ROBUST FACIAL POSE ESTIMATION USING LANDMARK SELECTION METHOD FOR BINOCULAR STEREO VISION

Jaeseong Park, Suwoong Heo, Kyungjune Lee and Sanghoon Lee

Yonsei University
Department of Electric and Electronic Engineering
Seoul, Republic of Korea, 120-749

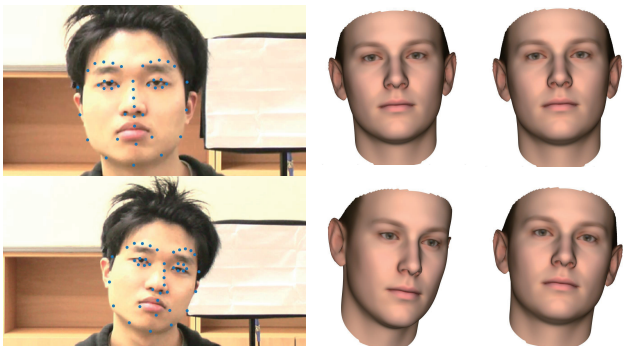
ABSTRACT

In this paper, we present a robust framework for facial pose estimation from binocular stereoscopic vision. Unlike prior work on the facial pose estimation that employs the whole landmarks even located in the wrong position, we propose a landmark selection method to remove the erroneous landmarks for better performance, especially in the large facial pose case. For this purpose, we train a convolutional neural network (CNN) in order to measure the confidence of each facial landmark detected by using a well-known landmark detection algorithm. Also, by fitting selected landmarks to 3D space, our framework becomes more robust even when a small number of landmarks are selected. Due to the absence of public dataset for the binocular stereo facial pose, we construct facial pose data sets using a motion sensor for performance validation. In our experiments, our method achieves the higher accuracy of the pose estimation than the previous method, especially for large facial pose cases.

Index Terms— Facial pose estimation, Binocular stereo vision, Facial landmarks, Stereo matching algorithm

1. INTRODUCTION

Facial pose estimation is a task which predicts the orientation of the target head with respect to some camera coordinate with three directions-pitch, yaw and roll. Murphy-Chutorian and Trivedi [1] distinguish facial pose estimation methods that require monocular vision and depth information from stereo vision. In the monocular vision, most of the facial pose estimation algorithms directly estimate the facial pose with machine learning methods [2, 3]. These methods estimate facial pose based on training several facial images [2, 3], part-based model [4] or 2D facial landmarks. However, due to the inherent problem of the facial image such as illumination, shadow, image quality [5, 6] and a lack of features such as occlusions, direct estimation of the facial pose is prone to be inaccurate. Recently, Yang et al. [7] trained a convolutional network with monocular face images. However, this method works unstably for a large pose.



(a) (b) (c)

Fig. 1: (a) A binocular image pair with landmarks obtained by the stereo camera system. (b) Facial pose estimation with whole landmarks. (c) Facial pose estimation by the proposed method.

Rather than directly estimating the facial pose from an image, fitting the landmarks of the 3D shape model on the 2D facial landmarks can be a more accurate and reliable for facial pose estimation [8]. By computing the projection of 3D points into the corresponding 2D coordinate system [9], the facial pose can be computed as 3 degrees of freedom (DOF) rotations consisting of rotation in pitch, roll and yaw directions. In most cases, calculating the exact correspondence of a 3D point to the 2D landmark is almost impossible. As shown in the second row of Fig. 1 (b), this problem is formidable when the variation of facial pose is so large that some of the landmarks are detected in the incorrect location. Thus, in order to estimate large facial pose, a small subset of landmarks should be taken into account. However, if the number of correspondences between 3D and 2D landmarks is not sufficient, facial pose estimation with the subset of landmarks would fail because of 3D-2D projection ambiguity. Fortunately, 3D information of 2D points can be obtained by stereo imaging or depth camera, which can help to resolve the problem caused by the projection ambiguity. There are several facial pose estimation methods using 3D data that provide depth information

