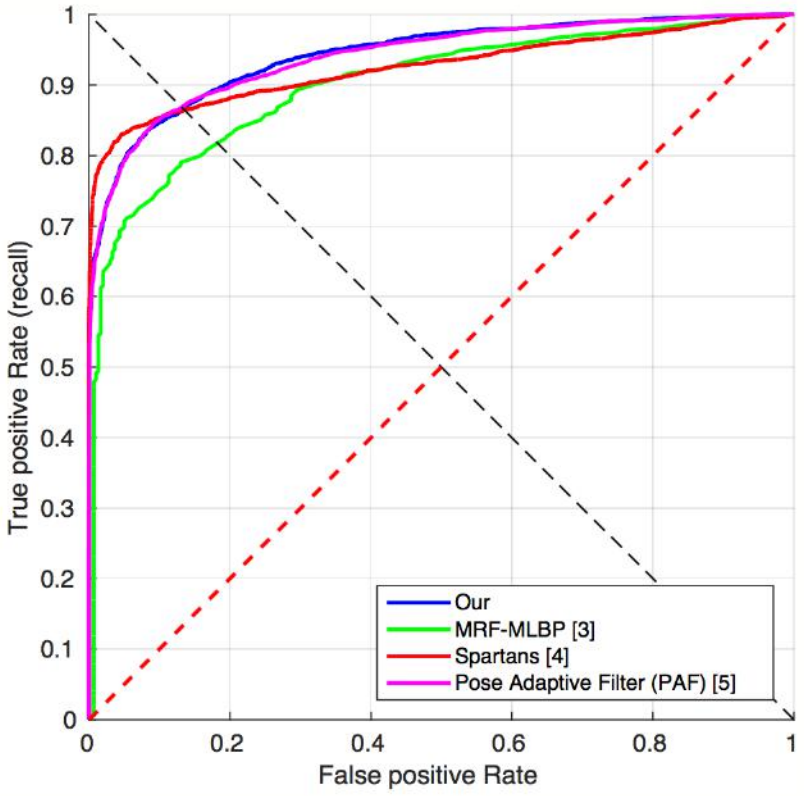


# Face Recognition on LFW

- An accurate 3DMM fitting allows good performance also for “in the wild” scenarios




Comparison with other frontalization algorithms




Comparison with the state-of-the-art

# AU Detection on FERA dataset

- Facial Action Units (AU) codify facial muscle movements based on the appearance changes e.g. eyebrow raiser, lips puller etc.

- Expressions are also formally defined as combination of action units activation

Regular grid Landmarks On the original image On the frontalized image

| AU | *F1-score* | | | | | | *AUC* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | DeGr | LM | PCA-(O) | DL-(O) | PCA-(F) | DL-(F) | DeGr | LM | PCA-(O) | DL-(O) | PCA-(F) | DL-(F) |
| 1 | 47.7 | 55.9 | 64.8 | 63.8 | 65.3 | 70.2 | 77.7 | 78.8 | 83.0 | 81.9 | 85.1 | 83.9 |
| 2 | 56.2 | 54.7 | 62.0 | 62.6 | 61.3 | 65.6 | 63.5 | 71.0 | 80.7 | 79.1 | 79.2 | 85.8 |
| 4 | 17.4 | 32.2 | 25.8 | 20.0 | 26.1 | 29.5 | 48.2 | 53.1 | 46.5 | 51.8 | 52.1 | 54.7 |
| 6 | 55.5 | 52.6 | 60.8 | 57.0 | 66.7 | 66.3 | 73.0 | 77.3 | 76.3 | 72.9 | 81.0 | 80.0 |
| 7 | 48.3 | 55.8 | 45.5 | 47.7 | 52.9 | 52.0 | 71.1 | 66.8 | 57.5 | 57.0 | 62.1 | 64.9 |
| 12 | 39.2 | 55.1 | 55.2 | 55.9 | 58.0 | 59.3 | 66.5 | 62.9 | 64.8 | 66.5 | 63.9 | 64.9 |
| 15 | 68.5 | 65.0 | 77.2 | 77.1 | 79.7 | 80.4 | 73.8 | 81.5 | 84.6 | 82.7 | 85.8 | 87.5 |
| 17 | 26.4 | 25.8 | 36.9 | 42.6 | 31.1 | 33.1 | 60.5 | 66.9 | 65.3 | 69.9 | 58.8 | 61.7 |
| Avg. | 44.9 | 49.6 | 53.5 | 53.4 | 55.1 | **57.1** | 66.8 | 69.8 | 69.8 | 70.2 | 71.0 | **72.9** |

