Real-time Facial Expression Recognition "In The Wild" by Disentangling 3D Expression from Identity (Supplementary Material)

Anonymous FG2020 submission
Paper ID 8

I. 3D Face Reconstruction From Videos

In this section, we provide more details about the proposed approach of 3D face reconstruction from videos that we adopt to create pseudo-ground truth in the training set.

A. Combined Identity and Expression 3D Face Modelling

Following [6], we exploit the fact that the current state of the art in facial landmarking can achieve highly-reliable landmark localization and therefore fuse the landmarks information with high-quality 3D face models as the one described in Eq. (1) of the main paper to achieve robust and accurate 3D face reconstruction results. We assume that the camera performs scaled orthographic projection (SOP) and that the identity parameters $\mathbf{i}$ are fixed (but unknown) over all the frames, letting however the expression parameters $\mathbf{e}_f$ as well as the camera parameters (scale and 3D pose) to vary from frame to frame. In brief, we minimize a cost function that consists of three terms: **a)** a sum of squared 2D landmark reprojection errors over all frames, **b)** a shape priors terms that imposes a quadratic prior over the identity and per-frame expression parameters, and **c)** a temporal smoothness term that enforces temporal smoothness of the expression parameters by using a quadratic penalty of the second temporal derivatives of the expression vector. In addition, to deal with outliers (e.g. frames with strong occlusions that cause gross errors in the landmarks), we also impose box constraints on the identity and per-frame expression parameters. Assuming that the camera parameters have been estimated in an initialization stage, the minimization of the cost function results in a large-scale least squares problem with box constraints, which we solve efficiently by using the reflective Newton method of [5].

B. Initialization Stage of Estimating Camera Parameters

To estimate the camera parameters during initialization, we assume that the shape to be recovered remains rigid over the whole video. As explained in the main paper, this is a simplistic yet effective assumption. Given this assumption, we seek to estimate the identity $\mathbf{i}$ and expression $\mathbf{e}$ parameters of the rigid facial shape as well as the per-frame camera parameters expressed as the SOP camera projection matrix $\Pi_f \in \mathbb{R}^{2\times 3}$ for every frame f (Π_f corresponds to the first 2 rows of the rotation matrix multiplied with the scale parameter). This estimation is implemented by solely considering and minimising the reprojection error term of the overall cost function described in Sec. I-A, which is the only term that depends on the camera parameters. This minimisation can be written as:

$$\begin{aligned}&\text{minimise } E_{land}(\Pi_1,..,\Pi_{n_f};\mathbf{i},\mathbf{e}) = \\ &\sum_{f=1}^{F}\sum_{j=1}^{L}\left\|\Pi_f\left(\bar{\mathbf{x}}^{(l_j)}+\mathbf{U}_{id}^{(l_j)}\,\mathbf{i}+\mathbf{U}_{exp}^{(l_j)}\,\mathbf{e}\right)-\ell_{j,f}\right\|^2\end{aligned} \quad (1)$$

where $\ell_{j,f}\in\mathbb{R}^2$ is the 2D location of the j-th facial landmark ($j=1,..,L$) in the f-th frame of the input video ($f=1,..,F$). In addition, $\bar{\mathbf{x}}^{(l_j)}\in\mathbb{R}^3$, $\mathbf{U}_{id}^{(l_j)}\in\mathbb{R}^{3\times n_i}$ and $\mathbf{U}_{exp}^{(l_j)}\in\mathbb{R}^{3\times n_e}$ are the 3 rows of $\bar{\mathbf{x}}$, $\mathbf{U}_{id}$ and $\mathbf{U}_{exp}$ respectively that correspond to the x, y and z coordinates of the vertex of the dense facial shape with index l_j, which is associated with the j-th landmark. In addition, as in the main 3D face reconstruction, we impose **box constraints** on the shape parameters $\mathbf{i}$ and $\mathbf{e}$.

