Robust Face Frontalization For Visual Speech Recognition*

Zhiqi Kang & Radu Horaud
Inria & Univ. Grenoble Alpes
Montbonnot Saint-Martin, France

Mostafa Sadeghi
Inria Nancy Grand-Est
Nancy, France

Abstract

Face frontalization consists of synthesizing a frontally-viewed face from an arbitrarily-viewed one. The main contribution of this paper is a robust frontalization method that preserves non-rigid facial deformations, i.e. expressions, to perform lip reading. The method iteratively estimates the rigid transformation (scale, rotation, and translation) and the non-rigid deformation between 3D landmarks extracted from an arbitrarily-viewed face, and 3D vertices parameterized by a deformable shape model. An important merit of the method is its ability to deal with large Gaussian and non-Gaussian errors in the data. For that purpose, we use the generalized Student-t distribution. The associated EM algorithm assigns a weight to each observed landmark, the higher the weight the more important the landmark, thus favoring landmarks that are only affected by rigid head movements. We propose to use the zero-mean normalized cross-correlation (ZNCC) score to evaluate the ability to preserve facial expressions. We show that the method, when incorporated into a deep lip-reading pipeline, considerably improves the word classification score on an in-the-wild benchmark.

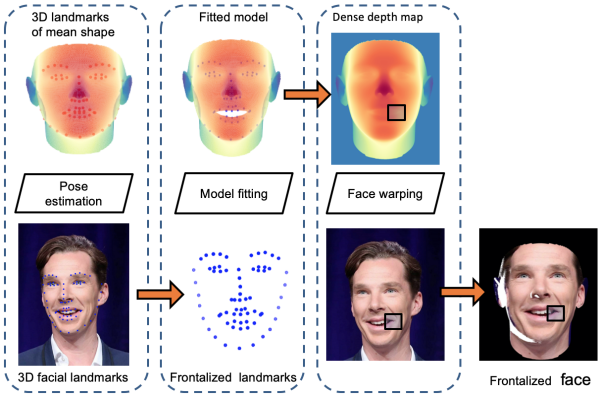


Figure 1: 3D landmarks extracted from a face (bottom-left) are aligned with 3D vertices associated with a frontal model (top-left). This deformable model is fitted to the frontalized landmarks (bottom-middle), yielding a deformed model aligned with the landmarks (top-middle). A dense depth map is computed by interpolating the 3D vertices of the triangulated mesh of the deformed model (top-right). This depth map is combined with the input face which is warped onto the frontal view (bottom-right).

1. Introduction

Face frontalization is the problem of synthesizing a frontal view of a face from an arbitrary view. Recent research has shown that face frontalization consistently boosts the performance of face recognition [53, 63, 4, 59, 60, 61]. A common feature of these methods is that they encourage *expression-free* face frontalization. In contrast, visual speech recognition (or lip reading) [17, 1, 35, 10] belongs to the larger class of facial expression analysis methods, e.g. [41], that require *expression-preserving* face frontalization. Indeed, lip and jaw motions are controlled by speech production as they are correlated with phonemes and with word pronunciation [45]. Nevertheless, lip- and jaw motion analysis is perturbed by inherent rigid head movements. It is therefore important to discriminate between rigid head movements and non-rigid facial/lip movements. Moreover, the capacity to properly analyse rigid and non-rigid movements separately plays a crucial role in audio-visual speech technologies, e.g. [14, 42, 52, 44, 48]. Visual speech is particularly useful when audio signals are corrupted by ambient noise, by competing speech sources, or by acoustic perturbations.

In this paper we address face frontalization as the problem of independently estimating rigid 3D pose and non-rigid 3D deformations of an arbitrarily-viewed face, Figure 1. In more detail, (i) we estimate the *scale, rotation, and translation* between a set of 3D facial landmarks extracted from an input face, and a set of 3D vertices associ-

*An extended version with supplemental materials can be found at https://team.inria.fr/robotlearn/rff-vsr/

ated with a mean statistical/deformable shape model, then (ii) we estimate the *statistical/deformation parameters* that best fit the shape model to the frontalized face. The one merit of the proposed method is its ability to deal both with landmark detection/localization noise and with large discrepancies between the facial landmarks and the associated model vertices. For the sake of robustness, we propose to use the generalized Student-t (gStudent) probability distribution function (pdf) – a heavy-tailed distribution that is able to deal both with Gaussian noise and with large errors in the data, by assigning a weight to each landmark-vertex pair [47, 19]. The data weights just mentioned measure the importance of each landmark-vertex pair and their role is to reduce the influence of those point pairs that are corrupted by large errors. The weights are treated as random variables whose realizations vary from zero to an arbitrarily large number (the higher the weight, the more important the point pair), hence they are an absolute measure of the relevance of each pair. This stays in contrast with the commonly used mixture model in point-set registration, namely a Gaussian mixture augmented with a uniform component, e.g. [37]. In the case of point-set registration the rationale is to classify as outliers those points in one set that don't have a match in the other set. The inlier/outlier discrimination is based on the posterior distribution to be an inlier, the values of these posteriors lying in the interval $[0, 1)$. Hence, and unlike the gStudent weights, the posteriors just mentioned are a relative measure of the importance of the point pairs.

We empirically evaluate the performance of the proposed method using the *zero-mean normalized cross-correlation* (ZNCC) coefficient between a frontalized face and the ground-truth frontal face. We also consider the task of isolated word recognition (IWR) based on lip movements. We embed the proposed face frontalization method into a state-of-the-art lip-reading deep model [35] and we show that IWR performance is improved by a considerable margin. These evaluations, based on direct comparison between predictions and ground-truth videos of faces, and on visual speech recognition, stand in contrast with evaluation based on face recognition, e.g. [53, 4, 59] that require as-neutral-as possible frontal faces.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes in detail the proposed frontalization method, analyses its performance and provides implementation details. Section 4 describes the proposed evaluation protocols and shows results obtained with our method and with several state-of-the-art methods. Section 5 draws some conclusions.