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2016R1A2B2014525)

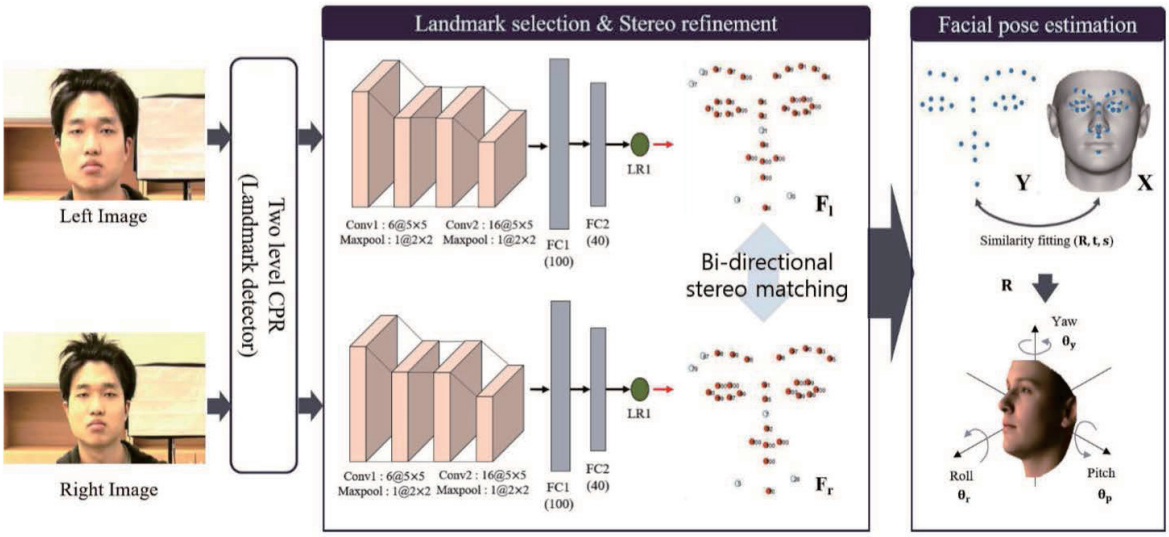


Fig. 2: Overview of the proposed framework. Conv, FC and LR are convolution, fully connected, and logistic regression layers respectively.

[10, 11].

In this paper, we present a robust large facial pose estimation using a selected small subset of landmarks. From facial images obtained by a binocular stereo camera, we detect a landmark set by a landmark selection method. After detecting landmarks, we select landmarks confidently located in their correct locations. Then the locations of landmarks in the set are refined through a bi-directional matching scheme. The 3D information of selected landmarks can be easily computed by using a stereo matching algorithm [12]. As shown in the second row of Fig. 1 (c), we can obtain a reliable and accurate facial pose by fitting landmarks on a shape model into computed 3D landmarks from input images. The strength our algorithm can be summarized into 2 points.

- By extracting 3D facial landmarks from stereo vision, we reliably resolve the projective ambiguity problem.
- Also, by selecting confident landmarks only, our method is less dependent on the performance of the landmark detector than existing methods.

