The problem of estimating the 3D shape of human faces from single images is of great interest and has attracted considerable research effort. Many approaches recently proposed to solve this problem could be considered extensions of Shape-from-Shading (SFS) methods, where a 3D shape is optimized to generate 2D renderings that match the input images [1, 5, 7]. Other methods in the literature propose to infer 3D face shape by fitting a set of feature points between the 2D image and the 3D model [3, 4, 6].

In this paper, we propose the Two-Fold Coupled Structure Learning (2FCSL) algorithm, which is capable of reconstructing 3D face models based on a sparse set of 2D landmarks that could be localized automatically by most of the recently proposed landmark detectors. By explicitly incorporating 3D-2D pose estimation and formulating the problem into a two-fold coupled structure learning problem, our method achieves better robustness to arbitrary pose variations and landmark localization noise.

Using a shape vector representation $\chi_{3D}$ of the dense 3D face, $N$ 3D training faces are stacked together to construct the 3D dense landmark (3DDL) model $\chi_{3D} = (\chi_{3D}^1, \ldots, \chi_{3D}^N)$. Similarly, 3D sparse landmark (3DSL) model is represented by $\chi_{3D}(x_{3D}) = (x_{3D}^1, \ldots, x_{3D}^N)$, where $x_{3D}$ is the vector representation of $M$ 3D landmarks. Given a 2D image, a sparse set of landmarks $x_{3D}$ is first detected with any off-the-shelf detector. Then, the 3D-2D projection matrix $P$ is estimated using least squares minimization, such that $x_{3D}^i = P x_{3D}^i$, where $x_{3D}$ is the mean of 3DSLs in the training database. By projecting each 3DSL via $P$, the corresponding 2D sparse landmark (2DSL) model $x_{2D}^i = (x_{2D}^i, \ldots, x_{2D}^N)$, where $x_{2D}$ is the vector representation of $M$ 2D landmarks, is generated on-line.

By applying PCA to the 3DSL and the 2DSL models, we derive a compact representations of the corresponding shapes $A_m$ and $A_n$, based on which a PLS regression $P_{PLS}$ [2] is learned, $A_m = A_n P_{PLS}$:

$$x_{3D} = \bar{x}_{3D} + \sum_{n=1}^{N-1} a_n u_{3D}^{i}$$

$$a_m = [a_{m1}^2, \ldots, a_{m(N-1)}^2]$$

In the training phase, the correlation between 3DSL and 3D model is implicitly learned in a coupled manner.

$$\arg\min_{\alpha, \lambda_{3D}, \lambda_{4D}} \left\| \begin{bmatrix} \beta_0 A_{3D}^d \\ \alpha^T \end{bmatrix} \right\|_2^2 s.t. \|\alpha\|_1 \leq \beta_1$$

$$\arg\min_{\alpha^*} \left\| \frac{X_{3D}^d - A_{3D}^d \alpha^*}{\beta_0} \right\|_2^2 + \beta_2 \|\alpha^*\|_1 + \beta_1 \|\alpha^*\|_2$$

By fitting $X_{3D}^d$ to $A_{3D}^d$, the shared coefficient $\alpha^*$ could be recovered by solving Eq. 4. Then, the final $y_{3D}^R$ is reconstructed via Eq. 5:

$$y_{3D}^R = \frac{A_{3D}^d \alpha^*}{\beta_0}$$

In the paper, we conducted several experiments using both synthetic data and real 2D face images from two face datasets. Compared with [6], our method demonstrates higher reconstruction accuracy and better robustness to face pose variations and landmark localization noise. Fig. 1 depicts the reconstructed 3D face of Mona Lisa using the famous painting by Leonardo da Vinci and the lifted texture in a pre-registered UV space.

Figure 1: (a) The input 2D image; (b) frontal view of the face shape reconstruction; (c) profile view of the face shape reconstruction; and (d) lifted UV texture.