We solve the above problem by adopting an alternating minimisation approach, with respect to shape parameters $\{\mathbf{i},\mathbf{e}\}$ and camera parameters We initialise the alternation by setting $\mathbf{i}$ and $\mathbf{e}$ to zero vectors, which corresponds to the mean shape $\bar{\mathbf{x}}$. Then we alternate

1) Keeping the shape parameters $\{\mathbf{i},\mathbf{e}\}$ fixed, we update the camera parameters $\{\Pi_1,..,\Pi_{n_f}\}$ by minimising E_{land} (1) with respect to $\{\Pi_1,..,\Pi_{n_f}\}$. This minimisation is decoupled for every frame and is approximated by using the extended POS approach of Bas et al. [1]
2) Keeping the camera parameters $\{\Pi_1,..,\Pi_{n_f}\}$ fixed, we update the shape parameters $\{\mathbf{i},\mathbf{e}\}$ by minimising E_{land} (1) with respect to $\{\mathbf{i},\mathbf{e}\}$ under the imposed box constraints. This is again a least squares optimisation with box constraints, which we solve efficiently by using the reflective Newton method of [5].

We have empirically observed that it is sufficient to apply only a few number of the above alternation iterations (e.g. 5), since after that we achieve convergence with negligible updates on the estimated parameters. As the final processing for this initialisation step, the sequence of estimated camera parameters (in the form of scale, rotation angles and translation parameters) is temporally smoothed using cubic smoothing splines. Please note that the estimation of the rigid shape parameters $\{\mathbf{i},\mathbf{e}\}$ in this initialisation stage plays only an auxiliary role, to facilitate the estimation of the camera parameters, which is the main goal and the final output of this stage.

II. Ground Truth Creation from a Large-scale Videos Dataset

In this section, we provide more details about the ground truth creation from a large-scale video dataset and more specifically about the video collection process and the automatic pruning process. In addition, we visualise some examples of generated pseudo-ground truth samples .

A. Details on the video collection process

We created our video collection by taking about 1,500 videos from the $2MF^2$ dataset [4] and adding about 4,500 more videos by downloading and cropping public videos from YouTube[1]. The videos have at least one person acting in front of a single monocular camera and none of them was shot from more than one camera at the same time. They have faces of variable resolutions and were taken under different conditions and set-ups, TV interviews, vlogs, movies, advertisements, conferences, etc.

B. Automatic pruning

As mentioned in the main paper, we are based on the fact that under the adopted 3D face modelling, the estimated identity vector $\mathbf{i}$ of each video is assumed to be drawn from a multi-variate normal distribution with zero mean and variance given by the identity matrix. Therefore, $\|\mathbf{i}\|^2$ yields a squared Mahalanobis distance between the current identity shape and the mean identity (following a chi-square distribution). This is a measure of plausibility of the estimated identity vector under the assumed face model. We thus classify as outliers and automatically prune the videos that correspond to an estimated value of $\|\mathbf{i}\|^2$ above a threshold θ. We select θ so that $\|\mathbf{i}\|^2$ is expected to be less than θ with a very high probability (e.g. 99%), under the assumed multi-variate normal distribution of $\mathbf{i}$.

C. Examples of pseudo-ground truth samples

Some frames selected randomly from our dataset along with their generated 3D pseudo-ground truth can be visualised in Figure 1.

III. Experimental Results

This section presents more results and visualisations regarding the 'Experimental Results' section of the main paper.

A. Evaluation of our 3D Face Reconstruction from Videos for Creating Pseudo-Ground Truth

In this section, we provide details on the evaluation of the intermediate step of creating pseudo-ground truth for training by using our approach on 3D face reconstruction. In more detail, the aim of this experiment is to assess the ability of our designed 3D face reconstruction in producing high-quality estimations on in-the-wild videos, and consequently validate the quality of pseudo-ground truth annotations in **FaceVid** dataset. For that, we test our 3D reconstruction method on the 4DFAB dataset [3], which contains ground truth information, in the same way followed on the **FaceVid** dataset. Since the **FaceVid** dataset has in-the-wild videos and 4DFAB was taken under lab (controlled) conditions, we added noise to the extracted landmarks from the videos of [3] based on the mean of the Cumulative Error Distribution (CED) reported in [7]. Next, we compute the vertex-to-vertex distance between our produced 3D reconstructions and the ground-truth 3D scans of [3]. This distance (error) was normalised by the inter-ocular distance of the neutral face of each sequence. We then compare our results with: 1) a baseline approach following a linear shape model fitting proposed in [10], [9], and 2) ITW [2], a state-of-the-art 3D reconstruction method from in-the-wild images. Figure 3 visualises the CED curves obtained by our method, ITW [2] and a baseline [10], [9]. We observe that our 3D face reconstruction approach achieves a better performance than the compared methods, with a mean normalised vertex to vertex error of 0.06.

