J. Vis. Commun. Image R. 48 (2017) 16–29

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Using 3D face priors for depth recovery ☆

Chongyu Chen [a], Hai Xuan Pham [b], Vladimir Pavlovic [b], Jianfei Cai [c], Guangming Shi [d], Yuefang Gao [e], Hui Cheng [a,*]

[a] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, Guangdong 510006, China
[b] Department of Computer Science, Rutgers University, NJ 08854-8019, USA
[c] School of Computer Engineering, Nanyang Technological University, 639798 Singapore, Singapore
[d] School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[e] College of Mathematics and Informatics, South China Agricultural University, Guangzhou, Guangdong 510642, China

ARTICLE INFO

Article history:
Received 7 November 2016
Revised 5 April 2017
Accepted 5 June 2017
Available online 7 June 2017

Keywords:
Depth recovery
Image restoration
Face model

ABSTRACT

For repairing inaccurate depth measurements from commodity RGB-D sensors, existing depth recovery methods primarily rely on low-level and rigid prior information. However, as the depth quality deteriorates, the recovered depth maps become increasingly unreliable, especially for non-rigid objects. Thus, additional high-level and non-rigid information is needed to improve the recovery quality. Taking as a starting point the human face that is the primary prior available in many high-level tasks, in this paper, we incorporate face priors into the depth recovery process. In particular, we propose a joint optimization framework that consists of two main steps: transforming the face model for better alignment and applying face priors for improved depth recovery. Face priors from both sparse and dense 3D face models are studied. By comparing with the baseline method on benchmark datasets, we demonstrate that the proposed method can achieve up to 23.8% improvement in depth recovery with more accurate face registrations, bringing inspirations to both non-rigid object modeling and analysis.

1. Introduction

Commodity RGB-D sensors such as Microsoft Kinect [1] have received significant attention in the recent years due to their low cost and the ability to capture synchronized color images and depth maps in real time. They have been successfully used in many applications such as game and 3D teleconferencing [2–4]. However, the depth measurements provided by commodity RGB-D sensors are far from perfect and often contain degradations such as noise and holes. To obtain high-quality depth maps, plenty of works have been done on depth recovery for commodity RGB-D sensors [5–10]. The common idea is to make use of some general prior information such as the spatial depth consistency, temporal depth continuity, and the coherency between the depth map and its aligned color image. However, as the distance between the camera and the object increases, the depth measurement error grows larger and the quality of color image decreases. In this case, general rigid prior information becomes insufficient for recovering high-quality depth maps, especially for common indoor scenes with non-rigid objects.

Among the non-rigid objects, human face is the most representative one that is widely studied. The space of face shapes is highly restrictive and usually parametric so that it provides high-level priors that are easy to use. Therefore, in this paper, we take depth recovery using face priors as the starting point of semantic prior guided depth recovery for non-rigid objects. Distinguished from the works on face model registration/fitting that adapt the 3D model to the raw input (either color image or depth map), we propose to simultaneously adjust the face model and refine the depth map. The benefits of the proposed framework are twofold. On one hand, 3D model fitting has been reported to be very important in many face analysis tasks such as face recognition [11–13] and facial expression tracking [14,15], and higher quality of depth map generally leads to higher accuracy and better robustness of face analysis. This work, which is designed to obtained facial depth maps of higher quality, can also lead to better performance of face analysis. On the other hand, model fitting methods are also used for rendering novel views of the face [16–18]. Since the facial depth map recovered by the proposed method is with per-vertex correspondences to the face model, it can be also used to render novel views, even with arbitrary face expressions. More importantly,

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: chenchy47@mail.sysu.edu.cn (C. Chen), hxp1@cs.rutgers.edu (H.X. Pham), vladimir@cs.rutgers.edu (V. Pavlovic), asjfcai@ntu.edu.sg (J. Cai), gmshi@xidian.edu.cn (G. Shi), gaoyuefang@scau.edu.cn (Y. Gao), chengh9@mail.sysu.edu.cn (H. Cheng).

http://dx.doi.org/10.1016/j.jvcir.2017.06.002
1047-3203/

the recovered face naturally connects to its surroundings, which makes rendering more realistic.

It should be pointed out that it is non-trivial to derive effective prior information for depth recovery from a 3D face model. First, the 3D model needs to be deformed and aligned with the input RGB-D data. Nevertheless, accurate alignment is hard to achieve due to the heterogeneous and quantized noise in the input data. Second, if the alignment is not accurate, the extracted face priors might provide inaccurate guidance to the depth recovery process. To address these issues, we propose a joint optimization framework to iteratively and alternatively refine the depth map and the face alignment. We consider face priors from both sparse and dense 3D face models as two alternatives and study their tradeoffs as priors for depth recovery. Extensive evaluations on benchmark and real-world data suggest that the proposed method with face priors clearly outperforms the baseline method that relies on generic smoothness constraints. We also show that face priors from a dense model outperforms simpler sparse priors without significant increase in computational cost. A preliminary version of this work, reported in [19], makes use of a sparse face model for face priors. In this paper, we extend our early work by employing a dense face model, performing more comparisons, and providing more inspiring discussions.

2. Related works

There have been plenty of works on depth recovery, face modeling, and non-rigid model registration, which form nice basis of this work.

