A Generative Model for Depth-based Robust 3D Facial Pose Tracking

Lu Sheng[1] Jianfei Cai[2] Tat-Jen Cham[2] Vladimir Pavlovic[3] King Ngi Ngan[1]
[1]The Chinese University of Hong Kong [2]Nanyang Technological University [3]Rutgers University
{lsheng, knngan}@ee.cuhk.edu.hk, {asjfcai, astjcham}@ntu.edu.sg, vladimir@cs.rutgers.edu

Abstract

We consider the problem of depth-based robust 3D facial pose tracking under unconstrained scenarios with heavy occlusions and arbitrary facial expression variations. Unlike the previous depth-based discriminative or data-driven methods that require sophisticated training or manual intervention, we propose a generative framework that unifies pose tracking and face model adaptation on-the-fly. Particularly, we propose a statistical 3D face model that owns the flexibility to generate and predict the distribution and uncertainty underlying the face model. Moreover, unlike prior arts employing the ICP-based facial pose estimation, we propose a ray visibility constraint that regularizes the pose based on the face model's visibility against the input point cloud, which augments the robustness against the occlusions. The experimental results on Biwi and ICT-3DHP datasets reveal that the proposed framework is effective and outperforms the state-of-the-art depth-based methods.

1. Introduction

Robust 3D facial pose tracking is an important topic in the fields of computer vision and computer graphics, with applications in facial performance capture, human-computer interaction, immersive 3DTV and free-view TV, as well as virtual reality and augmented reality. Traditionally, the facial pose tracking has been successfully performed on RGB videos [22, 3, 16, 4, 14, 15, 21, 33, 42] for well-constrained scenes, challenges posed by illumination variations, shadows, and substantial occlusions hamper RGB-based facial pose tracking systems from being employed in more typical *unconstrained* scenarios. The utilization of depth data from commodity real-time range sensors has led to more robust 3D facial pose tracking, not only by enabling registration along the depth axis, but also by providing cues for the occlusion reasoning.

Although promising results have been demonstrated by leveraging both RGB and depth data in unconstrained facial pose tracking, existing approaches are not yet able to reliably cope when the RGB data is poor due to inconsistent or poor lighting conditions. Furthermore, RGB data may not always be available in scenarios when privacy is a major concern. Therefore, it is meaningful to study robust 3D facial pose tracking using depth data alone, as a complementary alternative to traditional tracking systems.

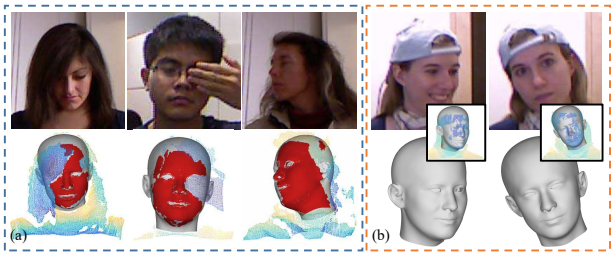


Figure 1. Our identity-adaptive facial pose tracking system is robust to occlusions and expression distortions. (a) Poses are estimated with heavy occlusions. The face models are overlaid with the input point clouds with the visible face points marked by red. (b) Poses are tracked under varying expressions. The estimated face identities are not interfered by the expressions.

Some of the new challenges that we want to address when solely tracking on depth data include: (1) coping with complex self-occlusions and other occlusions caused by hair, accessories, hands and *etc.*; (2) sustaining an always-on face tracker that can dynamically adapt to any user without manual calibration; and (3) providing stability over time to variations in user expressions. Unlike previous depth-based discriminative or data-driven methods [35, 7, 28, 18, 17, 30, 20] that require complex training or manual calibration, in this paper we propose a framework that unifies pose tracking and face model adaptation on-the-fly, offering highly accurate, occlusion-aware and uninterrupted 3D facial pose tracking, as shown in Fig. 1.

The contributions of this work are threefold. First, we introduce a novel analytical probabilistic formulation for a generative 3D face model, advancing the earlier 3D multilinear tensor model [12, 38] and encouraging groupwise pose estimation and expression-invariant face model updating. Second, we propose an occlusion-aware pose estimation mechanism based on minimizing an information-

theoretic *ray visibility score* that regularizes the visibility of the face model in the current depth frame. This is based on an underlying intuition that the visible face model points must either be co-located with the observed point cloud as visible points, or be located behind the point cloud as occluded points. Our pose estimation method does not need explicit correspondences to accurately estimate facial pose while handling occlusions well. Third, we present a tightly coupled online identity adaptation method which gradually adapts the face model to the captured user with sequential input depth frames. This is done by tracing the identity distribution during the tracking process in a generative process.

2. Related Work

With the popularity of the consumer-level depth sensors, apart from the RGB based facial pose tracking systems [22, 3, 16, 4, 14, 15, 21, 33, 42], a variety of 3D facial pose tracking and model personalization frameworks have been proposed. One category of approaches employed depth features, such as facial features defined by surface curvatures [35], nose detector [7], or triangular surface patch descriptors [28]. However, these methods may fail when such features cannot be detected under conditions of highly noisy depth data, extreme poses or large occlusions.

Another type of methods applied the discriminative methods based on the random forests [18, 17], the deep Hough network [30], or finding the dense correspondence field between the input depth image and a predefined canonical face model [20, 40]. Although these methods are promising and accurate, they require extensive and sophisticated supervised training with large scale datasets.

Another approach involves rigid and non-rigid registration of 3D face models to input depth images, either through the use of 3D morphable models [1, 11, 10, 9, 13, 29, 8, 6, 19, 26, 24, 34], or brute-force per-vertex 3D face reconstruction [37, 41, 23]. Although such systems may be accurate, most require offline initialization or user calibration to create face models specific to individual users. There are also some subsequent methods that gradually refine the 3D morphable model over time during active tracking [24, 26, 6, 19, 36]. Our proposed method falls into this category. Other than the existing multilinear face model [12, 38] that is discriminatively applied for tracking, we enhance this face model through a novel and complete probabilistic framework, in which the uncertainty due to expression variation is explicitly modeled while retaining user identity, thus increasing the stability of tracking.

A related problem is dealing with occlusion that arise during tracking. While occlusions may be discriminatively labeled through face segmentation [19, 32] or patch-based feature learning [17, 18, 30, 20], ICP-based face model registration frameworks do not handle correspondence ambiguities well when typical distance measures or normal vector compatibility criteria are used [31, 19, 41, 23]. Possible remedies include particle swarm optimization [27] for optimizing delicate objective functions [26]. Recently, Wang *et al*. [39] catered for partial registration of general moving subjects and handled occlusions better by considering multi-view visibility consistency. Our proposed ray visibility score incorporates a similar visibility constraint between the face model and the input point cloud but with a probabilistic formulation, which is able to more robustly handle uncertainties in the 3D face model, and is thus less vulnerable to local minima that are frequently encountered in ICP.

3. Probabilistic 3D Face Parameterization

In this section, we introduce the 3D face model with a probabilistic interpretation, which acts as an effective prior for facial pose estimation and face identity adaptation.