2. FACIAL POSE ESTIMATION FRAMEWORK

Our proposed framework is shown in Fig. 2. The input of our system is a set of stereo images. To acquire an initial landmark set, we use the two-level cascade pose regressor proposed by X. Cao et al. [13]. Next, a subset of landmarks is selected according to the confidence of each landmark by using the convolutional neural network (CNN) [14, 15, 16]. Also, bi-directional matching between the landmark points on the left and right images based on the normalized cross correlation measure is conducted in order to correct each landmark location. Then those 2D landmarks are converted into 3D landmarks by using a stereo matching algorithm. Let the target landmark set be $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_L\}$ which is obtained from the matching algorithm where $\mathbf{y}_i \in \mathbb{R}^{3\times1}$ is an ith element of the target landmark set. By solving the non-linear optimization problem, the transformation which contains rotation $\mathbf{R} \in \mathbb{R}^{3\times3}$, translation $\mathbf{t} \in \mathbb{R}^{3\times1}$ and anisotropic scale $\mathbf{s} \in \mathbb{R}^{3\times1}$ between the source landmark set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_L\}$ where $\mathbf{x}_i \in \mathbb{R}^{3\times1}$ and target landmark set $\mathbf{Y}$ is performed. From this transformation, the facial pose expressed by the rotation that consists of angles in the directions of pitch, yaw and roll can be obtained.

2.1. Facial landmark detection

We detect facial landmarks using two-level cascade pose regressor. Let the 2D facial shape be the set of landmarks detected from an image denoted as $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3 \ldots \mathbf{f}_L)$ where $\mathbf{f}_i = (u_i, v_i)$ is a position vector of each landmark. This method updates the initial shape $\mathbf{F}^0$ which is a mean shape of several augmented training shapes into actual shape $\mathbf{F}$ in a coarse to fine manner. This method shows an accurate result when the facial pose variation is small compared to the training data set. However, as shown in the second row of Fig. 1 (a), the initial shape does not always converge to the actual shape. In other words, the errous landmark positions can be obtained for the largely varying pose. Most of landmark detection algorithms update the facial shape from their initialization methods. Thus, if the initial shape is not located within the range of actual shape, the converged landmark would be located in a wrong position. So we estimate the posterior probability of whether each landmark is aligned on its proper location or not by using a CNN in the next

subsection.

2.2. Landmark selection method and stereo refinement

In order to choose aligned landmarks rigorously, we propose a landmark selection method by computing the posterior probability which we call confidence in this paper. For L landmarks, the confidence of the ith landmark can be expressed as posterior probability $p\left(y_{\mathrm{i}}|\mathbf{f}_i, \mathbf{\Phi}_i\right)$ where y_{i} is 1 if the ith landmark patch $\mathbf{\Phi}_i \in \mathbb{R}^{m\times m}$ centered at $\mathbf{f}_i \in \mathbb{R}^2$ is considered as aligned and is 0 otherwise.

In order to compute this, we designed a CNN architecture as shown in Fig. 2. The confidences $p_i = p\left(y_{\mathrm{i}} = 1|\mathbf{f}_i, \mathbf{\Phi}_i\right)$ and $(1-p_i) = p\left(y_{\mathrm{i}} = 0|\mathbf{f}_i, \mathbf{\Phi}_i\right)$ are computed in the logistic regression layer. In order to train the network, the training patches for each landmark are built by extracting $m \times m$ patches centered at different locations. For each landmark in training images, positive patches are extracted near the landmark location. The negative patches are constructed by randomly clipping patches far from the true landmarks with the same patch size. By using those patches, the CNN network is trained using a cross-entropy cost function with L2 regularization :

$$\sum_{i=1}^{N} \hat{y}_i \ln\left(p_i\right) + \left(1-\hat{y}_i\right)\left(1-\ln\left(p_i\right)\right) + \lambda||\mathbf{W}||_2^2, \quad (1)$$

where N is the the number of positive and negative patches, $\mathbf{W}$ is a weight vector of the network associated with the constant λ and $\hat{y}_i \in \{0, 1\}$ is the ground-truth label. Note that the true label is set to 1 if $\mathbf{\Phi}_i$ belongs to positive patches and set to 0 otherwise. Inspired by E. sariyadini et al. [12], the selection rule is designed to minimize the false positive rate. In runtime, after computing the confidence values for each landmark location, the landmark with confidence higher than the threshold τ which is determined by examining the receiver operating characteristic (ROC) curve is selected. As shown in Fig. 3, the landmarks within its reliable location are selected by using the proposed method.

