IDENTITY-INVARIANT FACIAL LANDMARK FRONTALIZATION FOR FACIAL EXPRESSION ANALYSIS

*Vassilios Vonikakis**

Amazon Web Services
Singapore
vonikakv@amazon.com

*Stefan Winkler**

School of Computing
National University of Singapore (NUS)
winkler@comp.nus.edu.sg

ABSTRACT

We propose a frontalization technique for 2D facial landmarks, designed to aid in the analysis of facial expressions. It employs a new normalization strategy aiming to minimize identity variations, by displacing groups of facial landmarks to standardized locations. The technique operates directly on 2D landmark coordinates, does not require additional feature extraction and as such is computationally light. It achieves considerable improvement over a reference approach, justifying its use as an efficient preprocessing step for facial expression analysis based on geometric features.

1. INTRODUCTION

Facial Expression Analysis (FEA) has attracted strong attention in recent years, both in the research community [1] as well as in real-world systems and applications [2]. However, faces are some of the most challenging objects to register. First, faces can exhibit a broad range of deformations which may significantly alter their shape and appearance, especially for high intensity expressions. Second, there is significant variability across individuals, resulting in considerable differences in appearance. This identity variability may further be accentuated by age, gender and race. Finally, perhaps the most important factor is the variability of faces across different viewpoints. As such, different head poses may result in dramatically different face appearance. Therefore, FEA systems operating in uncontrolled conditions have to address the head pose variability issue across different identities and expressions. Frontalization approaches offer a solution to this problem by recovering the frontal view of non-frontal facial images [3–6]. As such, training may take place on 'frontal' datasets, while during inference frontalization ensures that facial images are transformed to frontal before any analysis.

Frontalization can be classified into appearance-based and landmark-based approaches. Appearance-based frontalization attempts to recover the full frontal appearance of the face. Usually this is achieved by fitting a 3D model on the non-frontal face, projecting pixels on the fitted model and rotating it back to frontal using texture warping [3–5, 7]. Recently, Generative Adversarial Networks (GANs) have been successfully used to reconstruct the frontal view of non-frontal faces [8–10]. Landmark-based frontalization focuses on estimating only the location of facial landmarks as they would have looked from the front [6]; no pixel appearance is estimated. As such, these approaches tend to be considerably less computationally expensive, since they do not require pixel rendering or generative networks. In recent years, directly estimating 3D facial landmarks from 2D images has become possible [11–14], simplifying the estimation of the frontal landmark view. However, these methods are still computationally expensive and not as widely used in FEA systems. In many real-world scenarios, computationally light FEA systems are required, e.g. for real-time execution in edge devices. In such cases, generative networks and expensive 3D reconstruction methods are usually not an option. Instead, FEA systems solely based on geometric features (derived from facial landmarks) are a better choice. Even though appearance is not used in this approach, it has been shown to exhibit competitive performance [15].

In this paper we follow the latter approach. More specifically, we focus on the frontalization of faces for FEA, aiming to discount two of the three major sources of facial variability (identity and viewpoint), while preserving the third (deformations a.k.a. expressions). We also focus on a computationally light approach that allows for real-time landmark frontalization. To this end, we introduce a new fast facial landmark frontalization method that compensates for viewpoint and identity variations in FEA tasks.[1] The technique is data-driven and operates directly on the 2D landmark coordinates without the need for additional feature extraction. Once trained, it requires minimal overhead, since computation is reduced to a simple matrix multiplication. Furthermore, we introduce a new landmark normalization strategy, which minimizes identity variations by displacing facial parts to stan-

* This work was conducted while both authors were with the Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore. Contact: vonikakv@amazon.com

[1] A sample Python implementation of the method is available at https://github.com/bbonik/facial-landmark-frontalization

© IEEE 2020

dardized locations. This standardization further improves the performance of the proposed method. Experimental results demonstrate that our technique exhibits superior performance in comparison to an existing reference approach. As such, it is suitable as an efficient preprocessing step in landmark-based FEA systems, offering increased robustness to non-frontal headposes.