B. Evaluation of Estimation of 3D Facial Expression Parameters

Figure 2 depicts the cumulative distribution function (CDF) of the Mean Squared Error (MSE) resulting from comparison between obtained facial expression parameters using our trained CNN (**ExpressionNet**), ITW [2], and baseline [10], [9] on the test split of **FaceVid** dataset and the ground truth. Some of these results using the aforementioned three methods are shown in figure 4 when added to the mean face of the model, making the facial expression vector $\mathbf{e}$ as the only difference, see equation (1) in the main paper. Viewing all the 3D reconstructions closely, it can be noticed that our method generates more accurate visual results capturing the eyes opening/closure and eyebrows/mouth movements, unlike the ITW and baseline approaches which fail to do so and generate sometimes stiff results.

Facial expressions estimated by our trained **ExpressionNet** on RadFD[11], KDEF[13], RAF-DB[12], CFEE[8] can be viewed in figures 5 and 6. Inspecting those results in more details, it is noticeable that our proposed approach succeeds in capturing the facial expressions irrespective of the identity, which can be seen when paying close attention along each column. This largely helps the emotion classifier in robustly segregating the tested images based on their basic emotion, thanks to the salient facial expression features extracted by our **ExpressionNet**. However, due to the intensity of the posed emotion by the subject, it might be tricky, even for humans, to recognise the same emotion elicited by different subjects.

C. Testing the Overall Framework on Videos

To further inspect the performance of our framework, we 1) train our emotion classifier on all the facial expression features generated by our the **ExpressionNet** as a result of using all the four benchmarks jointly (RadFD[11], KDEF[13],

[1]We have only used the videos and not the facial landmarks that are provided by $2MF^2$, because we want to use the so-called 3D Aware 2D landmark configuration (3DA-2D) [6], which in contrast to the traditional 2D landmark configuration estimates the real projected positions of the 3D landmarks on the image plane: for example, it avoids placing landmarks of occluded facial regions on the visible face edge.