After cutting out the landmarks that have a confidence value of less than τ, the locations of $L_S\ (< L)$ landmarks are refined by using bi-directional pixel matching scheme [17] for robust stereo matching. Assuming that conjugate epipolar lines between two are collinear and parallel to one of the image axes, the search line for pixel matching can be set to a horizontal line with small bandwidth. With this assumption, the search range for stereo matching is restricted. The forward matches from landmark points in the left image to the right image are computed by using a normalized cross correlation measure. Then the backward matches are computed the same manner. For a correspondence, if the position of the point found in the backward match is more than 5 pixels from the position of the original landmark, other points inside the search window are examined. This process progresses iteratively until this constraint is satisfied.

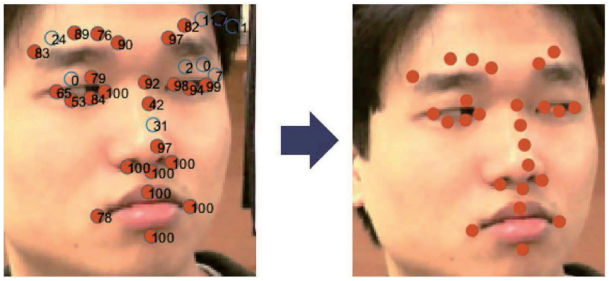


Fig. 3: The landmark selection method compute the confidence of each landmark expressed as percentage. In above example, a subset of landmarks in the left images are selected based on the threshold τ as in the right image.

2.3. Facial pose estimation

In the proposed framework, the facial pose is estimated by aligning the source landmark set $\mathbf{X}$ which is a subset of a 3D shape model into some target landmark set $\mathbf{Y}$. A target landmark point $\mathbf{y}_i$ is computed by using the disparity value of the 2D selected landmark points. That is, $\mathbf{y}_i = (X_i, Y_i, Z_i)$ can be obtained by using disparity $d_i = (u_{l,i} - u_{r,i})$:

$$\begin{aligned} X_i &= \frac{b}{2d_i}\left(u_{l,i} + u_{r,i} - u_0\right), \\ Y_i &= \frac{b}{d_i}\left(u_{l,i} - v_0\right), Z_i = \frac{bf}{d_i}, \end{aligned} \quad (2)$$

where b is the baseline of the camera system, f is the focal length of the stereo camera and $(u_0,\ v_0)$ is the principle point of binocular stereoscopic images.

Then the similarity transformation containing $\mathbf{R}$, $\mathbf{t}$ and $\mathbf{s}$ between $\mathbf{X}$ and $\mathbf{Y}$ is computed by solving following nonlinear least square optimization problem :

$$(\mathbf{R}, \mathbf{t}, \mathbf{s}) = \underset{\mathbf{R},\mathbf{t},\mathbf{s}}{\operatorname{argmin}} \sum_{i=1}^{L} \|(diag\left(\mathbf{s}\right)\mathbf{R}\mathbf{x}_i + \mathbf{t}) - \mathbf{y}_i\|_2^2, \quad (3)$$

where $diag\left(\cdot\right)$ is a digonalization operator for the vector. This optimization problem is solved by using the Levenberg-Marquardt algorithm [18]. For efficient initialization, the rotation $\mathbf{R}$ and translation $\mathbf{t}$ are initialized by using a singular value decomposition based method [19]. After the convergence of $\mathbf{R}, \mathbf{t}$ and $\mathbf{s}$, the facial pose can be obtained by the rotation matrix $\mathbf{R}$. Let the (i, j) element of $\mathbf{R}$ is r_{ij}. Then the facial pose is expressed as :

$$\begin{aligned} \theta_p &= \arctan\left(\frac{r_{32}}{r_{33}}\right), \\ \theta_y &= \arctan\left(\frac{-r_{31}}{\sqrt{r_{32}^2 + r_{33}^2}}\right), \\ \theta_r &= \arctan\left(\frac{r_{21}}{r_{11}}\right), \end{aligned} \quad (4)$$

where θ_p, θ_y and θ_r are the angles corresponding to the directions of pitch, roll and yaw direction respectively.