2. 2D LANDMARK FRONTALIZATION

Let ${}^{k}\mathbf{p}_i^j \in \mathbb{R}^{2N}$ be a vector containing the estimated x and y coordinates of N facial landmarks from subject i with facial expression $j \in \{1, \cdots, J\}$ from viewpoint $k \in \{0, \cdots, K\}$. Let ${}^{0}\mathbf{p}_i^j$ denote the frontal view of the j^{th} expression of subject i. Matrix $\mathbf{A}$ contains the transposed version of all landmark coordinate vectors ${}^{k}\mathbf{p}_i^j$, while $\mathbf{Y}$ is the 'ground truth' of their corresponding frontal view coordinates ${}^{0}\mathbf{p}_i^j$. With this, the required frontalization mapping $\mathbf{X}$ can be expressed through the following optimization:

$$\text{argmin}_{\mathbf{X}} \left\|\mathbf{Y} - \mathbf{A}\mathbf{X}\right\|_2 + \lambda \left\|\mathbf{X}\right\|_2, \tag{1}$$

which can be solved for $\hat{\mathbf{X}}$ in closed-form:

$$\hat{\mathbf{X}} = \left(\mathbf{A}^\top \mathbf{A} + \lambda \mathbf{I}\right)^{-1} \mathbf{A}^\top \mathbf{Y}, \tag{2}$$

where λ is the regularization term, and $\hat{\mathbf{X}} \in \mathbb{R}^{(2N+1)\times(2N)}$ represents the least-squares approximate solution, containing the frontalization weights for the mapping. It is important that matrices $\mathbf{A}$ and $\mathbf{Y}$ are filled in such a way that the landmark coordinates of every given non-frontal expression of a subject are mapped to the landmark coordinates of the corresponding frontal expression of the *same* subject. In the following example, single horizontal lines represent the transition to another expression of the same subject, while double horizontal lines represent the transition to a different subject.

$$\mathbf{A} = \begin{bmatrix} 1 & {}^{1}\mathbf{p}_1^{1\top} \\ \vdots & \vdots \\ 1 & {}^{K}\mathbf{p}_1^{1\top} \\ \hline \vdots & \vdots \\ \hline 1 & {}^{1}\mathbf{p}_1^{J\top} \\ \vdots & \vdots \\ 1 & {}^{K}\mathbf{p}_1^{J\top} \\ \hline\hline 1 & {}^{1}\mathbf{p}_2^{1\top} \\ \vdots & \vdots \\ 1 & {}^{K}\mathbf{p}_2^{1\top} \\ \hline \vdots & \vdots \\ \hline 1 & {}^{1}\mathbf{p}_2^{J\top} \\ \vdots & \vdots \\ 1 & {}^{K}\mathbf{p}_2^{J\top} \\ \hline\hline \vdots & \vdots \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} {}^{0}\mathbf{p}_1^{1\top} \\ \vdots \\ {}^{0}\mathbf{p}_1^{1\top} \\ \hline \vdots \\ \hline {}^{0}\mathbf{p}_1^{J\top} \\ \vdots \\ {}^{0}\mathbf{p}_1^{J\top} \\ \hline\hline {}^{0}\mathbf{p}_2^{1\top} \\ \vdots \\ {}^{0}\mathbf{p}_2^{1\top} \\ \hline \vdots \\ \hline {}^{0}\mathbf{p}_2^{J\top} \\ \vdots \\ {}^{0}\mathbf{p}_2^{J\top} \\ \hline\hline \vdots \end{bmatrix} \tag{3}$$

In order to train and evaluate the frontalization module, 5 different datasets listed in Table 1 are employed. These datasets contain portrait images of different people with varying facial expressions from multiple viewpoints. All available images from the 5 datasets were scanned using OpenCV's frontal and profile face detectors. All valid detected faces were analyzed using Supervised Descent Method (SDM) [16] in order to estimate the 2D facial landmarks. The total number of valid detected faces was 86,844, which were doubled to 173,688 after mirroring. Each set of raw coordinates was separately normalized for scaling, translation and rotation (see Section 2.1), before filling the $\mathbf{A}$ and $\mathbf{Y}$ matrices and estimating the frontalization weights $\hat{\mathbf{X}}$.

