Computer Vision Research with New Imaging Technology

Guangqi Hou*[a], Fei Liu[a,b], Zhenan Sun[a]

[a] Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences; [b] College of Engineering and Information Technology, University of Chinese Academy of Sciences

ABSTRACT

Light field imaging is capable of capturing dense multi-view 2D images in one snapshot, which record both intensity values and directions of rays simultaneously. As an emerging 3D device, the light field camera has been widely used in digital refocusing, depth estimation, stereoscopic display, etc. Traditional multi-view stereo (MVS) methods only perform well on strongly texture surfaces, but the depth map contains numerous holes and large ambiguities on textureless or low-textured regions. In this paper, we exploit the light field imaging technology on 3D face modeling in computer vision. Based on a 3D morphable model, we estimate the pose parameters from facial feature points. Then the depth map is estimated through the epipolar plane images (EPIs) method. At last, the high quality 3D face model is exactly recovered via the fusing strategy. We evaluate the effectiveness and robustness on face images captured by a light field camera with different poses.

Keywords: Light field imaging, 3D face model, depth estimation, EPIs

1. INTRODUCTION

The concept of light field was originally used in computer graphics as a powerful tool to record light rays from different directions. And recently it attracts more attention from computer vision community[1]. In computer graphics and computer vision researches, passive imaging methods are generally used based on the texture and shading cues of 2D images. In real world scenes, many surface regions are low-textured, even textureless, such as face, wall, table, and sky. However, current depth estimation algorithms for light field cannot obtain high quality reconstruction results. More researches must be carried out especially for dealing with poor texture regions as well as structure lighting imaging devices (e.g. Kinect).

Through computational photography methods on 4D light fields, new imaging properties can be derived by digital refocusing, high-dynamics range (HDR) and depth estimation. Light field imaging is recently applied for biometric application. Biometric technologies obtain great accuracy in recognizing person's identity, such as fingerprint, iris and face, among which face recognition is the most natural and popular biometric method. Furthermore, the intensity values and directions of rays can be captured simultaneously by the light field imaging. The additional rays information with more dimensions can be exploited for conventional face imaging process. Many researches have been done by making use of the light field imaging for biometric recognition, such as iris recognition[2] and face recognition[3,4]. The depth information is one important element for 3D scene display and understanding, which can be extracted from light field raw images. 3D face imaging *in the wild* can be applied in some popular applications, including automatic face recognition in surveillance video, network video meeting, entertainment and online 3D gaming.

In this paper, we import prior knowledge to enhance the light field depth estimation results. Two main differences on capturing setups exist between our work and Ralph[3]: (1) light field raw images are captured by a light field camera with a universal lens, rather than by camera arrays. The camera arrays usually cause large disparity, which lead to many tricky problems, such as non-ideal camera calibration, low image resolution, etc.; (2) we reconstruct high quality 3D face models based on light field depth estimation results, not from original 2D images.

The remainder of this paper is organized as follows. In the following section, we introduce the related works. The details of the proposed framework are described step by step in Section 3. More experimental results and discussions are shown in Section 4. Section 5 concludes our work.

*corresponding author: Guangqi Hou; email gqh@nlpr.ia.ac.cn;