3.1. Multilinear Face Model

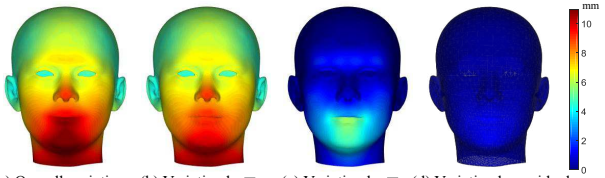
We apply the multilinear model [12, 38] to parametrically generate arbitrary 3D faces that are adaptive to different identities and expressions. It is controlled by a three dimensional tensor $\mathcal{C} \in \mathbb{R}^{3N_{\mathcal{M}} \times N_{\text{id}} \times N_{\text{exp}}}$ that each dimension corresponds to shape, identity and expression, respectively. The multilinear model represents a 3D face $\mathbf{f} = (x_1, y_1, z_1, \ldots, x_{N_{\mathcal{M}}}, y_{N_{\mathcal{M}}}, z_{N_{\mathcal{M}}})^\top$ consisting of $N_{\mathcal{M}}$ vertices $(x_n, y_n, z_n)^\top$ as

$$\mathbf{f} = \bar{\mathbf{f}} + \mathcal{C} \times_2 \mathbf{w}_{\text{id}}^\top \times_3 \mathbf{w}_{\text{exp}}^\top, \tag{1}$$

where $\mathbf{w}_{\text{id}} \in \mathbb{R}^{N_{\text{id}}}$ and $\mathbf{w}_{\text{exp}} \in \mathbb{R}^{N_{\text{exp}}}$ are linear weights for identity and expression, respectively. $\times_i$ denotes the i-th mode product. $\bar{\mathbf{f}}$ is the mean face in the training dataset. The tensor $\mathcal{C}$, or called the core tensor, encoding the subspaces that span the shape variations of faces, is calculated by high-order singular value decomposition (HOSVD) to the training dataset, *i.e.*, $\mathcal{C} = \mathcal{T} \times_2 \mathbf{U}_{\text{id}} \times_3 \mathbf{U}_{\text{exp}}$. $\mathbf{U}_{\text{id}}$ and $\mathbf{U}_{\text{exp}}$ are unitary matrices from the mode-2 and mode-3 HOSVD to the data tensor $\mathcal{T} \in \mathbb{R}^{3N_{\mathcal{M}} \times N_{\text{id}} \times N_{\text{exp}}}$. $\mathcal{T}$ is a 3D tensor that collects the offsets against the mean face $\bar{\mathbf{f}}$ from face meshes with varying identities and expressions in the training dataset. We employ the FaceWarehouse dataset [12] as the training dataset since it contains thousands of face meshes with a comprehensive set of expressions and a variety of identities including different ages, genders and races.

3.2. Proposed Statistical Face Model

Unlike conventional approaches, we do not employ a single face template with heuristically determined parameters to fit the target point cloud or track its motion, as doing so may lead to poor fitting to the user or be incompatible with local expression variations. Instead, we propose a face model in which the face shape can be probabilistically generated from a computed distribution, with the dynamics of



(a) Overall variation (b) Variation by $\boldsymbol{\Sigma}_{\mathcal{I}}$ (c) Variation by $\boldsymbol{\Sigma}_{\mathcal{E}}$ (d) Variation by residuals

Figure 2. The statistics of the face model trained in the FaceWarehouse dataset [12]. (a) Overall shape variation. (b)–(c) Shape variations by $\mathbf{w}_{\text{id}}$ and $\mathbf{w}_{\text{exp}}$, respectively. (d) Shape variation by the residual term in Eq. (2). The shape variation is set as one standard deviation of the marginalized per-vertex distribution.

the tracked face reliably predicted. Such a model essentially provides probabilistic priors for robust face pose tracking.

3.2.1 Identity and Expression Priors

It is reasonable to assume the identity weight $\mathbf{w}_{\text{id}}$ and expression weight $\mathbf{w}_{\text{exp}}$ follow two independent Gaussian distributions, $\mathbf{w}_{\text{id}} = \boldsymbol{\mu}_{\text{id}} + \boldsymbol{\epsilon}_{\text{id}}, \boldsymbol{\epsilon}_{\text{id}} \sim \mathcal{N}(\boldsymbol{\epsilon}_{\text{id}}|\mathbf{0}, \boldsymbol{\Sigma}_{\text{id}})$ and $\mathbf{w}_{\text{exp}} = \boldsymbol{\mu}_{\text{exp}} + \boldsymbol{\epsilon}_{\text{exp}}, \boldsymbol{\epsilon}_{\text{exp}} \sim \mathcal{N}(\boldsymbol{\epsilon}_{\text{exp}}|\mathbf{0}, \boldsymbol{\Sigma}_{\text{exp}})$. These prior distributions can be estimated from the training data. In particular, we learn that $\boldsymbol{\mu}_{\text{id}} = \frac{1}{N_{\text{id}}}\mathbf{U}_{\text{id}}^{\top}\mathbf{1}$ and $\boldsymbol{\mu}_{\text{exp}} = \frac{1}{N_{\text{exp}}}\mathbf{U}_{\text{exp}}^{\top}\mathbf{1}$. The variance matrices are set to identity matrices with scales, *i.e.*, $\boldsymbol{\Sigma}_{\text{id}} = \sigma_{\text{id}}^2\mathbf{I}$, and $\boldsymbol{\Sigma}_{\text{exp}} = \sigma_{\text{exp}}^2\mathbf{I}$, where $\sigma_{\text{id}}^2 = \frac{1}{N_{\text{id}}}$ and $\sigma_{\text{exp}}^2 = \frac{1}{N_{\text{exp}}}$ are empirically learned from the training set. Note that $\boldsymbol{\mu}_{\text{id}}$ (or $\boldsymbol{\mu}_{\text{exp}}$) should not be $\mathbf{0}$ as it will possibly let the face model $\mathbf{f}$ insensitive to $\mathbf{w}_{\text{exp}}$ (or $\mathbf{w}_{\text{id}}$) [5].

3.2.2 Multilinear Face Model Prior

The canonical face model $\mathcal{M}$ with respect to $\mathbf{w}_{\text{id}}$ and $\mathbf{w}_{\text{exp}}$ can be written in the form

$$\begin{aligned}\mathbf{f} = \bar{\mathbf{f}} + \mathcal{C} \times_2 \boldsymbol{\mu}_{\text{id}} \times_3 \boldsymbol{\mu}_{\text{exp}} + \mathcal{C} \times_2 \boldsymbol{\epsilon}_{\text{id}} \times_3 \boldsymbol{\mu}_{\text{exp}} \\ + \mathcal{C} \times_2 \boldsymbol{\mu}_{\text{id}} \times_3 \boldsymbol{\epsilon}_{\text{exp}} + \mathcal{C} \times_2 \boldsymbol{\epsilon}_{\text{id}} \times_3 \boldsymbol{\epsilon}_{\text{exp}}.\end{aligned} \tag{2}$$