Table 1. Datasets used for training (top 4 rows) and testing (bottom row).

Dataset	Pitch	Yaw
Radboud [17]	–	$0°$, $\pm45°$, $\pm90°$
Karolinska [18]	–	$0°$, $\pm45°$, $\pm90°$
CAS-PEAL [19]	$\pm15°$	$0°$, $\pm22°$, $\pm45°$, $\pm67°$, $\pm90°$
PIE [20]	$\pm15°$	$0°$, $\pm22°$, $\pm45°$, $\pm67°$, $\pm90°$
Multi-PIE [21]	$-30°$	$0°$, $\pm15°$, $\pm30°$, $\pm45°$, $\pm60°$, $\pm75°$, $\pm90°$

The computational overhead introduced by the proposed frontalization approach at runtime is minimal, since for N facial landmarks, this equates to a single vector-matrix multiplication between the landmark coordinate vector $\mathbf{p} \in \mathbb{R}^{2N+1}$ ($2N$ landmark coordinates plus the intercept) and the frontalization matrix $\hat{\mathbf{X}} \in \mathbb{R}^{(2N+1)\times(2N)}$.

2.1. Normalization of Landmarks

Let $\mathbf{P}_i^N \in \mathbb{R}^{N\times2}$ be a matrix containing the x and y coordinates of N facial landmarks of facial image i. In order to properly learn the frontalization matrix $\hat{\mathbf{X}}$, all coordinates in matrices $\mathbf{A}$ and $\mathbf{Y}$ should be appropriately standardized for translation, scaling and rotation. For this, we employ a non-isotropic version of the Procrustes transformation.

Let $n \subset N$ represent a subset of "stable" facial landmarks that change minimally with different expressions. These include facial points on the eyes, the nose, and the middle of the upper lip. Eyebrows and the majority of the mouth are excluded, since their coordinates may change significantly with different expressions, thus, introducing additional variations that may affect scale. In our case, $N = 49$ and $n = 13$. Let $\mathbf{P}_i^n \in \mathbb{R}^{n\times2}$ represent the coordinate matrix of the n stable points. Then the standardized coordinate vector $\hat{\mathbf{P}}_i^N$ is estimated as:

$$\hat{\mathbf{P}}_i^N = \left(\frac{\left\|\mathbf{P}_i^n - {}^{n}\overline{\mathbf{P}}_i^n\right\|_F}{\sqrt{n}}\right) \cdot \left(\mathbf{P}_i^N - {}^{N}\overline{\mathbf{P}}_i^n\right) \cdot \mathbf{R}_i, \tag{4}$$

where ${}^{x}\overline{\mathbf{P}}_i^n \in \mathbb{R}^{h\times 2}$ denotes a matrix containing the mean values of the columns of $\mathbf{P}_i^n$, replicated for x rows. $\|\cdot\|_F$ represents the Frobenius norm, while $\mathbf{R}_i \in \mathbb{R}^{2\times 2}$ is the following rotation matrix:

$$\mathbf{R}_i = \begin{bmatrix} \cos(a) & -\sin(a) \\ \sin(a) & \cos(a) \end{bmatrix}, a = \text{atan}\left(\frac{x_{eyeR} - x_{eyeL}}{y_{eyeR} - y_{eyeL}}\right)$$

with x_{eye}, y_{eye} denoting the coordinates of the centers of the left and right eyes in image i.

Equation (4) has 3 distinct terms, standardizing scaling, translation, and rotation, respectively. The scaling factor is essentially the average Euclidean distance of the stable landmarks to their corresponding mean. The fact that only the mean values of the stable points n are used for the calculation of the scaling and translation factors adds robustness to different facial expressions. Conversely, if all the N landmarks were used, then various facial expressions would affect the scale and the translation of the standardized landmark coordinates. This approach is also more robust to yaw changes, compared to the widely used intra-ocular (eye-to-eye) distance, since in non-frontal faces, the eye-to-eye distance becomes smaller, and thus, normalizing by it affects the scaling term.