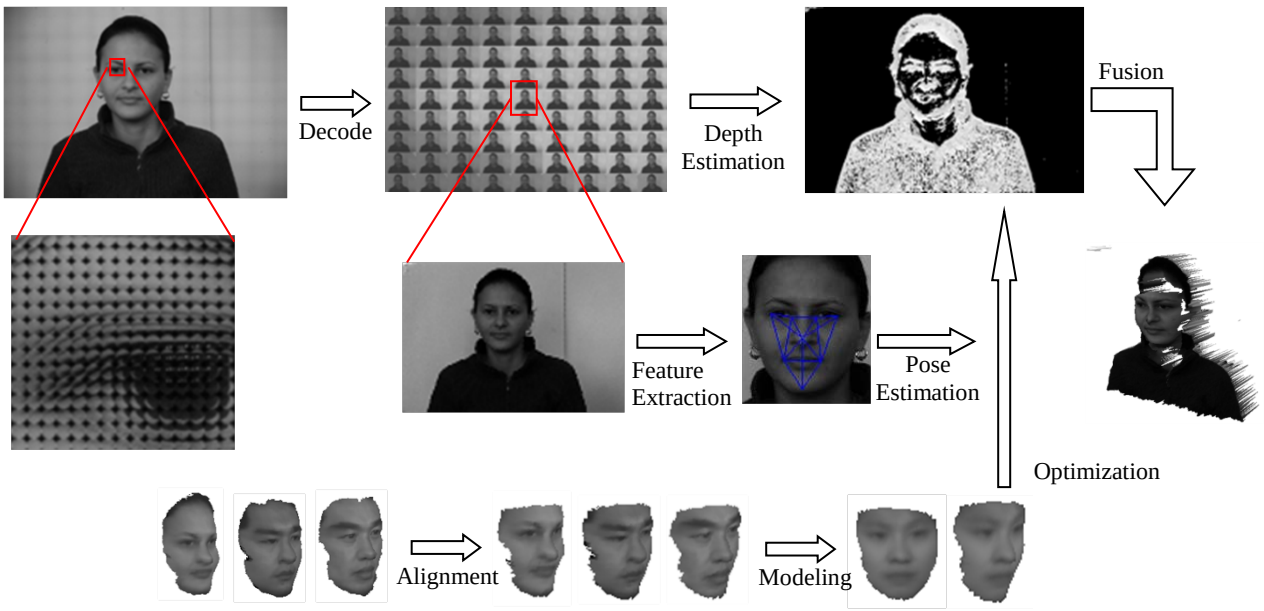

Figure 1: The whole flowchart of our method.

2. RELATED WORK

In the recent, light field cameras draw more attention due to its refocusing character at one snapshot moment. We build a 3D face morphable model based on the prior knowledge to enhance the depth estimation results from light field raw images. The following section will introduce the related research works of light field methods and facial morphable models.

2.1 Light field methods

Light field is prompted by Adelson and John[5] in 1992, and it can be parameterized as 4D light fields to decrease computation complexity by Levoy and Hanrahan[6]. Generation of epipolar plane images (EPIs) is adopted to represent 4D light fields from raw data captured by one light field camera with a single lens[7], and depth maps for the scenes are estimated from EPIs[8]. Ren proposes digital refocus concept and 4D Fourier Slice method[9] to accelerate the performance of refocus algorithms, and the hand-held light field camera is designed with enabling dynamic depth of field (DoF) extension[10]. The images rendered from light field cameras suffer from loss in resolution, and small micro-lens apertures lead to image blurring. To overcome these disadvantages, Bishop and Favaro[11] use an image formation model, and incorporate Lambertian model and texture into depth map estimation.

2.2 Facial morphable model

To achieve high-fidelity 3D models, it is necessary to utilize the prior knowledge of a statistical 3D face model. There are two well-known techniques: Active Appearance Models (AAMs) and 3D morphable Models (3DMMs). Although large rotation angles cannot be generated by AAMs[12] from traditional 2D images, 3DMMs are capable of overcoming the rotation transformation difficulty. 3DMMs by Blanz[13, 14] is built from the 3D face database, and is generated by performing Principal Component Analysis (PCA) on shape and texture vectors. The maximum a posteriori estimator (MAP) is implemented to minimize the energy between the input image and the rendering image. The disadvantages for 3DMMs is high computation complexity and sensitive to the original average face model. Zhou[15] proposes one morphing system, which builds a statistical model based on morphing in face recognition, and then fitting 3DMMs to 2D face images.

3. LIGHT FIELD FACE IMAGING FRAMEWORK

In the proposed light field face imaging framework, the prior knowledge is applied to build one generic 3D facial model with Principal Component Analysis (PCA) based on the 3D face database. Firstly, the light field raw image is decoded and the sub-aperture images are generated through interpolation and resampling from the toolbox[16] with our light field camera parameters. The central view sub-aperture image is used to accurately extract the facial feature points, and then the rotation and translation matrices are calculated with SoftPOSIT algorithm[17]. We treat the rotation and translation parameters as the facial pose, and minimize the cost energy function by the maximum a posteriori estimator (MAP). Then the new 3D face model is generated. Meanwhile, the original depth map is estimated from EPIs[8], which is with large holes and noisy for textureless and low-textured surface regions. By aligning the estimated depth map and the 3D face model, the smooth and reliable 3D face model is reconstructed by the fusion strategy with the central view sub-aperture image. Figure 1 is the whole flowchart of our proposed algorithm.