The last term in (2) is usually negligible on the shape variation, as visualized in Fig. 2. Therefore, $\mathcal{M}$ approximately follows a Gaussian distribution as

$$p_{\mathcal{M}}(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_{\mathcal{M}}, \boldsymbol{\Sigma}_{\mathcal{M}}), \tag{3}$$

where its neutral face is $\boldsymbol{\mu}_{\mathcal{M}} = \bar{\mathbf{f}} + \mathcal{C} \times_2 \boldsymbol{\mu}_{\text{id}} \times_3 \boldsymbol{\mu}_{\text{exp}}$, and its variance matrix is given by $\boldsymbol{\Sigma}_{\mathcal{M}} = \mathbf{P}_{\text{id}}\boldsymbol{\Sigma}_{\text{id}}\mathbf{P}_{\text{id}}^{\top} + \mathbf{P}_{\text{exp}}\boldsymbol{\Sigma}_{\text{exp}}\mathbf{P}_{\text{exp}}^{\top}$. The projection matrices $\mathbf{P}_{\text{id}}$ and $\mathbf{P}_{\text{exp}}$ for identity and expression are defined as: $\mathbf{P}_{\text{id}} = \mathcal{C} \times_3 \boldsymbol{\mu}_{\text{exp}} \in \mathbb{R}^{3N_{\mathcal{M}} \times N_{\text{id}}}$, $\mathbf{P}_{\text{exp}} = \mathcal{C} \times_2 \boldsymbol{\mu}_{\text{id}} \in \mathbb{R}^{3N_{\mathcal{M}} \times N_{\text{exp}}}$. $\boldsymbol{\mu}_{\mathcal{M}}$ is nearly the same as the mean face $\bar{\mathbf{f}}$, since $\|\boldsymbol{\mu}_{\mathcal{M}} - \bar{\mathbf{f}}\|_2 = \frac{1}{N_{\text{id}}N_{\text{exp}}}\|\mathcal{C} \times_2 (\mathbf{U}_{\text{id}}^{\top}\mathbf{1}) \times_3 (\mathbf{U}_{\text{exp}}^{\top}\mathbf{1})\|_2 \simeq 0$. It means that the priors of $\mathbf{w}_{\text{id}}$ and $\mathbf{w}_{\text{exp}}$ do not add biases to the face model $\mathcal{M}$ for the representation of the training dataset.

We are also interested in the identity adaptation that is invariant to the expression variations. The joint distribution of the face model and the identity parameter is

$$\begin{aligned}p(\mathbf{f}, \mathbf{w}_{\text{id}}) &= p_{\mathcal{M}}(\mathbf{f}|\mathbf{w}_{\text{id}})p(\mathbf{w}_{\text{id}}) \\ &= \mathcal{N}(\mathbf{f}|\bar{\mathbf{f}} + \mathbf{P}_{\text{id}}\mathbf{w}_{\text{id}}, \boldsymbol{\Sigma}_{\mathcal{E}})\mathcal{N}(\mathbf{w}_{\text{id}}|\boldsymbol{\mu}_{\text{id}}, \boldsymbol{\Sigma}_{\text{id}}),\end{aligned} \tag{4}$$

where the variance of the expression $\boldsymbol{\Sigma}_{\mathcal{E}} = \mathbf{P}_{\text{exp}}\boldsymbol{\Sigma}_{\text{exp}}\mathbf{P}_{\text{exp}}^{\top}$ is captured in the likelihood $p(\mathbf{f}|\mathbf{w}_{\text{id}})$. It is therefore robust to local shape variations led by expression, and the posterior of $\mathbf{w}_{\text{id}}$ will be less affected by the user's current expression. On the other hand, once the identity is adapted to current user, it will help adjust the expression variance $\boldsymbol{\Sigma}_{\mathcal{E}}$ and thus increases the robustness in pose estimation.

As shown in Fig. 2, the overall shape variation (represented as per-pixel standard deviation) is, unsurprisingly, most significant in the facial region as compared to other parts of the head. We further observe that this shape variation is dominated by differences in identities, as encoded by $\boldsymbol{\Sigma}_{\mathcal{I}} = \mathbf{P}_{\text{id}}\boldsymbol{\Sigma}_{\text{id}}\mathbf{P}_{\text{id}}^{\top}$. While as expected, the shape uncertainties by the expressions $\boldsymbol{\Sigma}_{\mathcal{E}}$ are usually localized around the mouth and chin, as well as the regions around cheek and eyebrow. More importantly, the variation by the residual term in Eq. (2) has a much lower magnitude than those caused solely by identity and expression.

4. Probabilistic Facial Pose Tracking

In this section, we present our probabilistic facial pose tracking. Fig. 3 shows the overall architecture, which consists of two main components: 1) robust facial pose tracking, and 2) online identity adaptation. The first component is to estimate the rigid facial pose $\boldsymbol{\theta}$, given an input depth image and the probabilistic facial model $p_{\mathcal{M}}(\mathbf{f})$. The pose parameters $\boldsymbol{\theta}$ include not only the rotation angles $\boldsymbol{\omega}$ and translation vector $\mathbf{t}$, but also the scale s, as the face model may not match the input point cloud due to scale differences. The purpose of the second component is to update the distribution of the identity parameter $\mathbf{w}_{\text{id}}$ and the probabilistic face model $p_{\mathcal{M}}(\mathbf{f})$, given the previous face model, the current pose parameter and the input depth image.

4.1. Robust Facial Pose Tracking

Prior to tracking or after tracking failure, we need to detect the position of the face in the first frame. We employ the head detection method by Meyer *et al.* [26], and crop the input depth map to get a depth patch centered at the detected head center within a radius of $r = 100$ pixels. Denote the point cloud extracted from this depth patch as $\mathcal{P}$.

The pose parameters $\boldsymbol{\theta} = \{\boldsymbol{\omega}, \mathbf{t}, \alpha\}$ indicate the rotation angles, translation vector, and the logarithm of the scale s, *i.e.*, $s = e^{\alpha} > 0, \forall \alpha \in \mathbb{R}$. A canonical face model point $\mathbf{f}_n$ is rigidly warped into $\mathbf{q}_n, n \in \{1, \ldots, N_{\mathcal{M}}\}$ as

$$\mathbf{q}_n = \mathbf{T}(\boldsymbol{\theta}) \circ \mathbf{f}_n = e^{\alpha}\mathbf{R}(\boldsymbol{\omega})\mathbf{f}_n + \mathbf{t}, \tag{5}$$

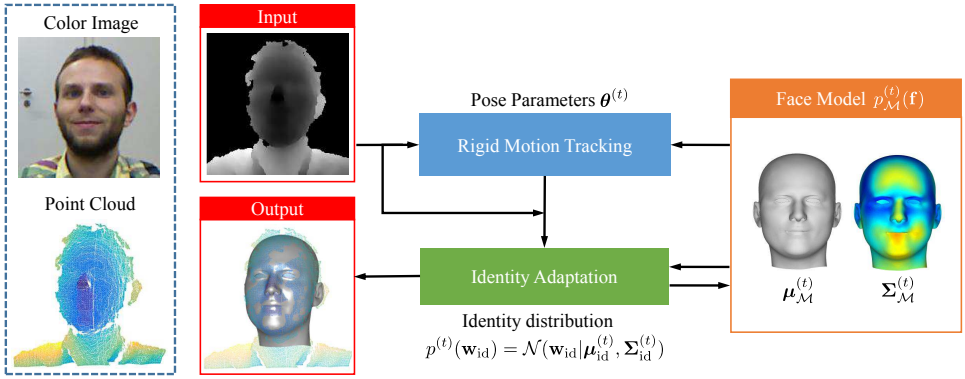


Figure 3. Overview of the propose probabilistic framework, which consists of two components: robust facial pose estimation and online identity adaptation. For both components, the generative model $p_{\mathcal{M}}^{(t)}(\mathbf{f})$ acts as the key intermediate and it is updated immediately with the feedback of the identity adaptation. The input to the system is the depth map while the output is the rigid pose parameter $\boldsymbol{\theta}^{(t)}$ and the updated face identity parameters $\{\mu_{\text{id}}^{(t)}, \boldsymbol{\Sigma}_{\text{id}}^{(t)}\}$ that encode the identity distribution $p^{(t)}(\mathbf{w}_{\text{id}})$. Note that the color image is for illustration but not used in our system.