2.1. Depth recovery

For depth recovery, previous researches focus on designing local filters that can reduce noises while preserving structural details. Numerous anisotropy filters are developed, including bilateral filter [20], joint bilateral filter [21], joint trilateral filter [22], and weighted mode filtering [5]. But these methods usually produces various artifacts because pre-defined filters are difficult to simultaneously handle both sensor diversity and scene complexity. Recent researches began to incorporate various priors for depth recovery. Representative methods include color-guided adaptive autoregressive model [8] that considers the coherence between color and depth, the sensor-oriented optimization model [9] that exploits the prior knowledge of sensor characteristics, and RGB-D fusion [10] that takes into account the scene illumination. It is demonstrated in these works that employing priors for depth recovery can lead to higher restoration quality. However, the employed priors are for static and rigid scenes. For dynamic and non-rigid objects, depth recovery becomes more challenging. Even the latest method [23] that considers temporal consistency still cannot handle the case of long camera-object distance where the depth map is with severe degradations.

2.2. Face modeling

As a representative non-rigid object, human face has been studied for decades. Face modeling usually includes the estimation of face pose, shape, and deformations, for which statistical models such as active shape models (ASMs) [24] and active appearance models (AAMs) [25,26] are common and effective approaches. There have been efforts to incorporate depth data into these techniques in recent years. For example, in [27] the depth frame is used as an additional texture to the traditional color texture in the ASM framework. In [28], the AAM framework is extended by fitting the 3D shape to the point cloud using the *Iterative Closest Point* (ICP) [29] separately after each AAM optimization iteration. However, their strong dependency on the training data limits their robustness in unconstrained environments. Motivated by the idea of model based image compression, researchers began to design face models with a number of vertices that can capture sufficient face expressions while being controlled by a small number of parameters, such as Candide-2 [30] and Candide-3 [31]. In particular, Candide-3 wireframe model can be easily extended to support depth input so that it can provide efficient face priors for depth recovery.

Recent advance in face modeling [32,15,14,33] also follows the idea of controlling rich expressions via small number of parameters, Typically, the 3D face model is controlled by a set of static shape deformation units (SUs) and action deformation units (AUs). In particular, SUs represent the face biometry of an individual, whereas AUs model the facial expressions. Since there are more vertices in these models, it is believed that they can provide more effective priors for facial depth recovery.

2.3. Non-rigid model registration

Methods for non-rigid registration between the face model and the input play a key role in utilizing high-level prior knowledge in depth recovery, which have been widely studied for decades. In the challenging tasks of pose-invariant face recognition and fine-scale face expression tracking, non-rigid registration methods are developed to synthesize the facial texture images or fit the face model to the facial depth maps.

In image based pose-invariant face recognition, face synthesis using 3D face models has achieved great success [34], which is unsurprising because 3D face models reveal the intrinsic physical factors of the face. Roughly two categories of face models are developed and fitted to the input color images, i.e., principal component analysis (PCA) based linear models [11,16,18] that learn statistical face characteristics from training data and generic elastic models (GEMs) [35,17] that abstract major face shapes. Both PCA-based models and GEMs are designed to produce novel facial images at desired views, which typically start with accurate estimation of sparse correspondences. Since novel facial images are unnecessary for depth recovery at the current view, there is no need of a complete implementation of these models for depth recovery.

Recent advance in depth based facial expression tracking [14,15,36] performs expression tracking by registering 3D face models to the input depth map. As the tracking continues, their methods estimate the deformation parameters in real time while changing the shape parameters of the pre-defined 3D face models for the current user. Such strategy is effective when the input depth map can still reveal the face shape. However, when there are strong noises in the depth maps, these methods will probably loose tracking, and the adjusted face model may be far from the user's face. Incorporating depth recovery into face tracking is believed to improve the tracking stability, which partially motivates this work.

3. Technical background

In this section, we introduce the components directly related to our method, including optimization based depth recovery and two face models.

3.1. Depth recovery baseline

The adopted baseline model for depth recovery is a simplified version of our previously work [9], which has general form and

practical effectiveness. In particular, depth recovery is formulated as an energy minimization problem. Given a degraded depth map Z and its corresponding color image I, the depth map is recovered by solving

$$\min_U \lambda E_d(U,Z) + E_r(U), \tag{1}$$

where U is the recovered depth map, λ is the trade-off parameter, E_d is the data term, and E_r is the regularization term. Both E_d and E_r are quadratic functions. In particular, the data term is defined as

$$E_d(U,Z) = \frac{1}{2}\sum_{i\in\Omega_d}\omega_i(U(i)-Z(i))^2, \tag{2}$$

and the regularization term is defined as

$$E_r(U) = \frac{1}{2}\sum_{i\in\Omega_s}\sum_{j\in\Omega_i}\alpha_{ij}(U(i)-U(j))^2, \tag{3}$$

where i stands for pixel index (e.g., $i=(i_x,i_y)$), Ω_d is the set of pixels with valid depth measurements, Ω_s is the set of pixels with sufficient surroundings, and Ω_i is the set of neighboring pixels of pixel i. Taking into account the empirical model of Kinect depth measurements [37], the distance-dependent weight ω_i is defined as

$$\omega_i = \begin{cases} \left(\frac{Z_{\max}-Z(i)}{Z_{\max}-Z_{\min}}\right)^2 & Z(i)\in[Z_{\min},Z_{\max}], \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $Z_{\min}=500$ mm and $Z_{\max}=5000$ mm are the minimum and maximum reliable working distances of Kinect-1 [1]. The weight α_{ij} is designed according to the color and depth similarities between pixel i and pixel j, which will be further explained in Section 4.1 (see [9] for details).

The effectiveness of this framework stems, in part, from the convexity of Eq. (1), implied by the specific forms of Eqs. (2) and (3). This additive energy formulation also makes it possible to include additional terms dependent on 3D shape prior.