2. Related Work

Recently, a successful face frontalization approach has been to train deep neural networks (DNNs) to learn non-linear 2D-to-2D mappings between an arbitrary view and a frontal view. Some of the best performing DNN-based frontalization methods use CNN/GAN architectures, e.g. [54, 27, 49, 59, 57, 43, 58, 55], which outperform CNN-only ones, e.g. [53]. These methods necessitate large collections of input/output pairs of face images. For that purpose [57, 55, 58] use two datasets that contain multiple-camera recordings in a controlled setup, i.e. [20, 21]. [57] proposed to learn dense pixel-to-pixel correspondences between the two faces. Subsequently, [58] proposed a semi-supervised GAN-based method that augments the paired face images of [21] with unpaired in-the-wild faces with large variations in identity, e.g. [26]; their adversarial and identity-preserving losses enhance face recognition performance. [55] proposed a dual-attention GAN architecture that captures long-term dependencies in image space, thus providing a means to preserve identity. Another way to estimate the non-linear 2D-to-2D mapping between a profile image and a canonical frontal image of a face is to use a rectification network that learns local homographies between a deformed grid, that corresponds to a profile view, and a regular grid, that corresponds to a frontal view [60]. These DNN-based methods are designed to predict as-neutral-as-possible frontal faces, i.e. expression-free faces, in order to improve the performance of face recognition. The main drawback of these methods is that they cannot guarantee that they preserve rigidity – a necessary feature for preserving facial expressions.

Other methods estimate the pose of an input face with respect to a frontal 3D face model, then use the pose parameters to warp the facial pixels from the input image onto a frontal one. This amounts to pose estimation based on 2D-to-3D fitting, e.g. [63, 22, 18, 4]. [22] used 2D facial landmarks extracted from the input face that are in correspondence with 3D model landmarks. This corresponds to the estimation of intrinsic and extrinsic camera parameters, followed by image warping. Similar methods were proposed in [63] and [4] where 2D landmarks are extracted from the input face and associated with 3D landmarks of a generic 3D model. The pose estimators of [63, 22, 18, 4] lack a built-in mechanism capable of making them robust against the mixed presence of large errors in landmark localization and of non-rigid facial deformations. Moreover, [63] purposely removes expressions in order to favor identity features. Recently, [61] proposed to combine the 2D-to-3D geometric fitting method of [64] with a state-of-the-art GAN-based style transfer model to fill in the occluded regions caused by 3D face rotation. This method yields state-of-the-art results for the task of face recognition, but there is

no theoretical guarantee that non-rigid facial deformations are preserved by the GAN-based transfer process.

A fundamental building block of the proposed method is to embed face frontalization into a robust estimator. Robustness refers to the capacity of the estimator to be weakly affected by outliers. For that purpose, we cast the problem at hand in the framework of maximum likelihood estimation (MLE). In MLE, the choice of the pdf is crucial. As already mentioned above, we opt for the generalized Student t-distribution which belongs to the larger class of heavy-tailed distributions that are well known to be robust to outliers [36, 47, 19]. The associated EM procedure evaluates the posteriors of the weights – there is a weight associated with each point pair. The weights are treated as random variables drawn from a gamma distribution. The realizations of these variables vary from zero to an arbitrary high number. Therefore, they are an absolute measure of the importance of the data pairs.

In addition to estimating the weights, the GStudent EM algorithm estimates a posterior covariance matrix. The evaluation of a *full covariance* is fundamental for taking into account an *anisotropic error distribution* of the landmarks' 3D coordinates, and hence for assessing how much one can trust them. This stays in strong contrast with the traditional methods that are used in computer vision to estimate the rigid transformation between two point sets, e.g. [24, 25, 15, 3, 51]. These methods assume an *isotropic covariance – spherical shaped*, with the advantage of yielding closed-form solutions for estimating the rotation parameters. Nevertheless, it has been shown that these estimators are biased if there are measurement errors in both point sets [31], and unbiased estimators based on convex optimization and on branch-and-bound algorithms were proposed [39]. Nevertheless, [31, 39] used an isotropic covariance and did not incorporate a robust pdf that ultimately weights the absolute importance of the point pairs.The fact that GStudent estimates a full covariance matrix – *ellipsoidal shaped*, introduces an additional complexity, namely a non-linear solver is required to estimate the rotation matrix. In [23] this was addressed via convex relaxation. In this paper, we simplify the optimization problem by using quaternions and propose to use sequential quadratic programming [6].

3. Robust Face Frontalization

The core idea of the method, Figure 1, is to estimate the 3D pose (scale, rotation, and translation) and 3D shape of an input face and to warp it onto a frontal view. The pose is robustly estimated from a set of 3D landmarks, associated with a set of 3D vertices of a deformable geometric model, i.e. a triangulated mesh. This allows to rigidly frontalize the input landmarks, hence to preserve facial expression, i.e. (7), then to fit the deformable model to these landmarks, thus yielding a frontalized and deformed 3D model of the input face, i.e. (11). Next, vertex interpolation enables to compute a dense depth map of the frontalized face, i.e. (12) and (13). Finally, the input face is warped onto a frontal view using the dense depth map, i.e. (39) and (41) of Appendix E.

3.1. Robust Pose Estimation

Let I_p be an *observed* image of a face in an unknown pose. A set $\mathcal{X}$ of $J = 68$ 3D landmarks is extracted from I_p with image-centered coordinates $\boldsymbol{X}_{1:J} = \{\boldsymbol{X}_j\}_{j=1}^J \subset \mathbb{R}^3$. Throughout the paper we adopt the notation $\boldsymbol{X}_j = (X_{j1}, X_{j2}, X_{j3})$ to designate the three coordinates of a point in $\mathbb{R}^3$. Let $\boldsymbol{Z}_{1:J} = \{\boldsymbol{Z}_j\}_{j=1}^J \subset \mathbb{R}^3$ be the 3D coordinates of a set of vertices, $\mathcal{Z}$, that correspond to the frontal view of a 3D deformable geometric model with a neutral expression, and let I_f be the frontal image to be predicted. Pose estimation consists of finding the rigid transformation that best maps $\boldsymbol{X}_{1:J}$ onto $\boldsymbol{Z}_{1:J}$, namely

$$\boldsymbol{Z}_j = \rho \mathbf{R} \boldsymbol{X}_j + \boldsymbol{t} + \boldsymbol{e}_j, \quad \forall j \in \{1 \dots J\}, \tag{1}$$

where $\rho \in \mathbb{R}^+$, $\mathbf{R} \in \mathbb{R}^{3\times3}$ and $\boldsymbol{t} \in \mathbb{R}^3$ are the scale, rotation matrix and translation vector, respectively, associated with the unknown pose. Because the landmark locations $\boldsymbol{X}_{1:J}$ are inherently affected by detection errors as well as by *non-rigid facial deformations*, it is suitable to use a robust rigid-parameter estimation technique. For this purpose, we assume that the residuals $\boldsymbol{e}_{1:J}$ are samples of a random variable $\boldsymbol{e}$ drawn from a robust probability pdf, $P(\boldsymbol{e}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters. Then, the problem is cast into maximum likelihood estimation (MLE) or, equivalently into the minimization of the following negative log-likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}_{1:J}, \boldsymbol{Z}_{1:J}) = -\sum_{j=1}^{J} \log P(\boldsymbol{e}_j; \boldsymbol{\theta}), \tag{2}$$