Figure 4. Samples of the occluded faces.

where the transformation $\mathbf{T}(\boldsymbol{\theta}) \circ \mathbf{f}_n$ describes this rigid warping, while $\mathbf{R}(\boldsymbol{\omega})$ is the rotation matrix. Thus, the warped face model $\mathcal{Q}$ owns a similar distribution for each $\mathbf{q}_n \in \mathcal{Q}$ that is closely related to (3):

$$p_{\mathcal{Q}}(\mathbf{q}_n; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{q}_n|\mathbf{T}(\boldsymbol{\theta}) \circ \boldsymbol{\mu}_{\mathcal{M},[n]}, e^{2\alpha}\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}), \quad (6)$$

where $\boldsymbol{\mu}_{\mathcal{M},[n]}$ and $\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}$ are the mean vector and the rotated variance matrix for point $\mathbf{f}_n$ in $\boldsymbol{\mu}_{\mathcal{M}}$ and $\boldsymbol{\Sigma}_{\mathcal{M}}$, respectively. Moreover, we have $\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})} = \mathbf{R}(\boldsymbol{\omega})\boldsymbol{\Sigma}_{\mathcal{M},[n]}\mathbf{R}(\boldsymbol{\omega})^{\top}$.

To find an optimal pose that matches the warped face model $\mathcal{Q}$ and the input point cloud $\mathcal{P}$, we expect the surface distribution of $\mathcal{P}$ to be within the range spanned by the distribution of face model $\mathcal{Q}$. However, in practical uncontrolled scenarios, we often encounter self-occlusions or object-to-face occlusions, where the occluded human faces are always behind the occluding objects, like hair, glasses and fingers/hands, as shown in Fig. 4. In these scenarios, even if the face model $\mathcal{Q}$ and the input point cloud $\mathcal{P}$ are correctly aligned, $\mathcal{Q}$ only partially fits a subset point cloud in $\mathcal{P}$ with the remaining points in $\mathcal{Q}$ are occluded.

Therefore, it is necessary to identify the non-occluded or visible parts of $\mathcal{Q}$ that overlap with $\mathcal{P}$, based on which we can robustly track the facial pose. For identifying visible parts, we did not follow a strict correspondence-based method like distance thresholding and normal vector compatibility check [19], since finding reliable correspondence is itself challenging. Instead, we propose a *ray visibility constraint* (RVC) to regularize the visibility of each face model point, based on our developed statistical face prior.

4.1.1 Ray Visibility Constraint

Formally we can specify the ray connecting the camera center to a face model point $\mathbf{q}_n$ as $\vec{v}_{(\mathbf{q}_n,\mathbf{p}_n)}$, by identifying $\mathbf{p}_n$ as the point in $\mathcal{P}$ nearest to this ray. This point can be found by matching pixel locations with $\mathbf{q}_n$ via a lookup-table [19, 23]. If $\mathbf{q}_n$ is visible, it should be closely located to the local surface extracted from $\mathcal{P}$. If $\mathbf{q}_n$ is not visible, it must be occluded by the surface and thus located further away than the surface. However, if $\mathbf{q}_n$ is in front of the surface point along the ray, it should suffer obligatory penalty that will push the face model $\mathcal{Q}$ farther away so as to let $\mathbf{q}_n$ be around the surface of $\mathcal{P}$. Eventually, the face model will tightly and/or partially fit $\mathcal{P}$ while leaving the rest of the points as occlusions. Please take Fig. 5 as examples.

The surface of $\mathcal{P}$ is locally defined by fitting planes to neighboring points. Thus if $\mathbf{q}_n$ is linked to $\mathbf{p}_n$ through the ray $\vec{v}_{(\mathbf{q}_n,\mathbf{p}_n)}$, the signed distance of $\mathbf{q}_n$ to the surface is

$$\Delta(\mathbf{q}_n; \mathbf{p}_n) = \mathbf{n}^{\top}(\mathbf{q}_n - \mathbf{p}_n) \quad (7)$$

where $\mathbf{n}_n$ is the normal vector of the local plane centered at point $\mathbf{p}_n$. Based on (6), the distribution of the signed distance $p_{\mathcal{Q}\to\mathcal{P}}(y_n; \boldsymbol{\theta})$ can be modeled as

$$\mathcal{N}\left(y_n|\Delta(\mathbf{T}(\boldsymbol{\theta}) \circ \boldsymbol{\mu}_{\mathcal{M},[n]}; \mathbf{p}_n), \sigma_o^2 + e^{2\alpha}\mathbf{n}_n^{\top}\boldsymbol{\Sigma}_{\mathcal{M},[n]}^{(\boldsymbol{\omega})}\mathbf{n}_n\right), \quad (8)$$

where σ_o^2 is the data noise variance of $\mathcal{P}$ taking into account the surface modeling error and the sensor's systematic error.

According to the ray visibility constraint, we can classify the point $\mathbf{q}_n$ with the label $\gamma_n = \{0, 1\}$:

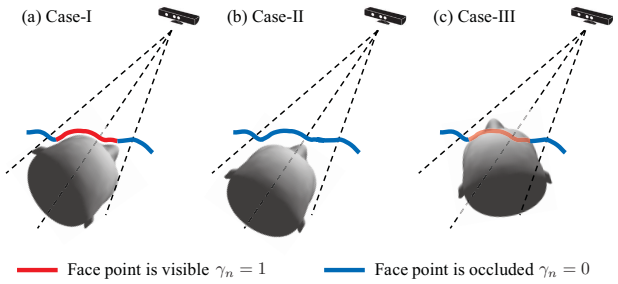


Figure 5. Illustration of the ray visibility constraint. A profiled face model and a curve on the point cloud are in front of a depth camera. (a) A part of face points fit the curve, while the rest points are occluded. (b) The face model is completely occluded. (c) An unrealistic case that the face model occludes the point cloud.

i) $\mathbf{q}_n$ is visible ($\gamma_n = 1$). If the point $\mathbf{q}_n$ is visible along the ray $\vec{v}_{(\mathbf{q}_n,\mathbf{p}_n)}$, $\mathbf{q}_n$ should be *around* or *in front of* the surface centered at $\mathbf{p}_n$. That is $\Delta(\mathbf{T}(\boldsymbol{\theta}) \circ \boldsymbol{\mu}_{\mathcal{M},[n]}; \mathbf{p}_n)$ should be within the bandwidth of $p_{\mathcal{Q}\to\mathcal{P}}(y_n)$ or negative[1]:

$$\Delta(\mathbf{T}(\boldsymbol{\theta}) \circ \boldsymbol{\mu}_{\mathcal{M},[n]}; \mathbf{p}_n) \leq \sqrt{\sigma_o^2 + e^{2\alpha}\mathbf{n}_n^\top \boldsymbol{\Sigma}^{(\boldsymbol{\omega})}_{\mathcal{M},[n]}\mathbf{n}_n}.$$

ii) $\mathbf{q}_n$ is occluded ($\gamma_n = 0$). Similarly, the point $\mathbf{q}_n$ is assumed to be occluded when its signed distance is positive and beyond the confidence interval of $p_{\mathcal{Q}\to\mathcal{P}}(y_n; \boldsymbol{\theta})$:

$$\Delta(\mathbf{T}(\boldsymbol{\theta}) \circ \boldsymbol{\mu}_{\mathcal{M},[n]}; \mathbf{p}_n) > \sqrt{\sigma_o^2 + e^{2\alpha}\mathbf{n}_n^\top \boldsymbol{\Sigma}^{(\boldsymbol{\omega})}_{\mathcal{M},[n]}\mathbf{n}_n}.$$

Theoretically, we are able to compute the posteriors for $\{\gamma_n\}_{n=1}^{N_\mathcal{M}}$ so as to favor a full Bayesian framework for pose estimation. But in practice, we find the binary labels are efficient and will not degrade performance too much.