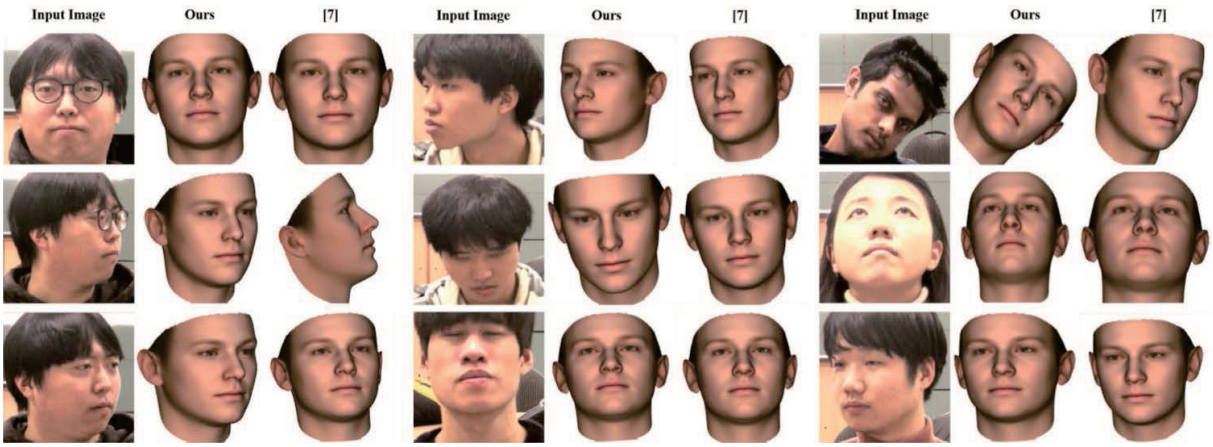


Fig. 4: Results of facial pose estimation using various facial poses. For the number of landmarks L, we set $L = 33$ without contour landmarks for our method and $L = 68$ for the method from [8].

3. EXPERIMENTAL RESULTS

Since it is difficult to obtain the binocular stereo database for facial pose estimation publicly available, we constructed the dataset by capturing 20 subjects by using a stereo camera. The stereo camera system used in this experiment is calibrated by MATLAB Stereo Camera Calibrator [20]. By using a motion sensor, the ground truth value corresponding to each facial pose is acquired. The landmark detector and the CNN for the landmark selection are trained with the Labeled Face Parts in the Wild (LFPW) dataset [21]. We use the Basel face model (BFM) [22] for 3D shape model. The parameters in our experiment are chosen as $m = 25$, $\lambda = 10^{-4}$ and $\tau = 30$. For each landmark, the CNN is trained with $N = 87,000$ samples. The final mean test accuracy of the trained networks for whole landmarks is about 95%.

The qualitative result of our method is presented in Fig. 4. The compared method estimates the facial pose by fitting the landmarks on the 3D shape model directly into the landmarks on both left and right images by using [8]. Since our method only uses a subset of the landmarks which are confidently located, the result of the proposed method is closer to the ground truth pose particularly for large pose cases in yaw and roll directions than the result by using [8]. Through our framework, a large range of facial poses can be obtained reliably.

For quantitative measure, the mean absolute error (MAE) between the ground truth pose measured by the motion sensor and the estimated pose in large and small variations are computed. Table 1 shows the mean and variances of estimation errors of the 20 subjects using our method with and without landmark selection and the method in [8]. In the table, each facial pose is categorized into the large pose with $60°$, $70°$ and $40°$ and small pose with $30°$, $20°$ and $20°$ in pitch, yaw and roll directions respectively. In the small pose case, since most of the landmarks are reliably located in its true position, the estimation error of both methods is similar and the effect of landmark selection method is not significant. However, in the large pose, the result by the framework with landmark selection is significantly improved. In summary, our method achieves consistent accuracy regardless of the direction and the variation of the facial pose.