3.2. Face shape models and their deformations

In this work, we make use of two popular 3D face models to generate the face priors for depth recovery, i.e., Candide-3 wireframe model [31] and blendshape face model developed from the FaceWarehouse database [33]. The former one is a sparse model, and the latter one is a dense model. We choose them because their deformation parameters are classified in a similar way.

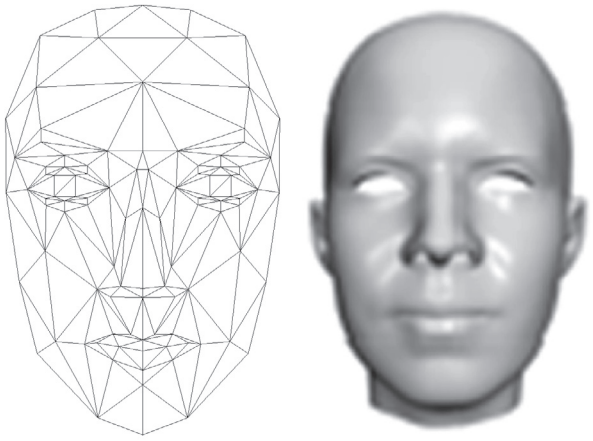
3.2.1. Candide-3 wireframe model

Candide-3 wireframe model consists of 113 vertices and 184 triangles, which are shown in Fig. 1(a). Every vertex $P_k\in\Re^3, k\in\Omega_p=\{1,\ldots,113\}$, of the 3D shape model is formed according to a low-dimensional subspace model:

$$P_k = P_k^0 + S_k\boldsymbol{\sigma} + A_k\boldsymbol{\alpha}, \tag{5}$$

where P_k^0 are the base coordinates of the vertex kth (corresponding to a reference neutral expression face), $S_k\in\Re^{3\times K_s}$ and $A_k\in\Re^{3\times K_a}$ are, respectively, the individual shape and action deformation bases (matrices) associated with the vertex, $\boldsymbol{\sigma}\in\Re^{K_s}$ is the vector of user-specific shape deformation parameters and likewise $\boldsymbol{\alpha}\in\Re^{K_a}$ is the vector of action deformation parameters. For Candide-3 model, $K_s=14$ and $K_a=73$. In this work, without loss of generality, we focus on the static individual shape deformation under the neutral face expression ($\boldsymbol{\alpha}=0$). Thus, the general transformation of a vertex given global rigid rotation R and translation $\boldsymbol{t}$ is defined as:

$$P_k = R\left(P_k^0 + S_k\boldsymbol{\sigma}\right) + \boldsymbol{t}. \tag{6}$$



(a) Candide-3 (b) Blendshape

Fig. 1. (a) Candide-3 Wireframe Model. (b) The blendshape face model with a neutral expression.

The geometry of the model is therefore determined by the base (average) shape P^0 and the user-specific shape deformation S, and is parameterized by the (rigid and non-rigid) deformation vector $\boldsymbol{\theta}=\{R,\boldsymbol{t},\boldsymbol{\sigma}\}$.

3.2.2. Blendshape face model

An example mesh of the utilized blendshape face model with neutral expression is shown in Fig. 1(b). It can be seen that compared to the Candide-3, the blendshape model can represent facial deformations more realistically because of larger number of mesh vertices.

As specified in [33], a facial expression of a person can be approximated by

$$F = C_r\times_2 w_{id}^T\times_3 w_{exp}^T \tag{7}$$

where C_r is a 3D matrix (called reduced core tensor) of size $(3\times N_v)\times N_{id}\times N_e$ (corresponding to number of vertices, number of identities and number of expressions, respectively), w_{id} is an N_{id}-dimension identity vector, and w_{exp} is an N_e-dimension expression vector. Eq. (7) basically describes tensor contraction at the 2nd mode by w_{id} and at the 3rd mode by w_{exp}.

Similar to [38], for real-time face tracking of one person, given his identity vector w_{id}, it is more convenient to reconstruct the N_e expression blendshapes for the person of identity w_{id} as

$$B_j = C_r\times_2 w_{id}^T\times_3 u_{\exp_j}^T \tag{8}$$

where $u_{\exp_j}$ is the pre-computed weight vector for the jth expression mode [33]. In this way, an arbitrary facial shape of the person can be represented as a linear sum of his expression blendshapes [38]:

$$F = B_0 + \sum_{j=1}^{N_e-1}(B_j - B_0)e_j, \tag{9}$$

where B_0 is the neutral shape, and $e_j\in[0,1]$ is the blending weight, $j=1,\ldots,N_e-1$. Finally, a fully transformed 3D facial shape can be represented as

$$P = R\cdot F(B,\boldsymbol{e}) + \boldsymbol{t}, \tag{10}$$

with the parameters $\boldsymbol{\theta}=\{R,\boldsymbol{t},\boldsymbol{e}\}$, where R and $\boldsymbol{t}$ respectively represent global rotation and translation, and $\boldsymbol{e}=\{e_j\}$ defined in (9) represents the deformation parameters. In this work, we keep the 50

most significant identity knobs in the reduced core tensor C_r, hence $N_v \times N_{id} \times N_e = 11,510 \times 50 \times 47$.

For the rest of the paper, we will present how to extract and use the prior information from these two face models to improve the depth recovery process. Note that we use the notations $\boldsymbol{\theta}$ as the parameter vector, P as the set of all the vertices, P_k as the kth vertex of P for both face models.