4.1.2 Ray Visibility Score

Here we develop a *ray visibility score* (RVS) to measure the compatibility between the distributions of face model points $\mathcal{Q}$ and the input point cloud $\mathcal{P}$.

Consider a ray $\vec{v}_{(\mathbf{q}_n,\mathbf{p}_n)}$ connecting a model point $\mathbf{q}_n$ and input point $\mathbf{p}_n$. Assume $\mathcal{Q}$ is correctly aligned, via the signed distance y_n, the distribution of $\mathbf{p}_n$ is modeled as

$$p_\mathcal{P}(y_n) = \mathcal{N}(y_n|0, \sigma_o^2)^{\gamma_n}\mathcal{U}_\mathcal{O}(y_n)^{1-\gamma_n}, \tag{9}$$

where $\mathcal{U}_\mathcal{O}(y_n) = U_\mathcal{O}$ is a uniform distribution. (9) takes into account the visibility labels. When $\mathbf{q}_n$ is visible, $\mathbf{p}_n$ has a compatible surface distribution of $\mathcal{N}(y_n|0, \sigma_o^2)$. However, if $\mathbf{q}_n$ is occluded, $\mathbf{p}_n$ can be arbitrary as long as it is in front of $\mathbf{q}_n$, which we model as a uniform distribution $\mathcal{U}_\mathcal{O}(y_n)$. Similarly, given $\mathcal{P}$, we can project the face model $p_\mathcal{Q}(\mathbf{q}_n; \boldsymbol{\theta})$ onto the local surface of $\mathbf{p}_n$'s, just as $p_{\mathcal{Q}\to\mathcal{P}}(y_n; \boldsymbol{\theta})$. $p_\mathcal{P}(y_n)$ can be regarded as a noisy face measurements contaminated by occlusions, while $p_{\mathcal{Q}\to\mathcal{P}}(y_n; \boldsymbol{\theta})$ denotes the face model with its own uncertainties, both of which are projected on the surfaces of $\mathcal{P}$.

The ray visibility score $\mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\theta})$ is to measure the similarity between $p_\mathcal{P}(\mathbf{y}) = \prod_{n=1}^{N_\mathcal{M}} p_\mathcal{P}(y_n)$ and $p_\mathcal{Q}(\mathbf{y}; \boldsymbol{\theta}) = \prod_{n=1}^{N_\mathcal{M}} p_{\mathcal{Q}\to\mathcal{P}}(y_n; \boldsymbol{\theta})$ by the Kullback-Leibler divergence,

$$\mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\theta}) = D_{KL}\left[p_\mathcal{Q}(\mathbf{y}; \boldsymbol{\theta})||p_\mathcal{P}(\mathbf{y})\right] \tag{10}$$

so that the more similar $p_\mathcal{P}(\mathbf{y})$ and $p_\mathcal{Q}(\mathbf{y}; \boldsymbol{\theta})$ are, the smaller $\mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\theta})$ is. Thus, the optimal pose parameter $\boldsymbol{\theta}^*$ is the one minimizing the ray visibility score:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta},\boldsymbol{\gamma}} \mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\theta}). \tag{11}$$

Note that (10) does not only account for the visible points but also penalizes occluded points to some extent, which avoids a degenerate solution with only a trivial number of perfectly aligned visible points, while the bulk of the points are labeled as occluded.

Algorithm 1: Robust 3D Facial Pose Tracking

Input : Input depth frame $\mathbf{D}_t$;
Previous pose parameters $\boldsymbol{\theta}^{(t-1)}$;
Output: Current pose parameters $\boldsymbol{\theta}^{(t)}$;

$\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}^{(t-1)}$;
if tracking failed **then** $\boldsymbol{\theta}_0 \leftarrow$ `head_detect`$(\mathbf{D}_t)$;
$\mathcal{P} \leftarrow$ `extract_point_cloud`$(\mathbf{D}_t, \boldsymbol{\theta}_0)$;
generate particles $\{\boldsymbol{\phi}_i\}_{i=1}^{N_{\text{particle}}}$ around initial pose $\boldsymbol{\theta}_0$;
for $\tau \leftarrow 1$ to N_{iter} **do**
  for $i \leftarrow 1$ to $N_{particle}$ **do**
    update $\boldsymbol{\phi}_i$ by optimizing $\mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\phi}_i)$ in Sec. 4.1.3
  particle swarm update of all particles
$\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\phi}_{\text{best}}$ where $\mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\phi}_{\text{best}})$ has the minimum score

4.1.3 Rigid Pose Estimation

Solving (11) is challenging since $\mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\theta})$ is highly nonlinear with no off-the-shelf closed-form solution. In this work, we apply a recursive estimation method to solve this problem. In particular, in each iteration, we alternatively estimate the intermediate $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\gamma}^{(t)}$. In the first subproblem, we apply the quasi-Newton update $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)} + \Delta\boldsymbol{\theta}$ using the trust region approach for $\mathcal{S}(\mathcal{Q}, \mathcal{P}; \boldsymbol{\theta}^{(t-1)})$ under the previous $\boldsymbol{\gamma}^{(t-1)}$. The second one is to update the visibility label set $\boldsymbol{\gamma}^{(t)} = \{\gamma_n^{(t)}\}_{n=1}^{N_\mathcal{M}}$ by examining the ray visibility constraint to all point pairs $\{\vec{v}_{(\mathbf{q}_n,\mathbf{p}_n)}\}_{n=1}^{N_\mathcal{M}}$ from the current pose $\boldsymbol{\theta}^{(t)}$ and face model $p_\mathcal{M}^{(t)}(\mathbf{f})$. The process repeats until convergence or beyond the predefined iteration numbers. Moreover, the particle swarm optimization

[1] We keep $\mathbf{n}_n$ pointing to the captured scene. Thus the negative signed distance y_n means $\mathbf{q}_n$ is in front of the surface.

	N_{seq}	N_{frm}	N_{subj}	Difficulties	$\boldsymbol{\omega}_{\max}$
BIWI [18]	24	~15K	25	occlusions expressions	$\pm 75^\circ$ yaw $\pm 60^\circ$ pitch
ICT-3DHP [1]	10	~14K	10	occlusions expressions	$\pm 75^\circ$ yaw $\pm 45^\circ$ pitch

Table 1. Facial Pose Datasets Summarization

(PSO) [27, 26] is further introduced to effectively eliminate the misalignment problem due to poor initialization, and rectify the wrong estimation when the optimization gets stuck in bad local minima of the RVS. A sketch of the rigid facial pose tracking is listed in Algorithm 1.

4.2. Online Identity Adaptation

In parallel with the rigid pose tracking, the face model is also progressively updated to adapt the user's identity. Because the identity is initially unknown when a new user first appears in the sensors, we begin with a generic face model, and then the identity is gradually personalized. In this work, local shape variations by expressions are effectively separated in our probabilistic model, thus the estimated identity is robust to the interferences by expression variations.