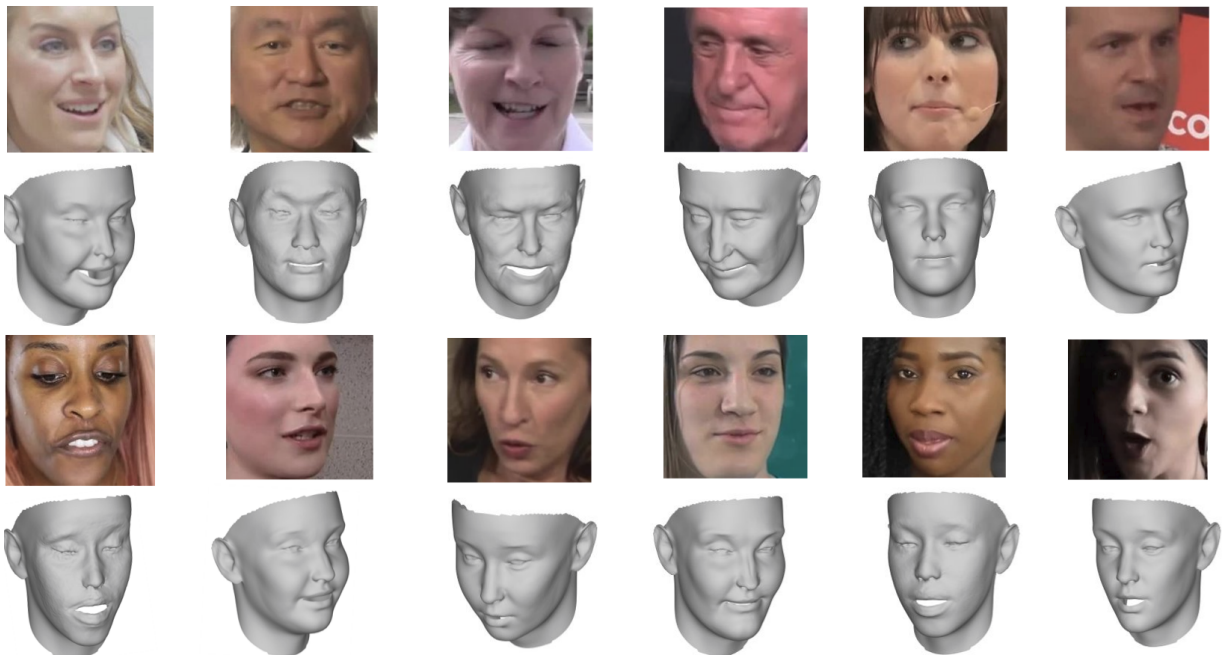

Fig. 1. Example frames from the collected dataset. Top row shows frames from different videos, while the generated 3D reconstructions are in the bottom row.

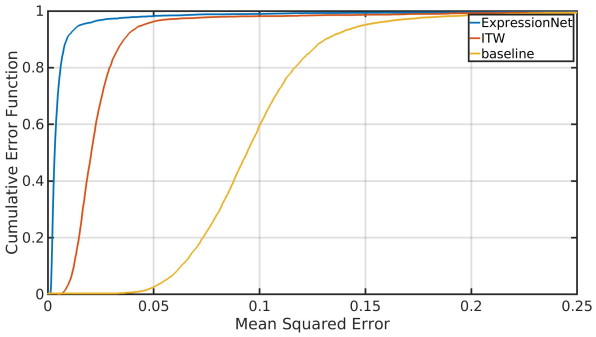

Fig. 2. Cumulative Error function of the mean squared error (MSE) between the facial expression coefficients estimation on the test split of **FaceVid** dataset by ExpressionNet, ITW [2], and baseline [10], [9] methods, and the ground truth. The average MSE values are 0.007, 0.026, 0.098 for the ExpressionNet, ITW, and baseline respectively.

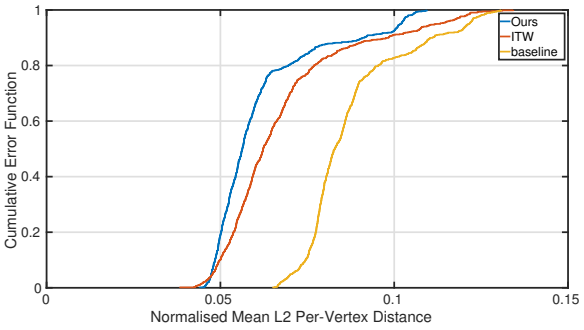

Fig. 3. Mean (over all vertices of every mesh) vertex-to-vertex error between the 3D reconstructions produced by our method, ITW and a baseline and the ground-truth 3D meshes of the 4DFAB dataset.

RAF-DB[12], CFEE[8]), 2) test the entire framework on new (previously unseen) video downloaded from Youtube after parsing it into frames and passing them one by one to our system. Please check out the resultant video in the supplementary material section of our submission. The generated result on this test video shows practically the very promising performance of our framework in the automatic recognition of facial emotions.

References

[1] 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 *ACCV*, pages 377–391. Springer, 2016.

[2] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Ploumpis, Y. Panagakis, and S. Zafeiriou. 3d reconstruction of "in-the-wild" faces in images and videos. *T-PAMI*, 2018.

[3] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou. 4dfab: A large scale 4D database for facial expression analysis and biometric applications. In *CVPR*.

[4] G. G. Chrysos, P. Favaro, and S. Zafeiriou. Motion deblurring of faces. *International Journal of Computer Vision*, pages 1–23, 2018.