4. Proposed method

Given a color image I and its corresponding (aligned) noisy depth map Z as input, our goal is to obtain a depth map of the face region with improved quality using the face priors derived from the general 3D deformable models. The pipeline of the proposed method is shown in Fig. 2. The first two components in Fig. 2 are pre-processing steps to roughly clean up the depth data and roughly align the general face model to the point cloud converted from the depth data. The last two components in Fig. 2 are the core of our proposed framework. For the component of the guided depth recovery, we fix the face prior and use it to update the depth, while for the last component, we fix the depth and update the face prior. The last two components alternatively and iteratively operate until convergence.

4.1. Energy model for depth recovery with face priors

To incorporate the face shape prior into the depth recovery process, we propose to recover the depth map U and obtain the deformation parameters $\boldsymbol{\theta}$ for the face model by solving the following optimization problem:

$$\min_{U,\boldsymbol{\theta}} E_r(U) + \lambda_d E_d(U) + \lambda_f E_f(U, \boldsymbol{\theta}), \tag{11}$$

where E_r and E_d are the regularization term and the data term as shown in Eq. (1), E_f is the term designed for the face prior (to be defined below), and λ_d and λ_f are the trade-off parameters.

The definition of E_d follows that of [9], defined in Eq. (2). For E_r, as defined in Eq. (3), we use the normalized weights α_{ij}:

$$\alpha_{ij} = \frac{\beta_{ij}}{\sum_{j\in\Omega_i}\beta_{ij}}, \tag{12}$$

with

$$-\log \beta_{ij} \propto \frac{\|i-j\|^2}{2l_s^2} + \frac{\|I(i)-I(j)\|^2}{2l_I^2} + \frac{(Z(i)-Z(j))^2}{2l_z^2}, \tag{13}$$

which are essentially the weights used in joint trilateral filtering taking into account the range distance, the color difference, and the depth difference [22], with l_s, l_I, and l_z the lengthscale constants for range, color, and depth, respectively.

We define the novel face prior E_f term as

$$E_f(U, \boldsymbol{\theta}) = \sum_{i\in\Omega_f} \eta_i (U(i) - Y(i))^2, \tag{14}$$

where Ω_f is the set of pixels of the face prior, η_i is the weight of the ith guidance depth value,

$$Y = T_f(P(\boldsymbol{\theta})) \tag{15}$$

is the guidance depth map generated from the face model, and T_f is a function that transforms the face model P defined by $\boldsymbol{\theta}$ to a dense depth map compatible with U. This term is critical to the recovery process and will be described in detail in the next section.

4.2. Shape priors for depth recovery

In our work, we use the Candide-3 and the blendshape face models to generate shape priors. The main reasons for choosing these models are twofold. First, their deformation parameters are classified in a similar way. That is, the parameters are either shape parameters that are related to shape changes or action parameters that are related to face actions. In our method, the action parameters are assumed to be fixed for simplicity. Second, these two models are representative sparse and dense models, respectively, which can be used for evaluating the performance of depth recovery with respect to the model density. Note that our proposed depth recovery is a generic method, which is also flexible to utilize any other face model as prior.

4.2.1. Shape prior using Candide-3 model

Considering that the guidance from the sparse vertices of the Candide-3 model are too weak to serve as the prior for the dense depth map U, we generate a dense synthetic depth map Y from the aligned face prior P using an interpolation process. It is possible to define different interpolation functions according to desired dense surface properties. In computer graphics, such models may use non-uniform rational basis spline (NURBS) to guarantee the surface smoothness. Here, for the purpose of a shape prior, we choose a simple piece-wise linear interpolation. Fig. 3 shows an example of the generated dense depth map from the sparse shape P.

To mitigate the effects of the piece-wise flat dense patches due to the linear interpolation, we introduce a weighting scheme defined through weights η_i in (14). In particular, for each pixel $Y(i)$, we use a normalized weight adaptive to the pixel's distances from the neighboring vertices of the sparse shape P. Let (a_i, b_i, c_i) be the barycentric coordinates of pixel i inside a triangle defined by its three neighboring vertices of P. Then, its weight is computed as

$$\eta_i = \sqrt{a_i^2 + b_i^2 + c_i^2}. \tag{16}$$

This suggests that the pixels corresponding to model vertices have the highest weight of 1 while the weights decline towards the center of each triangular patch. An illustration of the weights is given in Fig. 3(c), where bright pixels represent large weights.

4.2.2. Shape prior from a dense face model

Because the blendshape model [33] is dense enough to produce a complete face prior for depth recovery, we can render a dense depth map from the face model without any depth interpolation. However, only a subset of vertices can be used for depth recovery, because some parts of the 3D face model do not contribute as the prior. For example, the model includes the forehead part, which is usually covered by hair. Thus, the guidance from the forehead of the model is ineffective in recovering the depth values of this region of the input depth map. Therefore, we limit the effective face region Ω_f^* within a mask shown in Fig. 4(c). Similar to the case of the sparse face model, we also assign a weight to each pixel of the synthetic depth map to extract effective guidance information from the model. Considering the blendshape model is sufficiently dense that each pixel of the guidance depth map corresponds to a vertex, all the pixels in the face region Ω_f^* should have identical weights. As a result, we use binary weights for the dense model, i.e. $\eta_i = 1,\ \forall i \in \Omega_f^*$ and $\eta_i = 0$ otherwise.

The process of extracting the effective face region consists of three steps. First, we render the dense depth map Y from the face model, as shown in Fig. 4(b). Then, we extract some 2D landmark points on the face by projecting pre-defined 3D landmarks of the

Fig. 2. The pipeline of the proposed method.