As depicted in Sec. 3.2, the face model is personalized by the identity distribution $p^\star(\mathbf{w}_{\text{id}}) = \mathcal{N}(\mathbf{w}_{\text{id}}|\boldsymbol{\mu}_{\text{id}}^\star, \boldsymbol{\Sigma}_{\text{id}}^\star)$. However, the exact $p^\star(\mathbf{w}_{\text{id}})$ is unknown if no adequate depth samples are available. Thus the face identity requires sequential update such as the *assumed-density filtering* (ADF) [2], to approximate $p^{(t)}(\mathbf{w}_{\text{id}})$ from the posterior induced by the current likelihood $p_{\mathcal{L}}(\mathbf{y}^{(t)}|\mathbf{w}_{\text{id}}; \boldsymbol{\theta}^{(t)})$ and the previous best estimate $p^{(t-1)}(\mathbf{w}_{\text{id}})$.

The likelihood $p_{\mathcal{L}}(\mathbf{y}^{(t)}|\mathbf{w}_{\text{id}}; \boldsymbol{\theta}^{(t)})$ that models the distances from the visible face model points to the surface of $\mathcal{P}$, as well as the distances for the occluded points:

$$\prod_{n=1}^{N_{\mathcal{M}}} p_{\mathcal{Q}\to\mathcal{P}}(y_n^{(t)}|\mathbf{w}_{\text{id}}; \boldsymbol{\theta}^{(t)})^{\gamma_n} \mathcal{U}_{\mathcal{O}}(y_n^{(t)})^{1-\gamma_n}, \tag{12}$$

where the projection distribution $p_{\mathcal{Q}\to\mathcal{P}}(y_n^{(t)}|\mathbf{w}_{\text{id}}; \boldsymbol{\theta}^{(t)})$ is similar to $p_{\mathcal{Q}\to\mathcal{P}}(y_n^{(t)}; \boldsymbol{\theta}^{(t)})$ in (8) but with a different mean $m_n = \tilde{\Delta}\left(\mathbf{T}(\boldsymbol{\theta}^{(t)}) \circ (\bar{\mathbf{f}}_n + \mathbf{P}_{\text{id}}\mathbf{w}_{\text{id}}); \mathbf{p}_n\right)$ and variance $\xi_n^2 = \sigma_o^2 + e^{2\alpha^{(t)}} \mathbf{n}_n^{(t)\top} \boldsymbol{\Sigma}_{\mathcal{E},[n]}^{(t-1)} \mathbf{n}_n^{(t)}$. To eliminate the quantization errors in the input point cloud, we modify that $\tilde{\Delta}(\mathbf{q}_n; \mathbf{p}_n) = \text{sign}(\Delta(\mathbf{q}_n; \mathbf{p}_n)) \max\{|\Delta(\mathbf{q}_n; \mathbf{p}_n)| - \varepsilon, 0\}$.

The identity distribution $p^{(t)}(\mathbf{w}_{\text{id}}) = \mathcal{N}(\mathbf{w}_{\text{id}}|\boldsymbol{\mu}_{\text{id}}^{(t)}, \boldsymbol{\Sigma}_{\text{id}}^{(t)})$ is estimated by minimizing $D_{KL}[p^{(t)}(\mathbf{w}_{\text{id}})||p(\mathbf{w}_{\text{id}}|\mathbf{y}^{(t)})]$ [2]. Particularly, we compute the posterior following

$$p(\mathbf{w}_{\text{id}}|\mathbf{y}^{(t)}) \sim p_{\mathcal{L}}(\mathbf{y}^{(t)}|\mathbf{w}_{\text{id}}; \boldsymbol{\theta}^{(t)}) p^{(t-1)}(\mathbf{w}_{\text{id}}). \tag{13}$$

The parameters of $p^{(t)}(\mathbf{w}_{\text{id}})$ are estimated through the variational Bayes framework [2]. We empirically find that this process converges within $3 \sim 5$ iterations.

To fast capture a new user when the face model has been personalized, we add a relaxation to the variance matrix of $p^{(t)}(\mathbf{w}_{\text{id}})$ as $\boldsymbol{\Sigma}_{\text{id}}^{(t)} \leftarrow (\lambda+1)\boldsymbol{\Sigma}_{\text{id}}^{(t)}$ immediately after the identity adaptation. This process is analogous to adding more variances to $\boldsymbol{\mu}_{\text{id}}^{(t)}$ from the identity space $\boldsymbol{\Sigma}_{\text{id}}^{(t)}$, so that it will neither lose the ability to describe a new face, nor fail to preserve the shape of the estimated identity space.

5. Experiments and Discussions

5.1. Datasets And System Setup

We evaluate the proposed method on two public depth-based benchmark datasets, *i.e.*, the Biwi Kinect head pose dataset [18] and ICT 3D head pose (ICT-3DHP) dataset [1]. The dataset summaries are listed in Tab. 1.

Biwi Dataset: Biwi dataset contains over 15K RGB-D images of 20 subjects (different genders and races) in 24 sequences, with large ranges in rotations and translations. The recorded faces suffer the occlusions from hair and accessories and shape variations from facial expressions.

ICT-3DHP Dataset: 10 Kinect RGB-D sequences including 6 males and 4 females are provided by the ICT-3DHP dataset. The data contain similar occlusions and distortions like Biwi dataset. Each subject in this dataset also involves arbitrary expression variations.

System Setup: We implemented the proposed 3D facial pose tracking algorithm in `MATLAB`. The results reported in this paper were measured on a 3.4 GHz Intel Core i7 processor with 16GB RAM. No GPU acceleration was applied.

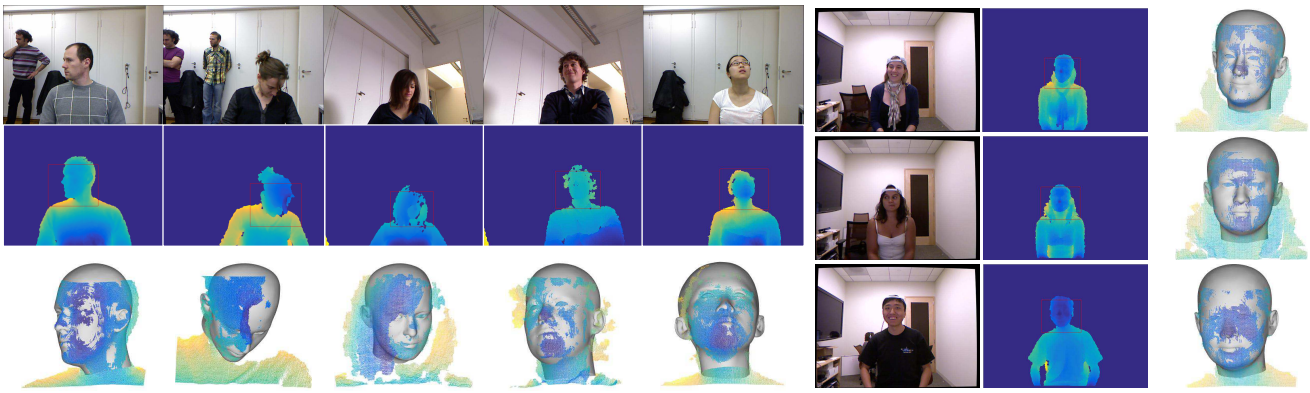
The dimension of the face model is $N_{\mathcal{M}} = 11510, N_{\text{id}} = 150, N_{\text{exp}} = 47$. In practice, we employ a truncated multilinear model with smaller dimensions as $\tilde{N}_{\text{id}} = 28, \tilde{N}_{\text{exp}} = 7$. We set the noise variance as $\sigma_o^2 = 25$, and the outlier distribution is characterized by $\mathcal{U}_{\mathcal{O}}(y) = U_{\mathcal{O}} = \frac{1}{2500}$ [2]. λ is empirically set to 0.25.