DL better captures local deformations

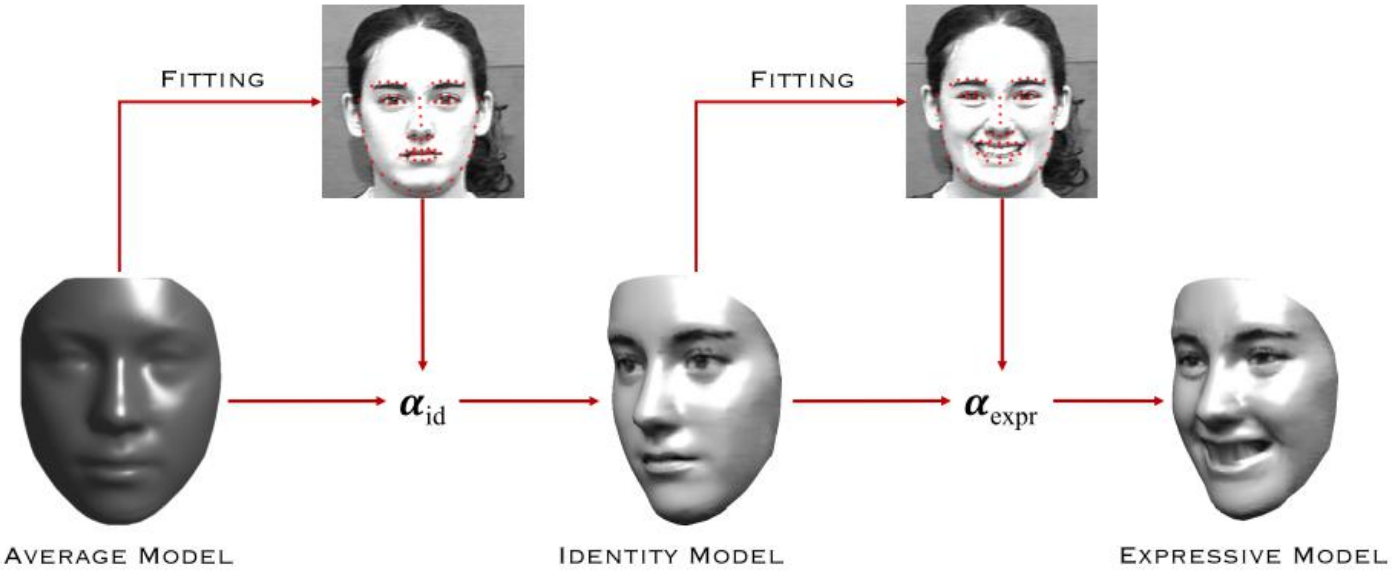
# Emotion Recognition on FERA dataset

- Similar results are obtained for the emotion recognition task
- The gap between PCA and DL is smaller since emotions usually involve complex movements of the whole face
- More complex task since emotions are subjective and different subject might show different expressions

| Emotion | DeGr | LM | PCA-(O) | DL-(O) | PCA-(F) | DL-(F) |
|---|---|---|---|---|---|---|
| Anger | 56.4 | 66.7 | 64.4 | 63.0 | 67.7 | 70.5 |
| Fear | 85.8 | 73.7 | 77.4 | 73.0 | 81.9 | 88.4 |
| Joy | 93.0 | 91.4 | 90.9 | 91.9 | 92.1 | 91.5 |
| Relief | 80.2 | 76.4 | 77.4 | 75.6 | 79.5 | 79.0 |
| Sadness | 81.1 | 78.0 | 81.0 | 80.7 | 86.2 | 81.5 |
| Avg. | 79.3 | 77.2 | 78.2 | 76.8 | 81.5 | **82.2** |

# Expression Transfer

- Another application of our DL-3DMM is the expression transfer
- The idea is that we can fit the 3DMM to neutral faces and use the subject-specific model to fit an expressive face of the same subject to learn expression-specific deformation parameters
- Given a set of training subjects, we can learn expression-specific deformation to be applied to a generic face image ( in neutral expression )




Ferrari, et al., *Rendering Realistic Subject-Dependent Expression Images by Learning 3DMM Deformation Coefficients*, *ECCV* workshops 2018

# Expression Transfer

- The expression-specific parameters are learned by means of simple statistical indicators ( mean, median … )
- This suggests that the dictionary is effective in separating between identity and expression components, even without explicit divisions of the training scans
- Some examples with applied expression are shown

# Practical Session

- Now that we have revised in detail the whole construction process, we can practice!