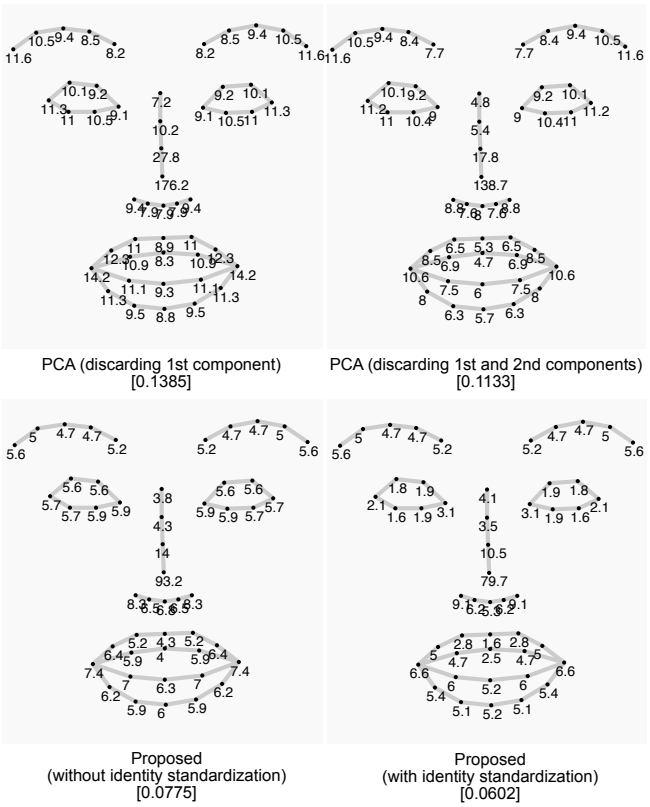

Fig. 1. Per landmark average normalized errors (not showing $\times 10^{-2}$) and total average normalized errors (bottom, square brackets) of different frontalization approaches. Normalized errors are unitless.

2.2. Identity Standardization

Although the above standardization approach eliminates variabilities related to scale, translation and rotation, a person's identity may still introduce unnecessary variations that can affect the frontalization performance. This is due to the fact that every person has different relative positions of landmark groups, e.g. distance between the eyes, distance of the upper mouth to the bottom of the nose, upper starting point of the nose etc. During our experiments we found that eliminating this type of identity variability further improves the performance of the frontalization module.

To this end, we introduce additional translation terms only to the frontal ground truth coordinates of matrix $\mathbf{Y}$. These translation terms displace whole groups of facial landmarks, such as eyes, mouth and nose, and re-position them in the corresponding group locations of a *template* face. By doing this, we force the optimization to learn a more *identity-invariant* frontalization transformation, from non-frontal landmarks towards a standardized frontal face, discounting individual face variabilities.

Let $\{\mathcal{E}_L, \mathcal{E}_R, \mathcal{M}, \mathcal{N}\} \subset N$ represent subsets of facial landmarks, corresponding to the left and right eye, the mouth and the nose, respectively. Let $\mathbf{P}_i^{\mathcal{E}_L}, \mathbf{P}_i^{\mathcal{E}_R}, \mathbf{P}_i^{\mathcal{M}}, \mathbf{P}_i^{\mathcal{N}}$ represent the landmark coordinates of the corresponding subsets for facial image i, and $\mathbf{P}_T^{\mathcal{E}_L}, \mathbf{P}_T^{\mathcal{E}_R}, \mathbf{P}_T^{\mathcal{M}}, \mathbf{P}_T^{\mathcal{N}}$ the landmark coordinates for the template face. Then the identity-standardized frontal coordinates $\tilde{\mathbf{P}}_i^x$ of matrix $\mathbf{Y}$ are given by the following equations:

$$\begin{aligned} \tilde{\mathbf{P}}_i^x &= \hat{\mathbf{P}}_i^x + \Delta\hat{\mathbf{P}}_i^x, \\ \Delta\hat{\mathbf{P}}_i^x &= \bar{\hat{\mathbf{P}}}_T^x - \bar{\hat{\mathbf{P}}}_i^x, \end{aligned} \tag{5}$$