The generalized Student-t distribution [19] writes:

$$P(\boldsymbol{e}; \boldsymbol{\theta}) = \int_0^\infty \mathcal{N}(\boldsymbol{e}; 0, w^{-1}\boldsymbol{\Sigma}) \mathcal{G}(w; \mu, \nu) dw, \tag{3}$$

where $\mathcal{N}(\cdot; 0, \boldsymbol{\Sigma})$ is the zero-mean normal distribution with covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$. The random latent variable $w \in \mathbb{R}^+$ is drawn from a gamma distribution $\mathcal{G}(\cdot; \mu, \nu)$, and it plays the role of a *precision*. Therefore, the variables $w_{1:J}$ (one for each data point) characterize the landmark-vertex pair: the higher the more reliable. The model's rigid and statistical parameters are gathered in the parameter vector $\boldsymbol{\theta} = \{\rho, \mathbf{R}, \boldsymbol{t}, \boldsymbol{\Sigma}, \mu, \nu\}$. Direct minimization of (2) is intractable. Expectation-maximization (EM) is therefore

adopted, namely the negative log-likelihood (2) is replaced with the *expected complete-data negative log-likelihood*:

$$\mathrm{E}_W[-\log P(\boldsymbol{X}_{1:J}, \boldsymbol{Z}_{1:J}, w_{1:J}|\boldsymbol{X}_{1:J}, \boldsymbol{Z}_{1:J}; \boldsymbol{\theta})]. \quad (4)$$

In practice, EM alternates between the estimation of the weight posteriors – the means $\overline{w}_{1:J}$ and the estimation of the parameters $\boldsymbol{\theta}$ via minimization of (4):

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{j=1}^{J} \left(\overline{w}_j \|\boldsymbol{Z}_j - \rho \mathbf{R}\boldsymbol{X}_j - \boldsymbol{t}\|^2_{\boldsymbol{\Sigma}} + \log|\boldsymbol{\Sigma}|\right). \quad (5)$$

The standard solution to (5) is to assume an isotropic covariance, $\boldsymbol{\Sigma} = \sigma \mathbf{I}_3$, which yields various closed-form solutions, e.g. [24, 51]. Indeed, after developing (5), one may verify that the term $\boldsymbol{X}'_j \mathbf{R} \boldsymbol{\Sigma}^{-1} \mathbf{R}^\top \boldsymbol{X}'^\top_j$ is proportional to $\boldsymbol{X}'_j \boldsymbol{X}'^\top_j$, where centered landmark coordinates were used. Nevertheless, the isotropic assumption cannot deal with anisotropic error distribution of the 3D landmark coordinates. In the case of a full covariance, the optimization of (5) with respect to the rotation parameters yields the following non-linear optimization problem:

$$\begin{cases} \min_{\boldsymbol{q}} & \operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}(\rho^2 \mathbf{R}(\boldsymbol{q})\mathbf{A}\mathbf{R}(\boldsymbol{q})^\top - 2s\mathbf{R}(\boldsymbol{q})\mathbf{B})\right) \\ \text{s.t.} & \boldsymbol{q}^\top \boldsymbol{q} = 1, \end{cases} \quad (6)$$

where $\boldsymbol{q}$ is a unit quaternion, thus reducing the number of parameters from nine to four. The six quadratic constraints and the quartic constraint needed to guarantee a rotation are replaced with one quadratic constraint, i.e. Appendix A.

While it iterates, Algorithm 1 (Appendix A) estimates (i) the posterior weight means, (ii) the pose parameters, (iii) the covariance, and (iv) the parameters of the gamma pdf. At convergence, the optimal parameters (denoted with a *) are applied to $\boldsymbol{X}_{1:J}$ in order to obtain *expression-preserving* frontalized landmarks, whose coordinates in I_f are denoted $\boldsymbol{Y}_{1:J} \subset \mathbb{R}^3$, namely:

$$\boldsymbol{Y}_j = \rho^* \mathbf{R}^* \boldsymbol{X}_j + \boldsymbol{t}^*, \quad \forall j \in \{1 \dots J\}. \quad (7)$$

3.2. Robust Deformable Shape Fitting

The next step is to fit a deformable 3D geometric model to these landmarks in order to obtain a *frontal dense depth map* of the face. For that purpose and without loss of generality, we consider a linear deformation model, e.g. the Basel Face Model (BFM) [40]. The latter consists of a 3D mesh with a set $\hat{\mathcal{V}}$ of N vertices, whose coordinates $\hat{\boldsymbol{V}}_{1:N} = \{\hat{\boldsymbol{V}}_n\}_{n=1}^N \subset \mathbb{R}^3$ are parameterized by a *statistical linear shape-model* in the following way, i.e. Appendix C:

$$\hat{\boldsymbol{V}}_n = \overline{\boldsymbol{V}}_n + \mathbf{W}_n \boldsymbol{s}, \quad \forall n \in \{1 \dots N\}, \quad (8)$$

where $\overline{\boldsymbol{V}}_{1:N} \subset \mathbb{R}^3$ are the vertices of a mean (neutral) shape, $\mathbf{W}_{1:N} \subset \mathbb{R}^{3\times K}$ are reconstruction matrices, and $\boldsymbol{s} \in \mathbb{R}^K$ is a low dimensional embedding of the vertex set, with $K \ll 3N$ (Appendix C). In order to fit this model to the frontalized landmarks $\boldsymbol{Y}_{1:J}$, we consider a subset of $J = 68$ ($\ll N$) vertices with coordinates $\hat{\boldsymbol{V}}_{1:J}$, associated one-to-one to the frontalized landmarks.[1] For that purpose, the statistical shape model must be deformed, such that the vertices $\hat{\boldsymbol{V}}_{1:J}$ are optimally aligned with the landmarks $\boldsymbol{Y}_{1:J}$. We thus obtain the residuals:

$$\boldsymbol{r}_j = \boldsymbol{Y}_j - \mathbf{Q}(\overline{\boldsymbol{V}}_j + \mathbf{W}_j \boldsymbol{s}) - \boldsymbol{d}, \quad \forall j \in \{1 \dots J\}, \quad (9)$$

where the rotation $\mathbf{Q}$ and the translation $\boldsymbol{d}$ align the image-centered and model-centered coordinate frames. As above, one can use the generalized Student-t distribution (3) and EM to robustly estimate the model parameters $\boldsymbol{s}$ and the weights $\overline{\pi}_{1:J}$. The shape parameters yield a closed-form expression, i.e. Appendix D:

$$\boldsymbol{s}^* = \left(\sum_{j=1}^{J} \overline{\pi}_j \mathbf{A}_j^\top \boldsymbol{\Gamma}^{-1} \mathbf{A}_j + \kappa \widetilde{\boldsymbol{\Lambda}}^{-1}\right)^{-1} \left(\sum_{j=1}^{J} \overline{\pi}_j \mathbf{A}_j^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{b}_j\right), \quad (10)$$

where $\boldsymbol{\Gamma}$ is a covariance, and with the notations $\mathbf{A}_j = \mathbf{Q}\mathbf{W}_j$ and $\boldsymbol{b}_j = \boldsymbol{Y}_j - \mathbf{Q}\overline{\boldsymbol{V}}_j - \boldsymbol{d}$; $\widetilde{\boldsymbol{\Lambda}}$ is a diagonal matrix containing the K principal eigenvalues of the shape embedding, and $\kappa \in \mathbb{R}^+$. All the N vertices of this *deformed shape* can now be mapped onto the frontal view, namely $\tilde{\boldsymbol{V}}_n = (\tilde{V}_{n1}, \tilde{V}_{n2}, \tilde{V}_{n3})$, with:

$$\tilde{\boldsymbol{V}}_n = \mathbf{Q}^*(\overline{\boldsymbol{V}}_n + \mathbf{W}_n \boldsymbol{s}^*) + \boldsymbol{d}^*, \quad \forall n \in \{1 \dots N\}. \quad (11)$$

3.3. Synthesizing a Frontal Face Image

A frontal dense depth map is then computed in the following way. Remember that the shape vertices form a 3D triangulated mesh; therefore the projection of $\tilde{\boldsymbol{V}}_{1:N}$ onto the image plane I_f form a 2D triangulated mesh whose vertices have $(\tilde{V}_{n1}, \tilde{V}_{n2})^\top_{1:N}$ as image coordinates. Let n_1, n_2 and n_3 be the indexes of the vertices of a mesh triangle. We now compute the barycentric coordinates, $\{\alpha_1, \alpha_2, \alpha_3\} \subset \mathbb{R}^+$ of a pixel $(a_1 \ a_2)^\top \in \mathbb{N}^2$ that lies inside that triangle, i.e. $0 \leq \alpha_1, \alpha_2, \alpha_3 \leq 1$. These barycentric coordinates are estimated by solving the following system of linear equations:

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \alpha_1 \begin{pmatrix} \tilde{V}_{n_1 1} \\ \tilde{V}_{n_1 2} \end{pmatrix} + \alpha_2 \begin{pmatrix} \tilde{V}_{n_2 1} \\ \tilde{V}_{n_2 2} \end{pmatrix} + \alpha_3 \begin{pmatrix} \tilde{V}_{n_3 1} \\ \tilde{V}_{n_3 2} \end{pmatrix}. \quad (12)$$

with the constraint $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Once the barycentric coordinates are computed, the depth $A_3 \in \mathbb{R}$ associated with pixel $(a_1 \ a_2)^\top$ is computed by linear interpolation, namely:

$$A_3 = \alpha_1 \tilde{V}_{n_1 3} + \alpha_2 \tilde{V}_{n_2 3} + \alpha_3 \tilde{V}_{n_3 3}. \quad (13)$$

[1] By abuse of notation, we use $\{1 \dots J\} \subset \{1 \dots N\}$.

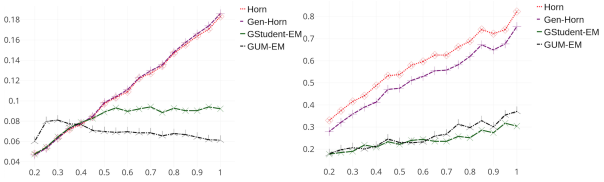
The above procedure is repeated for all the triangles and for all the points inside each triangle, thus obtaining a frontal dense depth map for each face pixel $\boldsymbol{A} = (a_1 \ a_2 \ A_3)^\top$. The final face frontalization step consists of synthesizing a frontal image of the input face. For this purpose we use a standard 3D shape visualization method, i.e. Appendix E.

3.4. Performance Analysis

In order to assess the performance of GStudent-EM, we compared it with the use of a Gaussian distribution and with a Gaussian-uniform mixture (GUM) distribution with a single Gaussian component [56]. The use of a uniform component in addition to Gaussian components in a mixture was initially proposed in [5]. In [37] it was proposed to be used in the context of point-set registration with the rationale of eliminating points in one set that don't have a match in the other set. One should note that in this paper we want to weight the relative importance of already registered point pairs and to reduce the influence of those point pairs that cannot be rigidly aligned.

In more detail, the likelihood $P(\boldsymbol{e};\boldsymbol{\theta})$ in (2) is replaced with: (i) a Gaussian distribution with isotropic covariance, i.e. [24], (ii) a Gaussian distribution with anisotropic (full) covariance, and (iii) a Gaussian-uniform distribution. The associated algorithms are referred to as Horn, Gen-Horn and GUM-EM, respectively. Starting with a frontal set of landmarks $\boldsymbol{Z}_{i:J}$ with image coordinates normalized in the interval $[0, 1]$, we randomly generated $P = 500$ poses (rotation, translation and scale). A landmark set $\boldsymbol{X}^p_{1:J}$ associated with a pose p is therefore simulated with $\boldsymbol{X}^p_j = s^p\mathbf{R}^p\boldsymbol{Z}_j + \boldsymbol{t}^p + \boldsymbol{e}^p_j(b)$, where $b > 0$ is a scalar that controls the level of noise and p is the trial (pose) index. b can be the variance associated with isotropic Gaussian noise $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, b\mathbf{I})$, the total variance associated with anisotropic Gaussian noise $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, or the volume associated with uniform noise $\boldsymbol{e} \sim \mathcal{U}(-b/2, b/2)^3$. For each trial we randomly split the samples into an equal number of inliers and outliers. Inliers are corrupted by anisotropic Gaussian noise with a fixed variance, while outliers are corrupted with either (i) uniform errors of increasing amplitude or (ii) anisotropic Gaussian errors of increasing total variance. The evaluation is based on the root mean square error (RMSE) between the estimated pose and the ground-truth pose, namely $\mathrm{RMSE}(\mathbf{R}) = 1/P(\sum_{p=1}^{P} \|\mathbf{R}^p - \tilde{\mathbf{R}}^p\|^2)^{1/2}$ as well as similar formulas for translation and scale. These RMSEs allow to directly compare robust and non-robust estimators. The RMSE curves are shown on Figure 2.