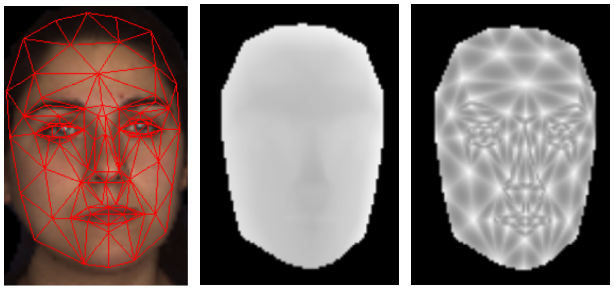


Fig. 3. The face priors from the sparse Candide-3 model. (a) The 3D wireframe model $P(\boldsymbol{\theta})$ drawn upon the texture frame. (b) The synthetic depth map Y generated from the 3D wireframe model. (c) The weights distribution associated with the synthetic dense depth map, where brighter means a larger weight.

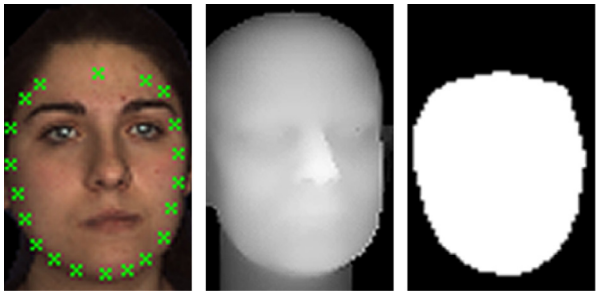


Fig. 4. The face priors from the dense face model. (a) The landmark points drawn upon the texture frame. (b) The depth map rendered from the dense face model. (c) The binary weight map (i.e region mask) for extracting effective guidance from the face model.

blendshape model to the image plane, as illustrated in Fig. 4(a). Finally, we use the contour defined by these landmark points to find the effective face region Ω_f^* and construct the aforementioned binary weight map, as shown in Fig. 4(c). In this way, we can easily extract the effective information from Y in Eq. (14).

4.3. Energy minimization

From the definitions of the energy functions in Eq. (11), it can be seen that the overall optimization of U is a convex task, for a given fixed prior Y. However, the optimization of the parameter set $\boldsymbol{\theta}$ might not be convex since it involves rigid and non-rigid deformations. Therefore, to tackle the global optimization task which includes both the depth map U and the transformation parameters $\boldsymbol{\theta}$, we resort to a standard iterative alternating optimization algorithm. In other words, we first optimize U while keeping $\boldsymbol{\theta}$ fixed, and then optimize $\boldsymbol{\theta}$ over the enhanced depth map U.

Specifically, we divide problem (11) into two well-studied subproblems: depth recovery and shape (rigid and non-rigid) registration. The subproblem of depth recovery is solved with fixed shape parameters $\boldsymbol{\theta}$,

$$\widehat{U} = \arg\min_U E(U,Y) = \arg\min_U E_r(U) + \lambda_d E_d(U) + \lambda_f E_f(U,Y), \tag{17}$$

where Y is the guidance depth map generated from the face model. Since the objective in Eq. (17) is quadratic with respect to U, its optimal solution can be easily found by solving a linear system. In this way, the subproblem of depth recovery converges.

After the depth map $\widehat{U}$ is obtained, we convert it into a point cloud and solve the registration of the shape prior P to this point cloud, which is determined by $\boldsymbol{\theta}$, using ICP approaches [29]. In particular, the point cloud is constructed by projecting the depth pixels from camera plane to 3D space according to the pinhole camera model. The shape registration problem is separated into subproblems of rigid and non-rigid registrations. The initial rigid registration is solved by standard ICP [29], and non-rigid registrations are solved with respect to the face model. In this work, we use $\boldsymbol{\sigma}$ and $\boldsymbol{e}$ to represent the deformation parameters of Candide-3 and the blendshape face models, respectively.

4.3.1. Candide-3 registration

For the Candide-3 model, parameters $\boldsymbol{\theta} = \{R, \boldsymbol{t}, \boldsymbol{\sigma}\}$ are solved by

$$\hat{\boldsymbol{\theta}} = \arg\min_{R,\boldsymbol{t},\boldsymbol{\sigma}} E_{f_1}(R,\boldsymbol{t},\boldsymbol{\sigma}) = \arg\min_{R,\boldsymbol{t},\boldsymbol{\sigma}} \sum_{k\in\Omega_p} (P_k(R,\boldsymbol{t},\boldsymbol{\sigma}) - \boldsymbol{d}(k))^2, \tag{18}$$

where $\boldsymbol{d}(k)$ represents the kth point of the point cloud that corresponds to the model vertex P_k, which are given by ICP correspondence search. Note that the complexity of such correspondence search is dependent on the number of sparse face vertices P.

The optimization in (18) is a non-linear least squares problem because of the multiplication of R and $\boldsymbol{\sigma}$, as shown in Eq. (6). Following the approach in [39], we divide this optimization into two subproblems: rigid registration where we minimize E_{f_1} w.r.t R and $\boldsymbol{t}$ given fixed $\boldsymbol{\sigma}$; and non-rigid registration where E_{f_1} is minimized w.r.t $\boldsymbol{\sigma}$ while keeping R and $\boldsymbol{t}$ constant. These two subproblems are equivalent to two linear least squares problems. The minimization of E_{f_1} reduces the distance between the Candide model and the point cloud, thus it also effectively minimize $E(U,Y)$ in Eq. (17).