3.1 3D facial generic model

We use the popular RGB-D camera—Kinect, to capture ten 3D face images for people standing at 1.2m from the device. Then light field raw images are obtained by a light field camera with micro-lens array structure. The model texture is represented as

$$I(x,y,z)=\left(R(x,y,z),G(x,y,z),B(x,y,z)\right)^T \tag{1}$$

The world coordinate system (x,y,z) is used instead of the cylindrical coordinate system[18]. We apply the grid calculation method to regulate all 3D face images. Then all faces are aligned with the Iterative Closest Point (ICP) algorithm after four basic vertices are selected manually, and the shape vector is defined as

$$S_i=\left(x_1,y_1,z_1,x_2,y_2,...,x_n,y_n,z_n\right)^T \tag{2}$$

The intensity images obtained by Kinect are stored in three color channels (R,G,B), and the light field camera can only obtain gray-scale images. In the proposed algorithm, the texture vector is represented as

$$T_i=\left(G_1,G_2,...,G_n\right)^T \tag{3}$$

We perform PCA for shape and texture vectors S_i and T_i separately. The eigenvectors $\{s_1,s_2,...,s_m\}$ and variances $\{\sigma_{S,1},\sigma_{S,2},...,\sigma_{S,m}\}$ are obtained. Then the new 3D morphable model can be calculated by

$$S=\overline{S}+\sum_{i=1}^{m-1}\alpha_i s_i,\quad T=\overline{T}+\sum_{i=1}^{m-1}\beta_i t_i \tag{4}$$

in which the model parameters α_i and β_i are deferent from each face sample. $\overline{S}$ and $\overline{T}$ are the average shape vector and texture vector separately.

The head pose transformation is treated as a rigid process, and the new coordinates for the vertex x_k is calculated by

$$\left(\omega_{x,k},\omega_{y,k},\omega_{z,k}\right)^T=s\cdot\Re_\gamma\Re_\theta\Re_\phi x_k+t_w \tag{5}$$

in which γ,θ,φ are the yaw, pitch and roll angles; s is the scale factor; t_w is the translation value.

The image $I_{model}(x,y)$ rendered from the new 3D face model can be obtained from a perspective projection. In this paper, we don't consider the illumination into our model. The process of solving α_i and β_i is one optimization problem. The energy function between the input image and the image rendered from the 3D model is described as

$$E_I=\sum_{x,y}\left\|I_{input}(x,y)-I_{model}(x,y)\right\|^2 \tag{6}$$

In order to achieve our goal, we use MAP method to minimize E. The whole cost function is as following

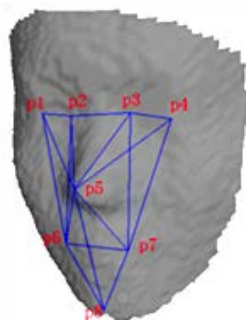

Figure 2: Description of eight facial feature vertexes.

Figure 3: Facial feature points for five poses.

$$E = \frac{1}{\sigma_N^2} E_I + \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{(\rho_i - \overline{\rho}_i)^2}{\sigma_{R,i}^2} \tag{7}$$

in which σ_N is the standard deviation for all images containing Gaussian noise. ρ_i is the transformation vector, which is set by the facial feature algorithm, not by hand[13]. $\gamma, \theta, \varphi, \overline{\rho}_i$ are set as their initial values.