Our method adapts the identity for a period of frames and it continues until the adapted face model converges, *i.e.*, the average point-wise difference between adjacent face models is smaller than a given threshold (*e.g.*, 5 mm). The online face adaptation is performed every 10 frames to avoid overfitting to partial facial scans.

5.2. Comparisons with the state-of-the-arts

We compare our method with a number of prior arts [18, 26, 25, 1, 28, 27, 24] for depth-based 3D facial pose tracking on the Biwi [18] and ICT-3DHP [1] datasets. Tab. 2 shows the average absolute errors for the rotation angles and the average Euclidean errors for the translation on the Biwi dataset. The rotational errors were further quantified with respect to the yaw, pitch and roll angles, respectively. Similarly in Tab. 3, we evaluate the average rotation errors on

[2]Note that the measurement unit used in this paper is millimeter (mm)



(a) Biwi dataset (b) ICT-3DHP dataset

Figure 6. (a) Tracking results on the Biwi dataset with the personalized face models. (b) Tracking results on the ICT-3DHP dataset. The extracted point clouds of the head regions are overlaid with the personalized face models. Our system is robust to the profiled faces and occlusions, and is also effective to the facial expression variations.

Method	Errors			
	Yaw (°)	Pitch (°)	Roll (°)	Translation (mm)
ours	**2.3**	**2.0**	**1.9**	**6.9**
RF [18]	8.9	8.5	7.9	14.0
Martin [25]	3.6	2.5	2.6	5.8
CLM-Z [1]	14.8	12.0	23.3	16.7
TSP [28]	3.9	3.0	2.5	8.4
PSO [27]	11.1	6.6	6.7	13.8
Meyer [26]	2.1	2.1	2.4	5.9
Li⋆ [24]	2.2	1.7	3.2	–

Table 2. Evaluations on Biwi dataset

Method	Errors		
	Yaw (°)	Pitch (°)	Roll (°)
ours	**3.4**	**3.2**	**3.3**
RF [18]	7.2	9.4	7.5
CLM-Z [1]	6.9	7.1	10.5
Li⋆ [24]	3.3	3.1	2.9

Table 3. Evaluations on ICT-3DHP dataset

the ICT-3DHP dataset. Note that the results of the reference methods are taken directly from those reported by their respective authors in literature.

On the Biwi dataset, the proposed method produces the overall lowest errors for rotation among the depth-based head pose tracking algorithms [18, 1, 25, 28, 24, 27, 26]. Although no appearance information is used, the proposed approach performs comparable with the state-of-the-art method [24] (marked with ⋆ in Tab. 2 and 3) that employed both RGB and depth data. Similar conclusions can also be drawn on the ICT-3DHP dataset, where the proposed method also achieves a superior performance on estimating the rotation parameters in comparison with the random forests [18] and CLM-Z [1]. Our performance is similar to Li [24] even though no color information is used.

As for the translation parameters, the proposed method also achieves very competitive performance on the Biwi dataset[3]. The sight degradation against Meyer *et al*. [26] shown in Tab. 2 may be because of the incompatibility of model configurations between the groundtruth face model in Biwi dataset and the proposed statistical multilinear face model based on [12, 38].

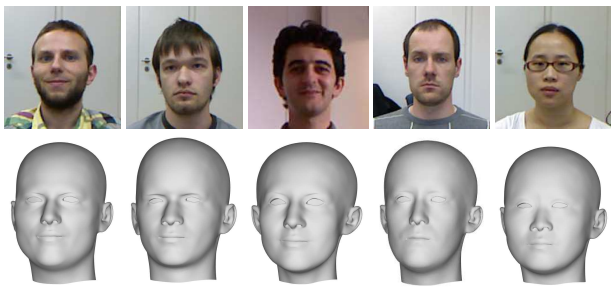
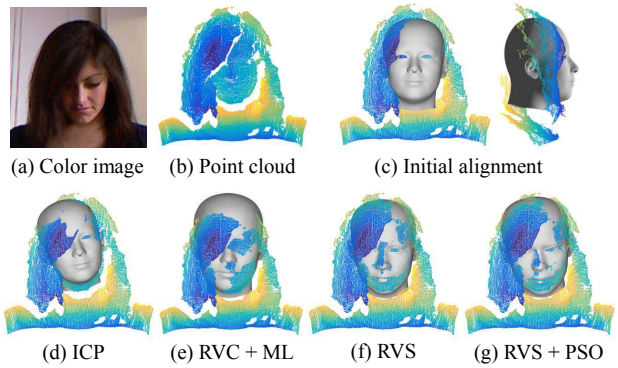


Figure 7. Examples of identity adaptation. Our method successfully adapts the generic model to different identities.

5.3. Visual Results

Fig. 6 shows some tracking results on Biwi and ICT-3DHP datasets based on the gradually adapted face models. Although using generic model can already achieve good performance over challenging cases like occlusions and expression variations with poor initial poses, as shown in Fig. 9, using personalized face model achieves even better results in both the rotation and translation metrics. Moreover, the personalized shape distributions enables the face model to fit compactly with the input point cloud, while the personalized expression distribution makes the estimated facial pose robust to changes in the personalized expressions. Fig. 7 reports a few personalized face models to visually validate the performance.

[3]Groundtruth translations are not available for ICT-3DHP datasets [1].



(a) Color image (b) Point cloud (c) Initial alignment

(d) ICP (e) RVC + ML (f) RVS (g) RVS + PSO

Figure 8. Comparison of the rigid pose estimation methods with the generic face model. (a) and (b): color image and its corresponded point cloud. (c): two views of the initial alignment. (d): result of ICP [31]. (e): result by maximizing the log-likeligood $\log p_{\mathcal{Q}\to\mathcal{P}}(\mathbf{y};\boldsymbol{\theta})$. (f): result by minimizing the RVS. (g): the augmented RVS method by PSO (RVS+PSO).

As for the rigid pose tracking, our method efficiently infers the occlusions caused by hairs, accessories and hand, as well as the self-occlusions like profiled faces, as shown in Fig. 6 by personalized face models, and Fig. 9 and 8 by generic face models. Fig. 1(a) also visualizes the visibility masks on the personalized face models, telling that the proposed method can effectively prune the occlusions out of the pose estimation as well as the face model adaptation.

In comparison with common techniques like iterative closest points (ICP) [31], the proposed method only needs the set of rays $\mathcal{V} = \{\vec{v}_{(\mathbf{q}_n,\mathbf{p}_n)}\}_{n=1}^{N_\mathcal{M}}$ but do not require explicit correspondences during estimation. In contrast, ICP [31] and its variants are not able to check the visibility of each matched point pair, thus cannot guarantee a reasonable pose. For example, as shown in Fig. 8(d), ICP matches the face model with the hairs but has not been aware of the fact that the face cannot occlude the input point cloud. Moreover, the RVS is less vulnerable to bad local minima, since it rewards a higher overlap of the probability distributions $p_\mathcal{P}(\mathbf{y})$ and $p_{\mathcal{Q}\to\mathcal{P}}(\mathbf{y};\boldsymbol{\theta})$ of depth data points and model points, respectively, rather than attempting to compute the maximum likelihood (ML) or maximum a posteriori (MAP) estimate, which is more sensitive to accurate estimation of distribution parameters. For example, maximizing the likelihood $p_{\mathcal{Q}\to\mathcal{P}}(\mathbf{y};\boldsymbol{\theta})$ in (8) may just seek a local mode that fails to catch the major mass of the distribution, as shown in Fig. 8(e). On the contrary, the Kullback-Leibler divergence in RVS ensures the face model distribution with the optimal $\boldsymbol{\theta}$ covers the bulk of information conveyed in $p_\mathcal{P}(\mathbf{y})$. Fig. 8 and 9 reveal the superiority of the RVS and RVS+PSO methods in handling unconstrained facial poses with large rotations and heavy occlusions, even with the generic face model.