4.3.2. Blendshape registration

For the dense blendshape model, parameters $\boldsymbol{\theta} = \{R, \boldsymbol{t}, \boldsymbol{e}\}$ are solved by

$$\hat{\boldsymbol{\theta}} = \arg\min_{R,\boldsymbol{t},\boldsymbol{e}} E_{f_2}(R,\boldsymbol{t},\boldsymbol{e}), \tag{19}$$

where the energy function E_{f_2} consists of two terms, i.e.,

$$E_{f_2}(R,\boldsymbol{t},\boldsymbol{e}) = \frac{1}{N_d}\sum_{k=1}^{N_d}((P_k(R,\boldsymbol{t},\boldsymbol{e}) - \boldsymbol{d}(k))\cdot\boldsymbol{n}_k)^2 + \omega_{2D}\frac{1}{N_l}\sum_{i=1}^{N_l}\left\|\Pi_p(P_i(R,\boldsymbol{t},\boldsymbol{e})) - l_i\right\|^2. \tag{20}$$

The first term (upper part) measures the 3D registration error in terms of point-to-plane distance, where N_d is the number of correspondences and $\boldsymbol{d}(k)$ is the corresponding data point with $\boldsymbol{n}_k$ as its normal. Note that a subsampling is applied to the blendshape model to reduce the computational cost, resulting in about two thousand vertices used in the correspondence search. The second term (lower

part) measures the 2D registration error, where N_l is the number of 2D landmarks recovered by an ASM method [40], l_i is a 2D landmark, and P_i is the corresponding 3D landmark of the model which is projected onto the image plane by projection Π_p.

The optimization problem defined in Eq. (19) is non-convex and non-linear because the variable $\boldsymbol{e}$ is bounded within $[0,1]$ and the definition of E_{f_2} includes a 2D projection. To solve this problem, we resort to a standard bound and linear equality/inequality constrained (BLEIC) quadratic solver. Similar to the case of using Candide-3 model, we solve for R and $\boldsymbol{t}$ with fixed $\boldsymbol{e}$. For non-rigid registration, $\boldsymbol{e}$ is optimized with fixed R and $\boldsymbol{t}$. Note that separating the whole problem into subproblems leads to faster convergence and less computational cost. The optimization of E_{f_2} guarantees local convergence. It also decreases the distance between the point cloud and the blendshape model, therefore it minimizes the global energy $E(U,Y)$.

The overall procedures for solving the optimization problem (11) is summarized in Algorithm 1. For clarity, we differentiate between the cases of sparse (Candide-3) and dense (Blendshape) models. Note that the blendshape face model also has expression parameters. But we only focus on neutral faces for simplicity. Because both two sub-problems have convergence guarantees, the proposed optimization process can also guarantee local convergence.

Algorithm 1. The proposed solving procedures

Input: Color image I and its depth map Z, the trade-off factors λ_d and λ_f, and the stopping thresholds ϵ_1 and ϵ_2.
Output: The refined depth map U and the model parameters θ.
Initialization: $\theta_1 \leftarrow \mathbf{0}, U_0 \leftarrow \mathbf{0}, U_1 \leftarrow Z, n \leftarrow 1$;
Initial face alignment: Estimate the initial model parameters θ_0 from I and Z;
Generate the initial prior $Y_0 = T_f(P(\theta_0))$;
Compute the weights ω_i and η_i for each pixel i;
while not $(\|\theta_n - \theta_{n-1})\|_2^2 \leqslant \epsilon_1$ and $\|U_n - U_{n-1}\|_2^2 \leqslant \epsilon_2)$ **do**
  $U_n = \arg\min_U E(U, Y_{n-1})$;
  Construct a point cloud from U_n;
  if Candide-3 model is used **then**
    Solve rigid ICP for R_n, t_n, and $\boldsymbol{d}$;
    $\hat{\boldsymbol{\sigma}}_n = \arg\min_{\boldsymbol{\sigma}} E_{f_1}(\boldsymbol{\sigma})$
    $\theta_n \leftarrow \{\hat{\boldsymbol{R}}_n, \hat{\boldsymbol{t}}_n, \hat{\boldsymbol{\sigma}}_n\}$;
  else
    $\hat{\boldsymbol{R}}_n, \hat{\boldsymbol{t}}_n = \arg\min_{\boldsymbol{R},\boldsymbol{t}} E_{f_2}(\boldsymbol{R}, \boldsymbol{t})$
    $\hat{\boldsymbol{e}}_n = \arg\min_{\boldsymbol{e}} E_{f_2}(\boldsymbol{e})$
    $\theta_n \leftarrow \{\hat{\boldsymbol{R}}_n, \hat{\boldsymbol{t}}_n, \hat{\boldsymbol{e}}_n\}$;
  end if
  Update the prior $Y_n = T_f(P(\theta_n))$;
  Compute ω_i and η_i for each pixel i;
  $n \leftarrow n + 1$;
end while

4.4. Implementation details

4.4.1. Preprocessing steps

The proposed guided depth recovery assumes starting with a roughly aligned face model. To get this rough registration, two pre-processing steps are performed prior to solving the optimization problem (11), as shown in Fig. 2. In the first step, depth denoising, we use the baseline method [9] to reduce the noise of the input depth map. In the second step, we use different schemes for the alignment of different face models.