During the whole iterated process, the parameters α_i, β_i, ρ_i need to be estimated and updated with the gradient descend method. We design different weight values λ as

$$\alpha_i \to \alpha_i - \lambda_\alpha \frac{\partial E}{\alpha_i},\ \beta_i \to \beta_i - \lambda_\beta \frac{\partial E}{\beta_i},\ \rho_i \to \rho_i - \lambda_\rho \frac{\partial E}{\rho_i} \tag{8}$$

3.2 Facial feature extraction and matching

The parameters for the 3D morphable model must be initialized automatically in light field face imaging. We consider that γ, θ, φ can be combined into the rotation matrix R. So, the pose transformation from the 3D model to the input image is the former step for obtaining the new 3D face model.

We design eight facial feature vertexes to calculate the pose transformation parameters. The eight vertexes in Figure 2 are: four eye corners $p1, p2, p3$ and $p4$ (right eye outer, right eye inner, left eye inner, left eye outer); the nose tip $p5$; left mouth corner and right mouth corner $p6, p7$; the chin tip $p8$.

The light field raw image can be decoded into 81 sub-aperture images. We select the central view sub-aperture image, which has little optical skewness (such as lens distortion and vignetting), as the input for facial feature algorithm.

We make use of the facial feature extraction algorithm from Yu et al.[19] for obtaining accurate facial feature points. These points are called as landmarks. This algorithm[19] implements a two-stage cascaded deformable shape fitting method to localize facial landmarks, and introduces the 3D shape model with optimized mixtures of parts.

While extracting facial landmarks, our feature points can also be obtained through indexing the 3D shape model. Then the facial pose is calculated by SoftPOSIT algorithm, which is an improved method from POSIT (Pose from Orthography and Scaling with ITerations).

The matching between the 3D face model and input image is projection from 2D image space to 3D world coordinate system. The transformation from the 3D shape model to the 2D image is perspective projection. Therefore, the relationship between them obeys the optical imaging principal. When generating the 2D image, its image plane will correspond to the input sub-aperture image.

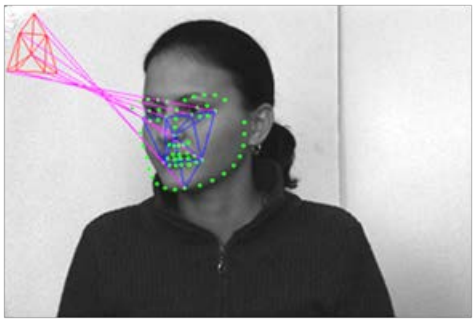

Figure 4: Matching the 3D model with the 2D input sub-image.

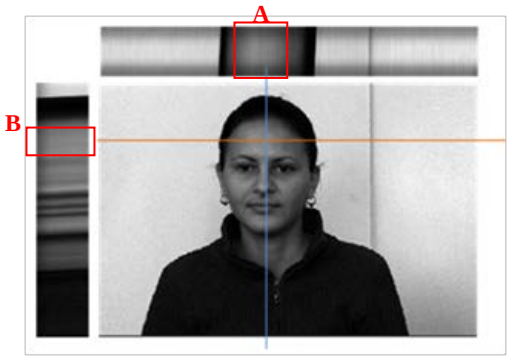

Figure 5: The epipolar plane images (EPIs).

3.3 Depth estimation from EPIs

The light fields captured by a light field camera can be parameterized as a 4D function $L(s,t,x,y)$, where (s,t) represent different viewpoints on the micro-lens plane and (x,y) represent spatial coordinates on the sensor plane. The epipolar plane images (EPIs) can be viewed as 2D slices of the light fields. In Figure 5, the top EPI is one of the 2D slices by fixing t and y, and the left EPI is by fixing s and x.

Active 3D imaging technologies, such as Kinect, Time-of-Flight and Laser Scanners, measure depth values of points on the object by emitting extra lighting. In general, the reliability, accuracy, and time-consuming of passive stereo vision cannot come up to active imaging. Light field imaging is an emerging passive 3D imaging method. Each 3D point is projected onto different views, which form one line in the EPI. The slope of each line is related to the depth value of this 3D point. Therefore, we can estimate the depth map by analysing EPIs. However, depth estimation for low-textured regions is still a challenging task, e.g. region A and B in Figure 5. There are numerous holes and ambiguities for these regions, which are so sensitive and noisy that the light field camera cannot be used for other applications normally. Figure 6 is the depth estimation and 3D reconstruction results based on the EPIs algorithm[8].