(a) Color image (b) Point cloud (c) Initial alignment (d) Ours

Figure 9. Examples of our rigid pose estimation by the generic face model. (a)–(b): color images and the corresponded point clouds. (c): initial alignment provided by the head detection method [26]. (d): the proposed rigid pose estimation results.

5.4. Limitations

The proposed system is inevitably vulnerable when the input depth video is contaminated by heavy noise, outliers and quantization errors. On the other hand, effective clues like facial landmarks are inaccessible due to the missing of the color information. Thus, difficult facial poses (with extreme large rotational angles or occlusions) receiving less confidence from the ray visibility constraint may still be unreliable. However, this problems could be relieved by constraining the temporal coherency of facial poses among adjacent frames like Kalman filtering and *etc*.

6. Conclusions

We propose a robust 3D facial pose tracking for commodity depth sensors that brings about the state-of-the-art performances on two popular facial pose datasets. The proposed generative face model and the ray visibility constraint ensure a robust 3D facial pose tracking that effectively handles heavy occlusions, profiled faces and expression variations, as well as online adapts face model without the interference from the face expression variations.

Acknowledges This research is partially supported by Singapore MoE AcRF Tier-1 Grants RG138/14 and the BeingTogether Centre, a collaboration between Nanyang Technological University (NTU) Singapore and University of North Carolina (UNC) at Chapel Hill. The BeingTogether Centre is supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

[1] T. Baltrušaitis, P. Robinson, and L.-P. Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2610–2617. IEEE, 2012. 2, 6, 7

[2] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 6

[3] M. J. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 374–381, Jun 1995. 1, 2

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 1, 2

[5] T. Bolkart and S. Wuhrer. Statistical analysis of 3d faces in motion. In *Proc. IEEE Int. Conf. 3D Vision*, pages 103–110, June 2013. 3

[6] S. Bouaziz, Y. Wang, and M. Pauly. Online modeling for realtime facial animation. *ACM Trans. Graph.*, 32(4):40, 2013. 2

[7] M. D. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister. Real-time face pose estimation from single range images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1–8. IEEE, 2008. 1, 2

[8] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer. Review of statistical shape spaces for 3d data with comparative analysis for human faces. *Computer Vision and Image Understanding*, 128:1–17, 2014. 2

[9] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang. 3d deformable face tracking with a commodity depth camera. In *Proc. Euro. Conf. Comput. Vis.*, pages 229–242. Springer, 2010. 2

[10] Y. Cai, M. Yang, and Z. Li. Robust head pose estimation using a 3D morphable model. *Mathematical Problems in Engineering*, 2015, 2015. 2

[11] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41, 2013. 2

[12] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graphics*, 20(3):413–425, 2014. 1, 2, 3, 7

[13] C. Chen, H. X. Pham, V. Pavlovic, J. Cai, and G. Shi. Depth recovery with face priors. In *Proc. Asia Conf. Comput. Vis.*, pages 336–351. Springer, 2014. 2

[14] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, (6):681–685, 2001. 1, 2

[15] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008. 1, 2

[16] D. DeCarlo and D. Metaxas. Optical flow constraints on deformable models with applications to face tracking. *International Journal of Computer Vision*, 38(2):99–127, 2000. 1, 2

[17] G. Fanelli, J. Gall, and L. Van Gool. Real time head pose estimation with random regression forests. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 617–624. IEEE, 2011. 1, 2

[18] G. Fanelli, T. Weise, J. Gall, and L. Van Gool. Real time head pose estimation from consumer depth cameras. In *Pattern Recognition*, pages 101–110. Springer, 2011. 1, 2, 6, 7

[19] P.-L. Hsieh, C. Ma, J. Yu, and H. Li. Unconstrained realtime facial performance capture. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1675–1683, 2015. 2, 4

[20] V. Kazemi, C. Keskin, J. Taylor, P. Kohli, and S. Izadi. Real-time face reconstruction from a single depth image. In *3D Vision (3DV), 2014 2nd International Conference on*, volume 1, pages 369–376. IEEE, 2014. 1, 2

[21] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1867–1874, 2014. 1, 2

[22] H. Li, P. Roivainen, and R. Forchheimer. 3-D motion estimation in model-based facial image coding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(6):545–555, Jun 1993. 1, 2

[23] H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.*, 32(4):42–1, 2013. 2, 4

[24] S. Li, K. Ngan, R. Paramesran, and L. Sheng. Real-time head pose tracking with online face template reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015. 2, 6, 7

[25] M. Martin, F. Van De Camp, and R. Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In *3D Vision (3DV), 2014 2nd International Conference on*, volume 1, pages 641–648. IEEE, 2014. 6, 7

[26] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz. Robust model-based 3d head pose estimation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3649–3657, 2015. 2, 3, 6, 7, 8

[27] P. Padeleris, X. Zabulis, and A. A. Argyros. Head pose estimation on depth data based on particle swarm optimization. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 42–49. IEEE, 2012. 2, 6, 7

[28] C. Papazov, T. K. Marks, and M. Jones. Real-time 3D head pose and facial landmark estimation from depth images using triangular surface patch features. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4722–4730, 2015. 1, 2, 6, 7

[29] H. X. Pham and V. Pavlovic. Robust real-time 3d face tracking from rgbd videos under extreme pose, depth, and expression variation. In *Proc. IEEE Int. Conf. 3D Vision*, pages 441–449, Oct 2016. 2

[30] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof. Hough networks for head pose estimation and facial feature localization. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. 1, 2

[31] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001. 2, 8

[32] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from RGB input. *arXiv preprint arXiv:1604.02647*, 2016. 2

[33] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. 1, 2

[34] M. Storer, M. Urschler, and H. Bischof. 3D-MAM: 3D morphable appearance model for efficient fine head pose estimation from still images. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 192–199. IEEE, 2009. 2

[35] Y. Sun and L. Yin. Automatic pose estimation of 3d facial models. In *Proc. IEEE Int. Conf. Pattern Recognit.*, pages 1–4. IEEE, 2008. 1, 2

[36] D. Thomas and R. I. Taniguchi. Augmented blendshapes for real-time simultaneous 3d head modeling and facial motion capture. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3299–3308, June 2016. 2

[37] S. Tulyakov, R.-L. Vieriu, S. Semeniuta, and N. Sebe. Robust real-time extreme head pose estimation. In *Proc. IEEE Int. Conf. Pattern Recognit.*, pages 2263–2268. IEEE, 2014. 2

[38] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. In *ACM Trans. Graph.*, volume 24, pages 426–433. ACM, 2005. 1, 2, 7

[39] R. Wang, L. Wei, E. Vouga, Q. Huang, D. Ceylan, G. Medioni, and H. Li. Capturing dynamic textured surfaces of moving targets. *arXiv preprint arXiv:1604.02801*, 2016. 2

[40] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016. 2

[41] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. In *ACM Trans. Graph.*, volume 30, page 77. ACM, 2011. 2

[42] X. Xiong and F. Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1, 2