In the light of these experiments, one concludes that both GStudent-EM and GUM-EM yield robustness against outliers. GUM-EM performs better than GStudent-EM when the outliers are drawn from a uniform distribution, Figure 2(a): this can be easily explained since the simulated



(a) Uniform outliers. (b) Gaussian outliers.

Figure 2: RMSE errors in rotation, in the presence of outliers; inliers (50%) are affected by anisotropic Gaussian noise with total variance 0.0025, while outliers (50%) are affected by (a) uniform noise of increasing amplitude in the interval $[0.2, 1.0]$ or by (b) Gaussian noise with total variance varying in the interval $[0.2, 1.0]$, i.e. Appendix B, Fig. 4 and Fig. 5.

outliers follow the same statistical distribution as the one of the model. It is worth to be noticed that the posteriors estimated by GUM-EM are normalized between 0 and 1, hence their values are a relative measure of the quality of the point pairs. In contrast, the data weights estimated with GStudent-EM are random variables whose realizations can vary from zero to a very high number (higher the better). Therefore, the weight values are an absolute measure of the relevance of data points. While GUM-EM yields very good results, GStudent-EM offers better theoretical guarantees and a better way to weight the data; its behavior is easier to predict because all the parameters are estimated by the EM algorithm.

3.5. Implementation Details

The proposed method starts with 3D facial landmark extraction, or 3D face alignment (3DFA). Recently there has been a flourishing literature on this topic, yielding several DNN models and associated software packages, e.g. [7, 62, 8, 16, 13, 64, 30, 50, 38]. Moreover, these models were thoroughly trained, validated, tested and benchmarked using very large datasets that are automatically annotated, thus limiting the inherent errors associated with manual intervention, e.g. [29, 28, 64, 12]. We thoroughly analysed and benchmarked four publicly available software packages associated with four 3DFA methods [8, 16, 64, 50]. We could not find a significant difference in performance between these four 3DFA methods. In practice, the results reported below were obtained with the method of [8].

All the computations inside Algorithm 1 and Algorithm 2 are in closed-form, with the notable exception of the estimation of the rotation. The minimization of (6) is carried out using the sequential least squares programming

(SLSQP) solver of *scipy.optimize*,[2] in combination with a root-finding software package [32]. The SLSQP minimizer found at the previous EM iteration is used to initialize the current EM iteration. At the start of EM, the closed-form method of [24] is used to initialize the rotation.

The proposed method also requires a deformable shape model (Appendix C). We consider a set $\mathcal{S}^I = \{\boldsymbol{S}_1^I, \dots, \boldsymbol{S}_m^I, \dots, \boldsymbol{S}_M^I\} \subset \mathbb{R}^{3N}$ of M shapes, where $\boldsymbol{S}_m^I$ is a concatenation of N 3D coordinates that corresponds to face identity m, and where each face was scanned in a frontal view and with a neutral expression. We also consider a set $\mathcal{S}^E = \{\boldsymbol{S}_1^E, \dots, \boldsymbol{S}_m^E, \dots, \boldsymbol{S}_M^E\} \subset \mathbb{R}^{3N}$ that contains M scans with facial expressions, associated one-to-one, with the M identities. Let $\boldsymbol{S}_m^\triangle = \boldsymbol{S}_m^E - \boldsymbol{S}_m^I$ be the expressive-neutral difference of face identity m, namely the expressive offset. A face $\boldsymbol{S}$ can be reconstructed from its identity and expression embeddings (Appendix C):

$$\hat{\boldsymbol{S}} = \overline{\boldsymbol{S}}^I + \overline{\boldsymbol{S}}^\triangle + \widetilde{\mathbf{U}}^I \boldsymbol{s}^I + \widetilde{\mathbf{U}}^\triangle \boldsymbol{s}^\triangle, \tag{14}$$

where $\overline{\boldsymbol{S}}^I$ and $\overline{\boldsymbol{S}}^\triangle$ are the means associated with the identity- and with the expressive-neutral difference, matrices $\widetilde{\mathbf{U}}^I$ and $\widetilde{\mathbf{U}}^\triangle$ contain the K principle eigenvectors, and $\boldsymbol{s}^I$ and $\boldsymbol{s}^\triangle$ are the corresponding embeddings. Note that we use the notation $\hat{\boldsymbol{S}}$ to emphasize that the reconstructed shape is an approximation of $\boldsymbol{S}$.

In practice we use the publicly available Basel Shape Model (BSM) [40] augmented with [9]. This provides registered face scans: $M = 200$ in a frontal view and with neutral expressions, corresponding to M different identities, and an additional set of M expressive scans of the same identities. A scan consists of a triangulated mesh with $N = 53490$ vertice. All the scans have the same number of vertices that are registered. The dimension of the embedding is $K = 200$, hence $K \ll 3N$. The landmarks $\boldsymbol{Z}_{1:J}$ correspond to the vertices of the mean identity $\overline{\boldsymbol{S}}^I$.

The processing time for a 256×256 face image is of 1.11 seconds on an Intel(R), Xeon(R) W-2145, 3.70GHz CPU equipped with a Quadro RTX 4000 GPU. This time decomposes as follows: 3D landmark extraction (0.48 s), pose estimation (0.02 s), model fitting (0.23 s), depth map interpolation and face warping (0.38 s).

4. Experiments

We now evaluate the performance of face frontalization for the task of lip reading. The evaluation is twofold. First, we use the OuluVS2 dataset [2] that contains pairs of frontal and profile videos of speaking participants for a large number of subjects. The evaluation consists of computing a metric between an image obtained by face frontalization of a profile view of a speaker, with an image containing a frontally-viewed face of the same speaker. It is important that the profile and frontal images are recorded with synchronized cameras in order to capture the same expression. Consequently, the proposed evaluation is based on image-to-image comparison. Several metrics were developed in the past for comparing two images, e.g. feature-based and pixel-based metrics. In this work we use the *zero-mean normalized cross correlation* (ZNCC) coefficient between two image regions, a measure that has successfully been used for stereo matching, e.g. [46]. ZNCC is invariant to differences in brightness and contrast between the two images, due to the normalization with respect to mean and standard deviation. Second we use a lip-reading network in conjunction with a dataset that contains short videos of speakers that utter a single word, together with the ground-truth annotations (word labels). We devise an experimental protocol that measures the effect of face frontalization on the word classification score.