In this paper, we make use of depth estimation results from EPIs as the intermediate inputs, and perform the depth enhancement through a fusion method with the 3D morphable model.

3.4 Depth enhancement by the fusion method

Till now, we already obtain the original depth maps and 3D morphable model from EPIs and MAP respectively. Then we seek to fuse them for reconstructing more smooth and stable 3D face models.

Firstly, we align the *x-axis* and *y-axis* of the 3D morphable model to the coordination grid of light field sub-aperture images. The depth estimation results from EPIs are enhanced by median filtering for filling small holes. And depth values of eight facial feature points will be extracted. We adopt a surrounding 7×7 patch to calculate the depth value of each feature point. The unreliable depth values are removed by setting a minimum threshold. Finally, the central sub-aperture image is used as texture for the reconstructed 3D face after the new morphable model is transformed by rotation and translation matrix (see Figure 7).

Figure 6: The depth estimation result from EPIs

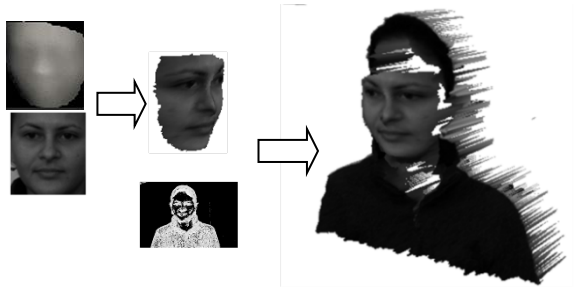

Figure 7: Fusing the original depth map and 3D morphable model.

4. EXPERIMENTS AND DISCUSSION

4.1 Capture 3D facial database and light field raw images

We capture 3D face database through Kinect and light field raw images by a light field camera with a universal camera lens at the same time. In our method, we do not perform camera calibration, and the image number of our Kinect database is much smaller than popular 3D face databases (e.g. FRGC), which means that weak prior knowledge is used.

4.2 Experiments

In order to evaluate the performance of our method on real world scenes, we capture 11 persons in 5 face poses by our designed light field camera. There are no constraints on the environment illumination in our capturing setup. Hence, the light field raw images contain a lot of noises due to the continuous variation of ambient lighting. The lens assembled on the camera is one generic Nikon AF 24-85mm product, which is locked at 85mm focal length. The resolution of our image sensor is 11Megapixel, and the micro-lens number is 399×266.

Figure 8 shows some results from the total 55 light field raw images. In the figure, the top row images are the central sub-aperture images exacted from light field raw images; the middle row images are depth estimation results from EPIs; the bottom row images are 3D face models reconstructed by our proposed algorithm. From these results, it is easy to see that the reconstructed 3D face models are relatively smoother and more reliable compared to the original depth maps, but the body parts are strongly noisy.

In addition, acceptable results in the database are 40 samples, and the other 15 results have large distortion in specific depth coordination and extremely face poses. The next section will give a detail discussion on the reason of the 15 poorly reconstructed results.

4.3 Discussion

In light field imaging, it is very hard to execute reliable depth map estimation on textureless or low-textured surface regions of the objects. We propose a 3D face imaging method with weak prior knowledge from light field raw images to solve this problem. The details of these 3D face images are coarse, and a large database with more samples is needed to obtain high quality details on the surface.

Seventy-two percent of our experiment results are acceptable for computer vision applications with smooth surfaces and low noises, and the remainder samples are reconstructed in large surface and structure deviations. There are two main reasons:

Figure 8: The experimental results.

(1) Facial feature extraction. In our experiments, the effectiveness and accuracy of facial feature points are very important for the fusion step. There is one bad result showed in Figure 9. The left image shows the extracted facial feature points, where the vertical coordination of the chin tip point is located with large deviation. The right image is its corresponding 3D face model. The deviation of only one feature point leads to the terrible result.

(2) 3D surface combination. The final reconstructed 3D face models are fused by the original depth estimation results from EPIs and the 3D morphable model. The depth direction, which can be defined as the *z-axis in its 3D coordinates*, must be accurately matched. And the better interpolating method of 3D curved surfaces is another key factor, which is essential for the reconstruction of reliable face boundaries in the final 3D face model.