[5] T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4):1040–1058, 1996.

[6] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018.

[7] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision*, 127(6-7):599–624, 2019.

[8] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[9] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

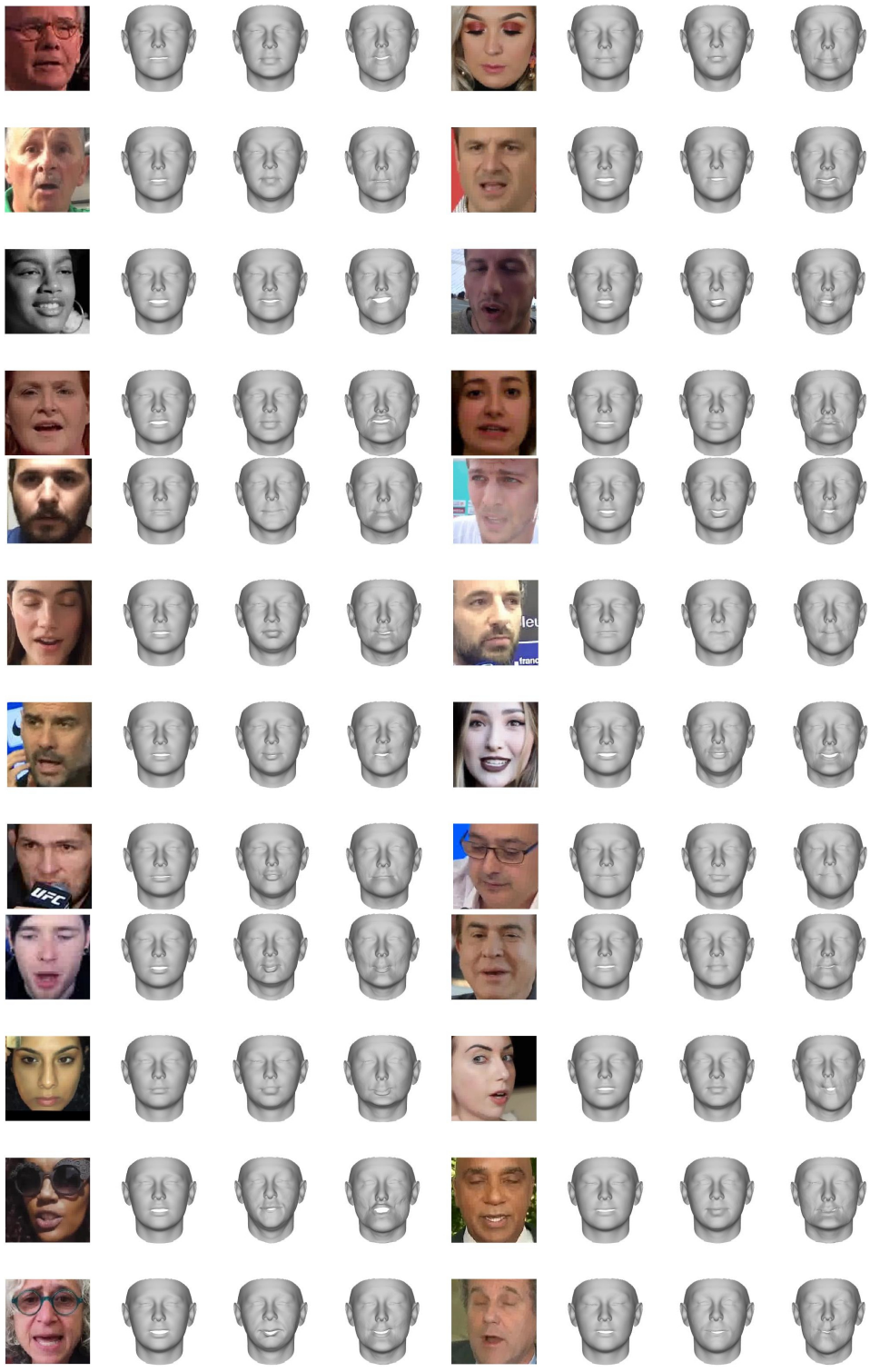

Fig. 4. Facial expression estimations generated by ExpressionNet, ITW [2], and baseline [10], [9] on some images selected randomly from the test split of **FaceVid**. The order of results visualisation is input image, ExpressionNet, ITW, baseline. All expressions were visualised on top of the mean face of the 3DMM.