For the blendshape model, obtaining a rough alignment is relatively simple. First, we use a classical face detector [41] to detect the face and an ASM alignment algorithm [40] to extract 2D landmark points. With the pre-processed depth map, we convert these 2D landmark points to 3D points. Then, the SVD based registration method [42] is used to estimate the initial R and $\boldsymbol{t}$. According to our experience, such initial alignment is good enough for extracting useful guidances from the blendshape model. Lastly, the subject identity, which is denoted as ω_{id} in Eq. (7), must be estimated in order to prepare the person-specific expression blendshapes $\{B_j\}$ in Eq. (8). To this end, we follow the technique in [33,38] to estimate ω_{id} by matching the 3D landmarks of the blendshape to 2D ASM landmarks, which is formulated as an optimization problem.

However, the alignment of Candide-3 model obtained by the SVD-based method is not sufficiently good and sometimes even harmful for guiding depth recovery due to its limited representation capability. Fig. 5(a) shows such a case. To address this problem, we further refine the 3D face registration by using point-to-plane ICP [43] constrained by the small set of correspondences used in the initial alignment. In particular, we solve

$$\min_{R,\boldsymbol{t}} \sum_{i\in\Omega_p} \Big((R\boldsymbol{p_0}(i) + \boldsymbol{t} - \boldsymbol{d}(i))^T \boldsymbol{n}(i) \Big)^2 + w_a \sum_{j\in\Omega_s} \|R\boldsymbol{p_0}(j) + \boldsymbol{t} - \boldsymbol{d}(j)\|^2 \tag{21}$$

to update R and $\boldsymbol{t}$. The major difference between Eqs. (20) and (21) is the second term that represents the point-to-point distance function for the set of vertices used to estimate head pose by the SVD method [42]. Minimizing the second term helps prevent the 3D shape model from sliding away too much, as shown in Fig. 5(b).

After initial pose estimation, some parts of Candide-3 model may not well match that of the input data, especially for chin and nose. We address this small problem empirically by looking for these points based on our observations. Fig. 5 gives an intuitive illustration for this coarse-to-fine alignment. Note that the current implementation of our method may not work well for challenging cases such as large head rotations. Discriminative and generative techniques from face tracking may handle such cases. However, related discussions are not included because handling such cases is beyond the scope of this paper.

4.4.2. Parameter selection

To determine a reasonable value range of λ_f, we conduct small-scale experiments using the BU4DFE database and evaluate the difference between the ground truth and the depth map recovered by the proposed method. Experimental setup is identical to that described in Section 5.1. As shown in Fig. 6, the testing error firstly decreases and then increases as λ_f increases, which means that the weight of face prior should be neither too small nor too large for the optimal restoration performance. Too small λ_f makes the proposed method performs like the baseline method, while too large λ_f changes the proposed method into depth replacement where the recovered facial depth is replaced by the prior facial depth. By comparing the optimal values of λ_f indicated in Fig. 6(a)–(c), it can be found that the optimal choice of λ_f increases as the camera-object distance increases. Although the optimal values of λ_f only work for the testing data, we still adopt them in our experiments because such setup already demonstrates the benefit of the face prior.

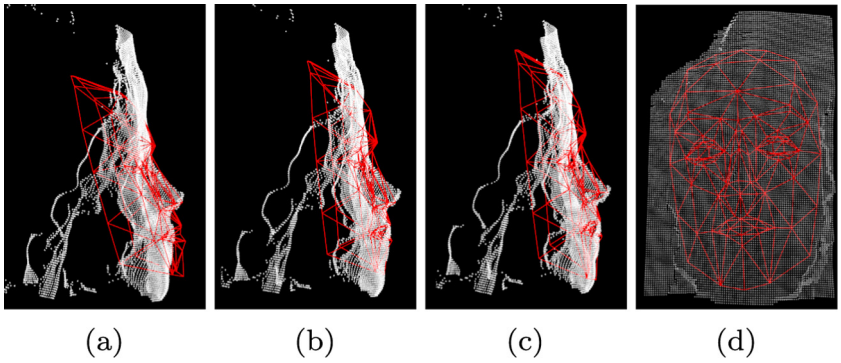


Fig. 5. The coarse-to-fine face alignment. (a) The alignment after SVD-based pose estimation. (b) The alignment refined by the point-to-plane ICP with regularization. (c) and (d) The alignment after estimating initial shape parameters.

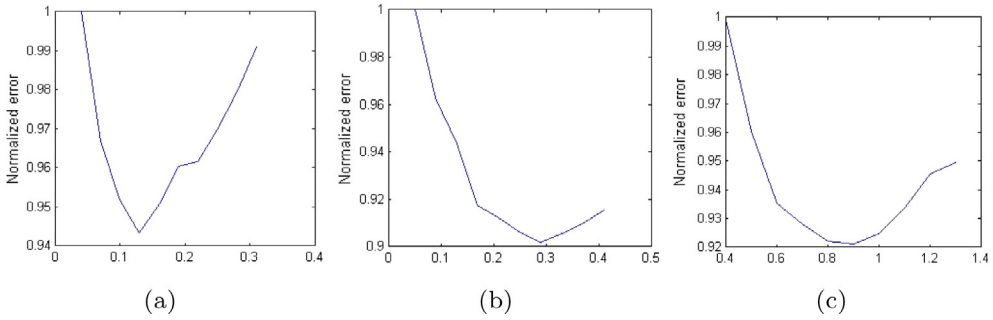


Fig. 6. The testing errors with respect to λ_f at different distances. Note that the errors are normalized w.r.t. the maximum value of error. The testing depth maps are rendered at the distances of (a) 1.5 m, (b) 1.75 m, and (c) 2.0 m.