Let $R_f(h, v) \subset I_f$ be a region of size $H \times V$ whose center coincides with pixel location (h, v) of a frontalized image I_f. Similarly, let $R_t(h, v) \subset I_t$ be a region of the same size and whose center coincides with pixel location (h, v) of a ground-truth image I_t. The ZNCC coefficient between these two regions writes:

$$\text{ZNCC}(h, v, \delta h', \delta v') = \max_{\delta h, \delta v} \left\{ \frac{\text{Cov}\left[R_f(h, v), R_t(h + \delta h, v + \delta v)\right]}{\text{Var}\left[R_f(h, v)\right]^{1/2} \text{Var}\left[R_t(h + \delta h, v + \delta v)\right]^{1/2}} \right\}, \tag{15}$$

where $\text{Cov}\left[\cdot, \cdot\right]$ is the centered covariance between the two regions, $\text{Var}\left[\cdot\right]$ is the centered variance of a region, δh and δv are horizontal and vertical shifts, and $\delta h'$ and $\delta v'$ are the horizontal and vertical shifts that maximize the ZNCC coefficient. ZNCC lies in the interval $[0, 1]$.

The OuluVS2 dataset [2] targets the understanding of visual speech perception – the analysis of non-rigid lip motions that are associated with speech production. The dataset was recorded in an office with ordinary (artificial and natural) lighting conditions. The recording setup consists of five synchronized cameras (2 MP, 30 FPS) placed in different points of view: $0°$, $30°$, $45°$, $60°$, $90°$. The dataset contains 5×780 videos recorded with 53 participants. Each participant was instructed to read loudly several text sequences displayed on a computer monitor placed slightly to the left and behind the $0°$ (frontal) camera. While participants were asked to keep the head still, natural uncontrolled head movements and body position changes were inevitable. As a consequence the actual head pose varies from one participant to another and there is no exact match between the head and the camera orientations.

[2] https://docs.scipy.org/doc/scipy/reference/optimize.html

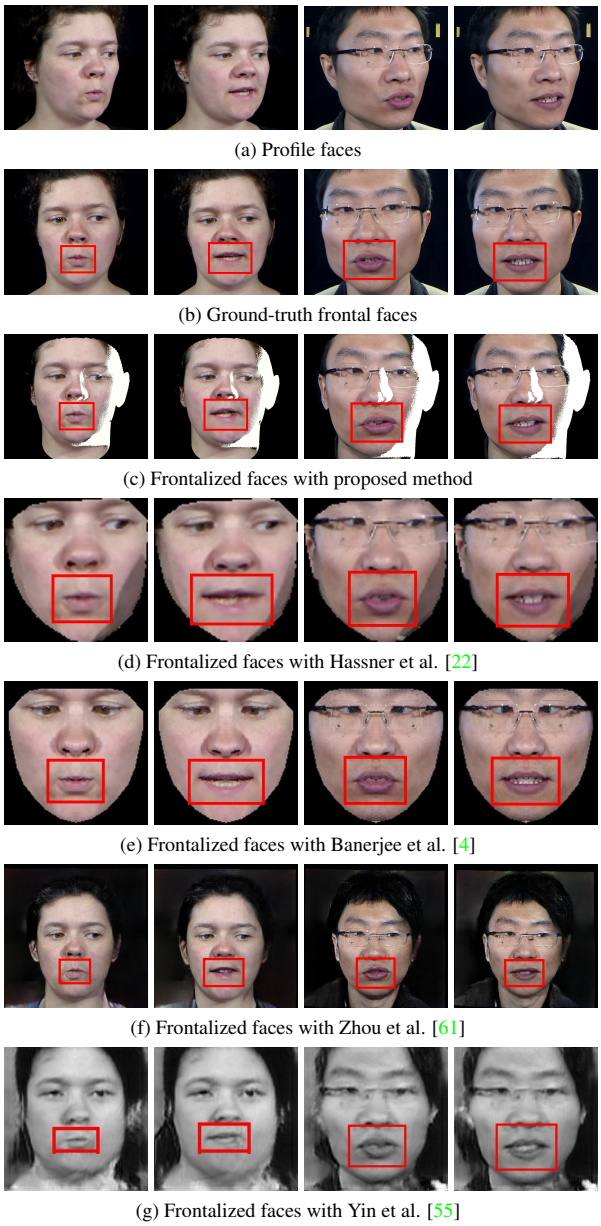
Method	Principle	ZNCC
Hassner et al. [22]	2D-to-3D fitting+symmetry	0.780
Banerjee et al. [4]	2D-to-3D fitting+symmetry	0.739
Zhou et al. [61]	2D-to-3D fitting + GAN	0.801
Yin et al. [55]	2D-to-2D GAN	0.773
Proposed	3D-to-3D robust fitting	**0.841**

Table 1: Mean ZNCC coefficients for 15 participants from the OuluVS2 dataset. ZNCC lies in the interval $[0, 1]$.

Part.	Yaw	[22]	[4]	[61]	[55]	Prop.
#31	19.1	***0.905***	0.856	0.822	0.875	**0.927**
#01	23.5	***0.915***	0.893	0.884	**0.921**	0.909
#02	24.9	0.888	0.878	***0.929***	0.881	**0.956**
#10	29.0	0.805	***0.812***	**0.873**	0.792	***0.812***
#23	30.0	0.810	**0.857**	0.819	0.817	***0.847***
#27	32.9	0.685	**0.852**	***0.824***	0.772	0.787
#19	37.8	***0.752***	0.650	0.662	0.677	**0.755**
#12	38.5	0.731	0.713	***0.755***	0.683	**0.770**
#21	40.6	0.632	***0.743***	0.653	0.673	**0.766**
Mean		0.791	0.801	***0.802***	0.787	**0.836**

Table 2: ZNCC scores for nine participants as a function of estimated yaw angle (in degrees) that corresponds to the horizontal head orientation computed with the proposed 3D head-pose estimator. For each participant, the best scores are in **bold** and the second best are in ***slanted bold***.

We compared the proposed methods with four state-of-the-art methods for which the code is publicly available, [22, 4, 61, 55]. We applied the frontalization to images extracted from the videos recorded with the 30° camera (I_p) and compared the results with the "ground-truth", namely the corresponding images extracted from the videos recorded with the 0° camera (I_t). Notice that videos recorded with higher viewing angles, i.e. 45°, 60° and 90°, can be hardly exploited for lip reading. For each frontalized image I_f we extract the mouth region R_f and we search in the associated ground-truth image I_t for the best-matching region R_t. This provides a ZNCC coefficient (15) for each query image I_p. Notice that (15) only cares about the horizontal and vertical shifts in the image plane and assumes that the frontalized face and the corresponding ground-truth frontal face share the same scale. In practice, different frontalization algorithms output faces at different scales. For this reason and for the sake of fairness, prior to applying (15), we extract facial landmarks from both the frontalized and ground-truth faces and we use a subset of this set of landmarks to estimate the scale factor between the two faces.