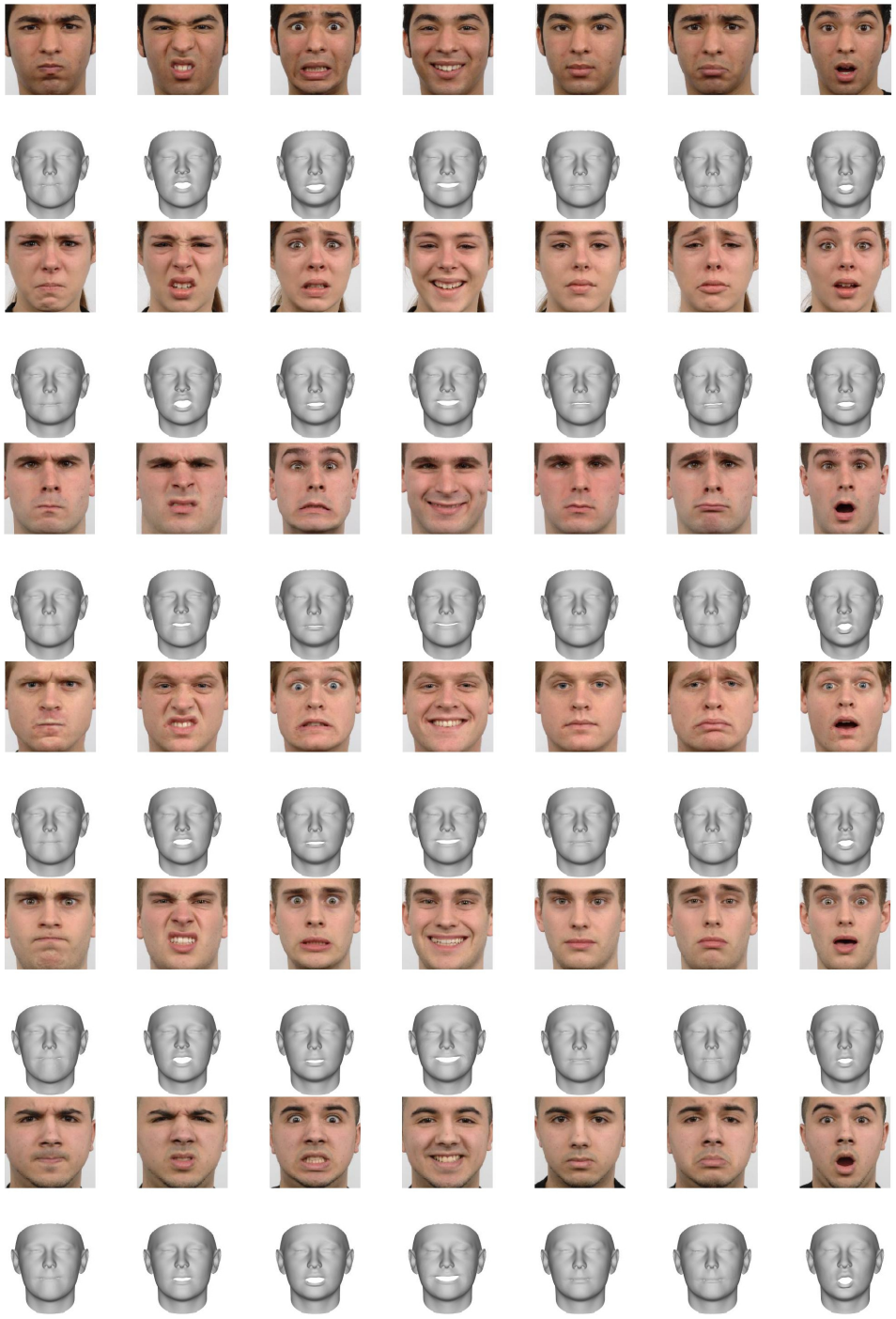

Fig. 5. Facial expression estimations generated by ExpressoinNet on some randomly selected images from the RadFD [11] dataset. All expressions were visualised on top of the mean face $\bar{\mathbf{x}}$ of the 3DMM. The order of visualised emotions from the left is angry, disgusted, fearful, happy, neutral, sad, surprised