with $x \in \{\mathcal{E}_L, \mathcal{E}_R, \mathcal{M}, \mathcal{N}\}$ and $\bar{\hat{\mathbf{P}}}^x$ representing an anchor point in the landmarks of subset x. For subsets $\mathcal{E}_L, \mathcal{E}_R$ and $\mathcal{N}$, the anchor point used was the mean of their standardized coordinates $\hat{\mathbf{P}}_i^x$. For $\mathcal{M}$, the upper middle landmark was used as anchor point, since different mouth deformations can affect its coordinate mean considerably, introducing fluctuations. Matrix $\Delta\hat{\mathbf{P}}_i^x$ adds a displacement in the standardized coordinates, such that the different landmark subsets are always located in the same location as in the template face T. In our experiments, the mean frontal face of the training datasets was used as template face.

The displacement term does not affect the shape of the landmark subsets, but only their *relative* position within the face. For example, a smiling mouth will maintain its shape, and only its position within the face will change. Unlike all other facial parts, eyebrows are omitted from the identity stan-

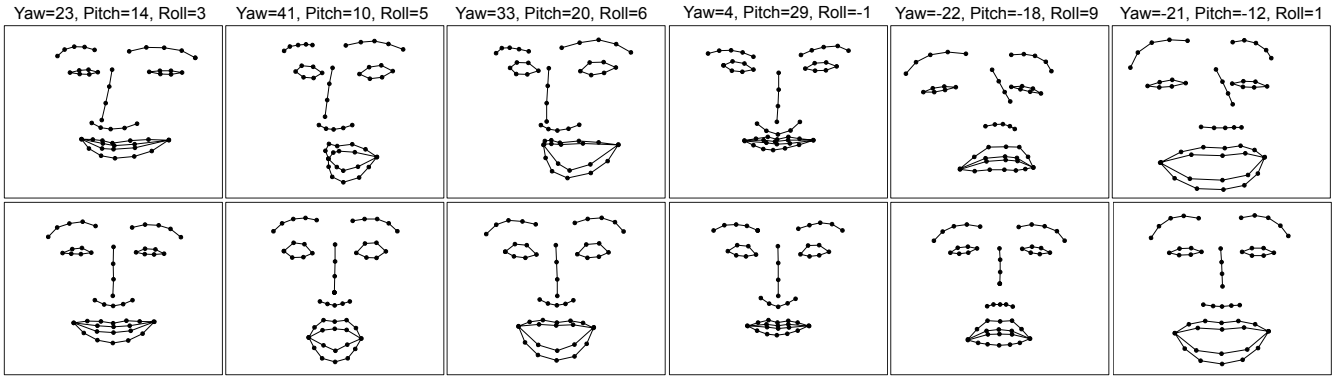

Fig. 2. Example outputs of the frontalization module on unseen non-frontal images of various head poses and expressions. Top: Original landmarks and head pose estimated by SDM. Bottom: Output of the proposed frontalization module.

dardization, because their relative position conveys important expression information; e.g. in a surprised expression, eyebrows tend to move up away from the eyes. As such, standardizing their relative position would discount this important information for FEA.

The proposed identity standardization takes place only once, during training of the frontalization module. At runtime, all facial landmarks are standardized only for scale, rotation and translation, using Equation (4), while the frontalization module will estimate their frontal view, suppressing individual identity differences and maintaining the expression.

3. EXPERIMENTAL RESULTS

In order to quantitatively evaluate the performance of the proposed frontalization approach, a frontalization module was trained using 4 out of the 5 datasets of Table 1 (Radboud, Karolinska, CAS-PEAL, PIE) and tested on the holdout dataset (Multi-PIE). The normalized least square residual $\left\| f({}^k\mathbf{p}_i^j) - {}^0\mathbf{p}_i^j \right\| / \left\| {}^0\mathbf{p}_i^j \right\|$ was used to measure frontalization performance, which is the same metric used in face alignment [16]. $f({}^k\mathbf{p}_i^j)$ represents the frontalized view of the non-frontal landmarks ${}^k\mathbf{p}_i^j$, ${}^0\mathbf{p}_i^j$ denotes its actual ground truth frontal landmarks, provided by the test set, while their ratio is a unitless number.