	Pose	Pitch error	Yaw error	Roll error
Ours (L_S)	Large	$7.6°/2.8°$	$10.9°/3.7°$	$11.3°/4.4°$
	Small	$5.7°/2.7°$	$7.2°/3.4°$	$5.9°/2.8°$
Ours (L)	Large	$9.2°/4.3°$	$15.5°/8.6°$	$10.4°/5.9°$
	Small	$5.9°/3.3°$	$8.1°/4.1°$	$5.8°/2.6°$
[8]	Large	$13.9°/7.5°$	$18.2°/10.7°$	$15.3°/8.1°$
	Small	$5.3°/3.2°$	$7.5°/2.9°$	$5.2°/3.3°$

Table 1: Quantitative results of our proposed framework with (L_S) and without (L) landmark selection compare to the method in [8]. The values in the table represent a mean and standard deviation of MAE in degree.

4. CONCLUSION

In this paper, we present a robust framework for facial pose estimation from binocular stereoscopic vision. The proposed framework is robust to the self-occlusion because the proposed landmark selection method removes the landmarks located far from its true position. Also, our method show the high accuracy even when using a small number of landmarks since it directly fits the images to the 3D space rather than projecting 3D shapes into images. The experiments demonstrate that our method accurately estimates the facial pose even in relatively large pose cases. In addition, our method can be extended to the data captured from RGB-D sensors. Though, it is still difficult to cope with object occlusion, which is out of scope in this paper. We hope to deal with this issue in a further paper.

5. REFERENCES

[1] E. Murphy-Chutorian and MM Trivedi, "Head pose estimation in computer vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[2] S. Z. Li, X. G. Lu, X. Hou, X. Peng, and Q. Cheng, "Learning multiview face subspaces and facial pose estimation using independent component analysis," *IEEE Transactions on Image Processing*, vol. 14, no. 6, pp. 705–712, 2005.

[3] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[4] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[5] Jongyoo Kim, Taewan Kim, Sanghoon Lee, and Alan Conrad Bovik, "Quality assessment of perceptual crosstalk on two-view auto-stereoscopic displays," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4885–4899, 2017.

[6] Heeseok Oh, Sewoong Ahn, Jongyoo Kim, and Sanghoon Lee, "Blind deep s3d image quality evaluation via local to global feature aggregation," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4923–4936, 2017.

[7] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson, "Face alignment assisted by head pose estimation," *arXiv preprint arXiv:1507.03148*, 2015.

[8] A. Bas, W. A. Smith, T. Bolkart, and S. Wuhrer, "Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 377–391.

[9] D. F. Dementhon and L. S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision*, vol. 15, no. 1-2, pp. 123–141, 1995.

[10] M. D. Breitenstein, J. Jensen, C. Høilund, T. B. Moeslund, and L. Van Gool, "Head pose estimation from passive stereo images," in *Scandinavian conference on image analysis*. Springer, 2009, pp. 219–228.

[11] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 617–624.

[12] E. Saryanidi, H. Gunes, and A. Cavallaro, "Robust registration of dynamic facial sequences," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1708–1722, 2017.

[13] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[16] Jongyoo Kim and Sanghoon Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, 2017.

[17] W. Li, D. Cosker, and M. Brown, "An anchor patch based optimization framework for reducing optical flow drift in long image sequences," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 112–125.

[18] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*, pp. 105–116. Springer, 1978.

[19] O.Sorkine, "Least-squares rigid motion using svd," *Technical notes*, vol. 120, no. 3, pp. 52, 2009.

[20] Zhengyou Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[21] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[22] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*. IEEE, 2009, pp. 296–301.