[10] P. Huber, P. Kopp, W. Christmas, M. Rätsch, and J. Kittler. Real-time 3d face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441, 2017.

[11] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.

[12] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

[13] D. Lundqvist, A. Flykt, and A. Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630, 1998.

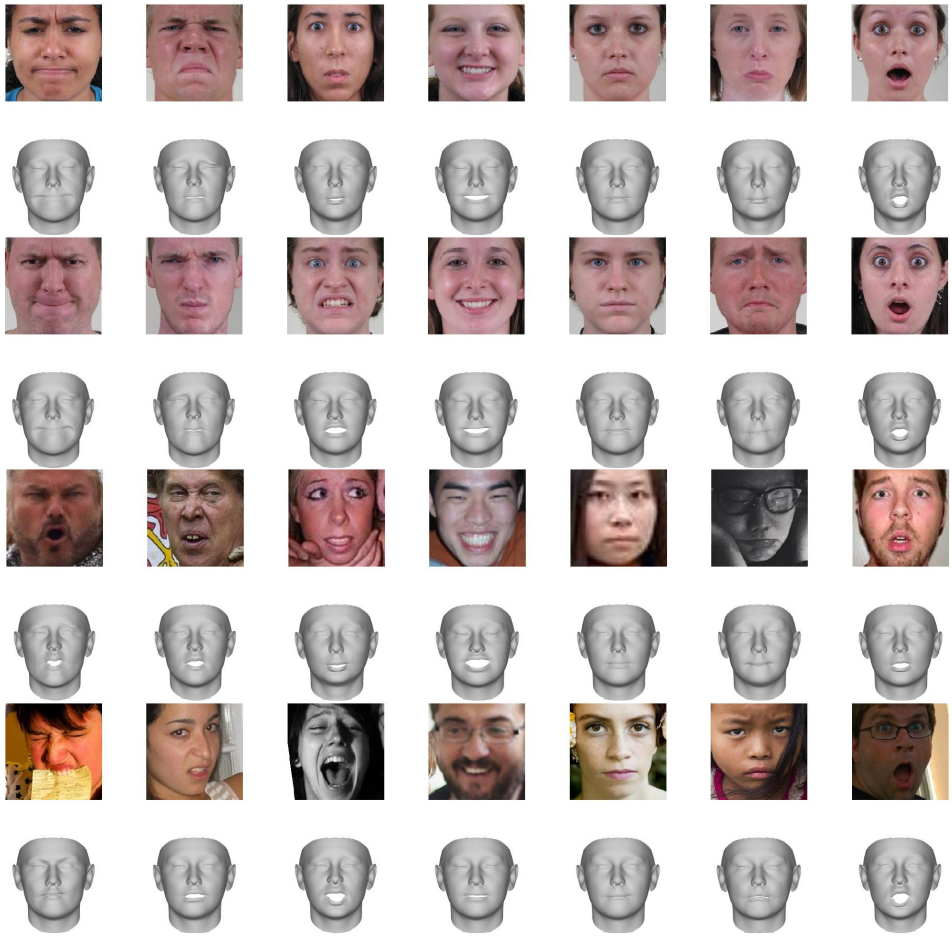

Fig. 6. Facial expression estimations generated by ExpressoinNet on some randomly selected images from the CFEE [11], and RAF-DB[12] datasets. All expressions were visualised on top of the mean face $\bar{\mathbf{x}}$ of the 3DMM. The order of visualised emotions from the left is angry, disgusted, fearful, happy, neutral, sad, surprised.