(a) Profile faces

(b) Ground-truth frontal faces

(c) Frontalized faces with proposed method

(d) Frontalized faces with Hassner et al. [22]

(e) Frontalized faces with Banerjee et al. [4]

(f) Frontalized faces with Zhou et al. [61]

(g) Frontalized faces with Yin et al. [55]

Figure 3: Face frontalization examples for participants #02 and #01 from the OuluVS2 dataset.

We randomly selected 30 video pairs, recorded with the 30° and 0° cameras, respectively, associated with 15 participants from the OuluVS2 dataset. Each video contains 160 images, hence there are $30 \times 160 = 4800$ image pairs in our benchmark. The mean ZNCC coefficients obtained with four state-of-the-art methods and with the proposed method are displayed in Table 1.

We noticed that there were important discrepancies in

Training \ Testing	[22]	[61]	[55]	[33]	Prop.
Pre-trained [33]	60	59	20	87	82
Fine-tuned [33]	64	66	24	88	85

Table 3: Word classification scores (WCSs) in %. *First row:* Pre-trained lip reading model [33]. *Second row:* Fine tuned lip reading model, where our frontalization replaces the built-in frontalization [35]. The pre-trained and fine-tuned models were tested using several frontalization methods.

method performance across participants, and this for all tested methods. In order to better understand this phenomenon, we computed the mean ZNCC coefficient for nine participants and displayed these means as a function of the yaw angle (in degrees), i.e. the horizontal head orientation estimated with the proposed method, Table 2. One may notice that there is a wide range of yaw angles, from $19°$ to $40°$, and that the performance gracefully decreases as the yaw angle increases. The proposed method yields results that are more consistent than the other methods, as the yaw angle increases. Nevertheless, the proposed method yields the best scores for participants #19 and #12 and the second best score for participant #21, and this is without using mirror-symmetric information or inpainting. Examples of face frontalization obtained with our method and with four other methods, [22, 4, 61, 55], are shown on Figure 3: (a) input mages recorded with the $30°$ camera, (b) ground-truth images recorded with the $0°$ camera, (c)-(g) frontalization results. The ZNCC correlation scores correspond the mouth region, shown in red.

We also evaluated the ability of our method to improve the performance of lip reading and we compared it with other frontalizers. For this purpose, we considered the isolated word recognition (IWR) task. The LRW (lip reading in the wild) dataset [11] consists of $500,000$ videos of 500 English words uttered by $1,000$ different speakers. Each video is 29 frame long and each target word is surrounded by context words. There are large inter-speaker variations in head motion. The best performing method for this 500-IWR task is based on the temporal convolutional network (TCN) model of [35, 33, 34] which achieves a word classification score of 87%. This lip-reading model and its variants use the same built-in face frontalization for training, validation and test. This frontalizer estimates a 3D mapping between the input face and a generic face model, [35]. Unfortunately, the authors don't provide a detailed description of the frontalization method that they use.

We performed two sets of experiments. The first experiment uses the pre-trained model of [33] with its built-in frontalizer, as it is available online. The second experiment replaces their frontalization with ours and fine-tunes the lip reading model. For each one of the 500 words, we used 200 videos for training and 20 videos for validation. We tested the pre-trained and fine-tuned models with the same test set that consists of 20 videos for each one of the 500 word vocabulary. As for test, we replaced the built-in frontalization with ours as well as with the other frontalizations under comparison. The results summarized in Table 3 show an increase in performance when our frontalization method is used to fine tune the model, instead of the one used in [11].

5. Conclusions

In this paper we proposed a face frontalization method that preserves non-rigid facial deformations. This stays in contrast with several state-of-the-art frontalization methods that are designed to boost the performance of face recognition by predicting as-neutral-as-possible frontal faces. We conducted a series of experiments in order to analyze the effect of frontalization on the task of visual speech recognition, whose success heavily relies on the analysis of non-rigid mouth motions, i.e. lip reading. For this purpose, we used two datasets.

We proposed an evaluation pipeline that consists of measuring the ZNCC score between a frontalized face and a frontal view of the same face. We compared our method with four state-of-the-art methods that use various geometric and DNN models. This benchmark reveals that the proposed method better preserves the shape of the mouth by a significant margin, and this without making recourse to facial symmetry or to DNN-based inpainting techniques to fill in the occluded areas.

The LRW dataset contains videos of persons uttering speech. Unlike the OuluVS2 participants who keep their heads still, the LRW participants perform large and unexpected head motions. We plugged our frontalization model into a DNN-based lip reading model and we thoroughly analyzed its effect on the word classification score. These experiments empirically demonstrate that the proposed robust frontalization improves these scores significantly, Table 3.

As already outlined, face frontalization may well be viewed as a process of discriminating between rigid head movements and non-rigid facial deformations, hence it can be used to eliminate head motions that naturally accompany speech production. The preliminary lip reading experiments described above are clear evidence that expression-preserving frontalization boosts the performance of visually-augmented speech technologies.

References

[1] Ahsan Adeel, Mandar Gogate, Amir Hussain, and William M Whitmer. Lip-reading driven deep learning approach for speech enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2019. 1

[2] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis. In *International Conference on Automatic Face and Gesture Recognition*, volume 1, pages 1–5. IEEE, 2015. 6

[3] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, September 1987. 3

[4] Sandipan Banerjee, Joel Brogan, Janez Krizaj, Aparna Bharati, Brandon Richard Webster, Vitomir Struc, Patrick J Flynn, and Walter J Scheirer. To frontalize or not to frontalize: Do we really need elaborate pre-processing to improve face recognition? In *IEEE Winter Conference on Applications of Computer Vision*, pages 20–29, 2018. 1, 2, 7, 8

[5] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993. 5

[6] Joseph-Frédéric Bonnans, Jean Charles Gilbert, Claude Lemaréchal, and Claudia A Sagastizábal. *Numerical optimization: theoretical and practical aspects*. Springer Science & Business Media, 2006. 3, 2

[7] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAW) challenge. In *European Conference on Computer Vision Workshops*, pages 616–624. Springer, 2016. 5

[8] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *IEEE International Conference on Computer Vision*, pages 1021–1030, 2017. 5