Fig. 1 depicts the frontalization errors of two versions of the proposed method (with and without identity standardization), as well as that of a PCA-based approach [6].[2] Our proposed approach exhibits considerably smaller errors than the PCA-based approach (roughly by a factor of two). Additionally, identity standardization further reduces the error across all landmarks, resulting in a more robust frontalization.

Fig. 2 depicts example outputs of the final frontalization approach (trained with all 5 datasets of Table 1) on unseen images captured by a web camera. Based on our qualitative tests, the frontalization module is quite robust to yaw ranges within $\pm 45^\circ$ and pitch ranges within $\pm 15^\circ$. Beyond those ranges, performance drops, but still remains reasonable for yaw ranges in $\pm[45^\circ, 60^\circ]$ and pitch ranges in $\pm[15^\circ, 30^\circ]$. It was not possible to test the performance above $\pm 60^\circ$ for yaw or $\pm 30^\circ$ for pitch, since this is beyond the operating range of the face and landmark detectors; faces with more extreme head poses would either yield unreliable facial landmarks or would not be detected at all.

We also observe that there is an imbalance in performance between yaw and pitch. This could be due to two possible reasons. First, the datasets that we used had considerably fewer training examples for pitch variations compared to yaw. Second, yaw frontalization is inherently easier compared to pitch, since the natural left-right symmetry of the face helps with it.

4. CONCLUSIONS

We presented a new technique for frontalizing 2D facial landmarks, which is designed for facial expression analysis. Our approach employs a new point normalization strategy that aims to minimize identity variations and shifts different facial parts to standard locations. The technique operates directly on 2D landmark coordinates and does not require additional feature extraction. As such, it adds minimal computational overhead making it suitable for real-time systems. Benchmarking with several face datasets shows that it outperforms a PCA-based reference approach by a substantial margin, approximately halving average errors.

[2] According to [6], the first principal components with the largest capacity to explain variance are associated with yaw and pitch. Consequently, if discarded during reconstruction, yaw and pitch are discounted and thus the landmarks are frontalized.

5. REFERENCES

[1] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, June 2015.

[2] V. Vonikakis and S. Winkler, "Emotion-based sequence of family photos," in *Proc. 20th ACM International Conference on Multimedia*. ACM, 2012, p. 1371–1372.

[3] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 4295–4304.

[4] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Effective 3D based frontalization for unconstrained face recognition," in *Proc. 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 1047–1052.

[5] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1701–1708.

[6] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic, "FERA 2017 – addressing head pose in the third facial expression recognition and analysis challenge," in *Proc. 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2017, pp. 839–847.

[7] W. Deng, J. Hu, Z. Wu, and J. Guo, "Lighting-aware face frontalization for unconstrained face recognition," *Pattern Recognition*, vol. 68, pp. 260–271, 2017.

[8] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 4010–4019.

[9] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 9843–9850.

[10] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1283–1292.

[11] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D video for real-time use," *Image and Vision Computing*, vol. 58, pp. 13–24, 2017.

[12] S. Tulyakov, L. A. Jeni, J. F. Cohn, and N. Sebe, "Viewpoint-consistent 3D face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 9, pp. 2250–2264, Sept 2018.

[13] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 1021–1030.

[14] R. Zhao, Y. Wang, and A. M. Martinez, "A simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 3059–3066, Dec 2018.

[15] V. Vonikakis, Y. Yazici, V. D. Nguyen, and S. Winkler, "Group happiness assessment using geometric features and dataset balancing," in *Proc. 18th ACM International Conference on Multimodal Interaction (ICMI)*. ACM, 2016, pp. 479–486.

[16] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 532–539.

[17] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the Radboud faces database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[18] E. Goeleven, R. D. Raedt, L. Leyman, and B. Verschuere, "The Karolinska directed emotional faces: A validation study," *Cognition and Emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.

[19] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 38, no. 1, pp. 149–161, Jan 2008.

[20] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, Dec 2003.

[21] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.