5. CONCLUSION

The goal of this paper is to exploit the light field imaging technology on a classical computer vision problem--3D face modeling. The face images are captured *in the wild* by our designed light field camera with a universe lens. We propose a light field face imaging framework based on the 3D morphable model. The face pose parameters are estimated from the extracted facial feature points. By Fusing the estimated depth maps based on EPIs and the 3D morphable model, the final 3D face models are reconstructed with the MAP optimization method.

In our experiments, we deal with light field raw images from different people with five poses. The smooth and reliable 3D face results of most samples confirm the effective and accuracy of our proposed framework. Furthermore, some bad reconstruction results show that we need to pay more attention on robust facial feature extraction and accurate 3D curved surface interpolating algorithms in the future research.

Figure 9: One failing fusion sample due to large deviation of one facial feature point

ACKNOWLEDGMENT:

This work is funded by National Natural Science Foundation of China Youth Fond (Grant No.61302184), National Natural Science Foundation of China Youth Fond (Grant No.61305007) and National Natural Science Foundation of China Major Instrument Special Fund (Grant No. 61427811).

REFERENCES

[1] SvenWanner, S. Meister, and B. Goldluecke, "Datasets and Benchmarks for Densely Sampled 4D Light Fields," in Vision, Modelling and Visualization (VMV), 2013, 2013.
[2] C. Zhang, G. Hou, Z. Sun *et al.*, "Light Field Photography for Iris Image Acquisition," in Chinese Conference on Biometric Recognition(CCBR), 2013.
[3] R. Gross, I. Matthews, and S. Baker, "Appearance-Based Face Recognition and Light-Fields," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, 2004.
[4] R. Raghavendra, B. Yang, K. B. Raja *et al.*, "A New Perspective - Face Recognition with Light-Field Camera," in Biometrics (ICB), 2013 International Conference on, 2013.
[5] E. H. Adelson, and J. Y. A. Wang, "Single Lens Stereo with a Plenoptic Camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, 1992.
[6] M. Levoy, and P. Hanrahan, "Light Field Rendering," in SIGGRAPH '96 Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996.
[7] S. Wanner, J. Fehr, and B. Jahne, "Generating EPI Representations of 4D Light Fields with a Single Lens Focused Plenoptic Camera," in International Symposium on Visual Computing (ISVC) 2011, 2011.
[8] S. Wanner, and B. Goldluecke, "Globally Consistent Depth Labeling of 4D Lightfields," in CVPR'12, 2012.
[9] R. Ng, "Fourier Slice Photography," *ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2005*, 2005.
[10] R. Ng, M. Levoy, M. Bredif *et al.*, *Light Field Photography with a Hand-held Plenoptic Camera*, 2005.
[11] T. E. Bishop, and P. Favaro, "The Light Field Camera: Extended Depth of Field, Aliasing, and Superresolution," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2012.
[12] G. J. E. Timothy F. Cootes, Christopher J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681-685, 2001.
[13] V. Blanz, S. Romdhani, and T. Vetter, "Face Identification across Different Poses and Illumination with a 3D Morphable Model," in Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02), 2002.
[14] V. Blanz, and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Transactions on Pattern Analysis and Machine Intellegence*, 2003.
[15] X. Zou, J. Kittler, and J. Tena1, "A Morphing System for Effective Human Face Recognition," in Visual Information Engineering, 2008. VIE 2008. 5th International Conference on, 2008.
[16] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on 2013.
[17] P. David, D. DeMenthon, R. Duraiswami *et al.*, "SoftPOSIT: Simultaneous Pose and Correspondence Determination," in 7th European Conference on Computer Vision Copenhagen, Copenhagen, Denmark, 2002.
[18] V. Blanz, and T. Vetter, "A Morphable Model for The Synthesis of 3D Faces," in Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH '99), 1999.
[19] X. Yu, J. Huang, S. Zhang *et al.*, "Pose-free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model," in IEEE International Conference on Computer Vision, 2013.