[9] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 6

[10] Shiyang Cheng, Pingchuan Ma, Georgios Tzimiropoulos, Stavros Petridis, Adrian Bulat, Jie Shen, and Maja Pantic. Towards pose-invariant lip-reading. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4357–4361, 2020. 1

[11] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103, 2016. 8

[12] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos 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. 5

[13] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zaferiou. Cascade multi-view hourglass model for robust 3d face alignment. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 399–403. IEEE, 2018. 5

[14] Stéphane Dupont and Juergen Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, 2000. 1

[15] Olivier D Faugeras and Martial Hebert. The representation, recognition, and locating of 3-d objects. *The International Journal of Robotics Research*, 5(3):27–52, 1986. 3

[16] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision*, pages 534–551, 2018. 5

[17] Adriana Fernandez-Lopez and Federico M Sukno. Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, 78:53–72, 2018. 1

[18] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Effective 3D based frontalization for unconstrained face recognition. In *International Conference on Pattern Recognition*, pages 1047–1052. IEEE, 2016. 2

[19] Florence Forbes and Darren Wraith. A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing*, 24(6):971–984, 2014. 2, 3, 1

[20] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2007. 2

[21] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. 2

[22] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015. 2, 7, 8

[23] Radu Horaud, Florence Forbes, Manuel Yguel, Guillaume Dewaele, and Jian Zhang. Rigid and articulated point registration with expectation conditional maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):587–602, 2010. 3

[24] Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987. 3, 4, 5, 6, 2

[25] Berthold KP Horn, Hugh M Hilden, and Shahriar Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society of America A*, 5(7):1127–1135, 1988. 3

[26] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2008. 2

[27] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis. In *IEEE International Conference on Computer Vision*, pages 2439–2448, 2017. 2

[28] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3D face alignment from 2D video for real-time use. *Image and Vision Computing*, 58:13–24, 2017. 5

[29] László A Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, and Jeffrey F Cohn. The first 3D face alignment in the wild (3DFAW) challenge. In *European Conference on Computer Vision*, pages 511–520. Springer, 2016. 5

[30] Lei Jiang, Xiao-Jun Wu, and Josef Kittler. Dual attention mobdensenet (damdnet) for robust 3D face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 5

[31] Ken-ichi Kanatani. Unbiased estimation and statistical analysis of 3-d rigid motion from two views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(1):37–50, 1993. 3

[32] Dieter Kraft. A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, DLR German Aerospace Center – Institute for Flight Mechanics, Koln, Germany, 1988. 6

[33] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic. Towards practical lipreading with distilled and efficient models. *arXiv preprint arXiv:2007.06504*, 2020. 8

[34] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic. Lipreading with densely connected temporal convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2021. 8

[35] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6319–6323, 2020. 1, 2, 8

[36] G.J. McLachlan and D. Peel. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000. 3

[37] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, 2010. 2, 5

[38] Xin Ning, Pengfei Duan, Weijun Li, and Shaolin Zhang. Real-time 3d face alignment using an encoder-decoder network with an efficient deconvolution layer. *IEEE Signal Processing Letters*, 27:1944–1948, 2020. 5

[39] Carl Olsson, Fredrik Kahl, and Magnus Oskarsson. The registration problem revisited: Optimal solutions from points, lines and planes. In *Computer Vision and Pattern Recognition*, volume 1, pages 1206–1213. IEEE, 2006. 3

[40] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 4, 6

[41] Ercheng Pei, Meshia Cedric Oveneke, Yong Zhao, Dongmei Jiang, and Hichem Sahli. Monocular 3d facial expression features for continuous affect recognition. *IEEE Transactions on Multimedia*, 2020. 1

[42] Bertrand Rivet, Laurent Girin, and Christian Jutten. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):96–108, 2006. 1

[43] Changle Rong, Xingming Zhang, and Yubei Lin. Feature-improving generative adversarial network for face frontalization. *IEEE Access*, 8:68842–68851, 2020. 2

[44] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Audio-visual speech enhancement using conditional variational auto-encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1788–1800, 2020. 1

[45] Tanja Schultz, Michael Wand, Thomas Hueber, Dean J Krusienski, Christian Herff, and Jonathan S Brumberg. Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2257–2271, 2017. 1

[46] Changming Sun. Fast stereo matching using rectangular subregioning and 3D maximum-surface techniques. *International Journal of Computer Vision*, 47(1-3):99–117, 2002. 6

[47] J. Sun, A. Kabán, and J. M. Garibaldi. Robust mixture clustering using Pearson type VII distribution. *Pattern Recognition Letters*, 31(16):2447–2454, 2010. 2, 3

[48] Fei Tao and Carlos Busso. End-to-end audiovisual speech recognition system with multitask learning. *IEEE Transactions on Multimedia*, 2020. 1

[49] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017. 2

[50] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. 3d face reconstruction from a single image assisted by 2d face images in the wild. *IEEE Transactions on Multimedia*, 23:1160–1172, 2020. 5

[51] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, April 1991. 3, 4

[52] Pingping Wu, Hong Liu, Xiaofei Li, Ting Fan, and Xuewu Zhang. A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. *IEEE Transactions on Multimedia*, 18(3):326–338, 2016. 1

[53] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–684, 2015. 1, 2

[54] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3990–3999, 2017. 2

[55] Yu Yin, Songyao Jiang, Joseph P Robinson, and Yun Fu. Dual-attention GAN for large-pose face frontalization. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 24–31. IEEE Computer Society, 2020. 2, 7, 8

[56] Andrei Zaharescu and Radu Horaud. Robust factorization methods using a Gaussian/uniform mixture model. *International Journal of Computer Vision*, 81(3):240, 2009. 5

[57] Zhihong Zhang, Xu Chen, Beizhan Wang, Guosheng Hu, Wangmeng Zuo, and Edwin R Hancock. Face frontalization

using an appearance-flow-based convolutional neural network. *IEEE Transactions on Image Processing*, 28(5):2187–2199, 2019. 2

[58] Zhihong Zhang, Ruiyang Liang, Xu Chen, Xuexin Xu, Guosheng Hu, Wangmeng Zuo, and Edwin R Hancock. Semi-supervised face frontalization in the wild. *IEEE Transactions on Information Forensics and Security*, 16:909–922, 2021. 2

[59] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2018. 1, 2

[60] Erjin Zhou, Zhimin Cao, and Jian Sun. Gridface: Face rectification via learning local homography transformations. In *Proceedings of the European Conference on Computer Vision*, 2018. 1, 2

[61] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5911–5920, 2020. 1, 2, 7, 8

[62] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: a 3D solution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 5

[63] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 1, 2

[64] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):78–92, 2019. 2, 5