5. Experiments

In this section, we conduct experiments to evaluate the performance of the proposed method. The priors from both sparse and dense face models are used and compared in these experiments. The BU4D Facial Expression database [44] is used for quantitative evaluations. Considering that Kinect is the most popular commodity RGB-D sensor, we add some Kinect-like artifacts according to [37] to the depth maps generated from the BU4DFE database, as described in Section 5.1. By using synthetic data, we are able to obtain the ground truth for quantitative evaluation. Section 5.2 shows the comparisons between using and not using face priors as well as the comparisons between sparse and dense face priors. In Section 5.3, we also show qualitative results on real-world data captured by Kinect-1 sensor. Since face tracking is beyond the scope of this work, we focus on a simple case that the face is always with neutral expression and mainly report the recovery performances.

5.1. Generating data for quantitative evaluations

According to [37], distance-dependent noise and quantization error are the two main characteristics of the data captured by Kinect. We simulate these two artifacts in our experiments. In particular, the distance-dependent noise is simulated by

$$Z'(i) = Z_0(i) + n(i), \tag{22}$$

where i stands for pixel index, Z_0 is the depth map generated from the face model, $n(i)$ is a random sample of a Gaussian distribution $N(0, cZ_0^2(i))$, and $c = 1.43 \times 10^{-5}$ is Kinect-oriented constant [37]. The quantization artifact is simulated by quantizing the noisy depth map using quantization steps computed from the camera parameters of Kinect. An example of the added artifacts is shown in Fig. 7.

5.2. The effectiveness of employing face priors

To show the effectiveness of our idea of utilizing the prior face information, we compare the proposed method with the baseline method [9]. For a fair comparison, the parameters l_s, l_l, l_z, and λ_d are set according to [9] for both the proposed and the baseline methods. For the proposed methods, we empirically set $\epsilon_1 = 0.5$ and $\epsilon_2 = 2$. It should be noted that the proposed method is not sensitive to these parameters because its performance remains similar when the parameters change within a reasonable range. Considering that the reliability of the input depth map decreases as the distance increases, we use a relatively small value for λ_f at close distances, and a relative large value at far distances. Specifically, λ_f ranges from 0.1 to 0.9 for both Candide-3 model and the blendshape model. The value range is determined according to our experiments on a small subset of RGB-D data. We recommend conducting small-scale experiments to determine the range of λ_f before utilizing the proposed method.

BU4DFE database contains more than 600 sequences of 3D face expressions. For the evaluation of depth recovery, we choose 266 sequences whose 1st frame is with neutral (or nearly neutral) expression and render their first frames as RGB-D data sets. Each depth map is rendered at four different camera-object distances: 1.5 m, 1.75 m, 2.0 m, and 2.3 m. Note that as the camera-object distance increases, the resolution of face region decreases. Fig. 8 shows the face region of "F001" data rendered at 4 distances with resolutions ranging from 30×40 to 50×60. According to the analysis of the weight map in Section 4.2.2, in the ideal case that each depth pixel corresponds to a vertex of the face model, for a face region of M pixels, the maximum number of required vertices is M. As shown in Fig. 7, the face region of the synthetic data has less than 3000 pixels, which means a 3D model with more than 3000 vertices for the face region is enough. The utilized blendshape have

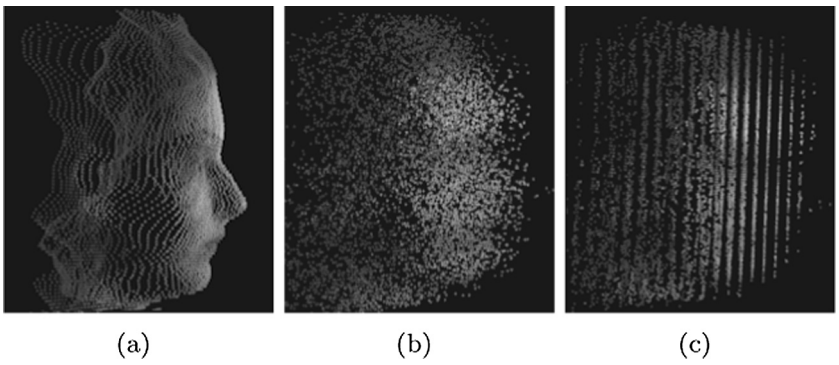


Fig. 7. An example of adding Kinect-like artifacts. (a) The noise-free depth map. (b) The depth map with distance-dependent noise. (c) The depth map with both noise and quantization error.

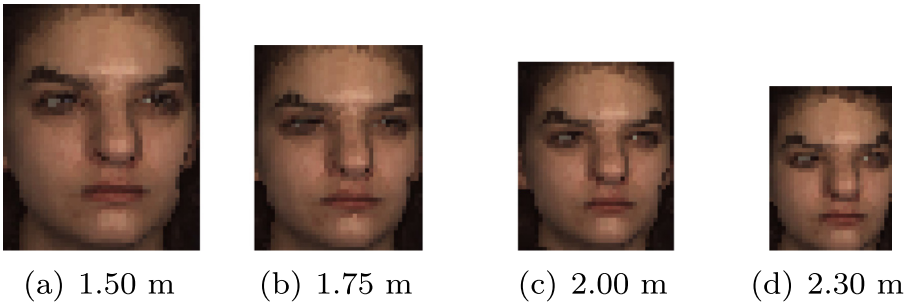


Fig. 8. Face region of "F001" data rendered at 4 distances, which are shown in their original resultions: (a) 48×60, (b) 41×50, (c) 38×46, and (d) 30×40 pixels.

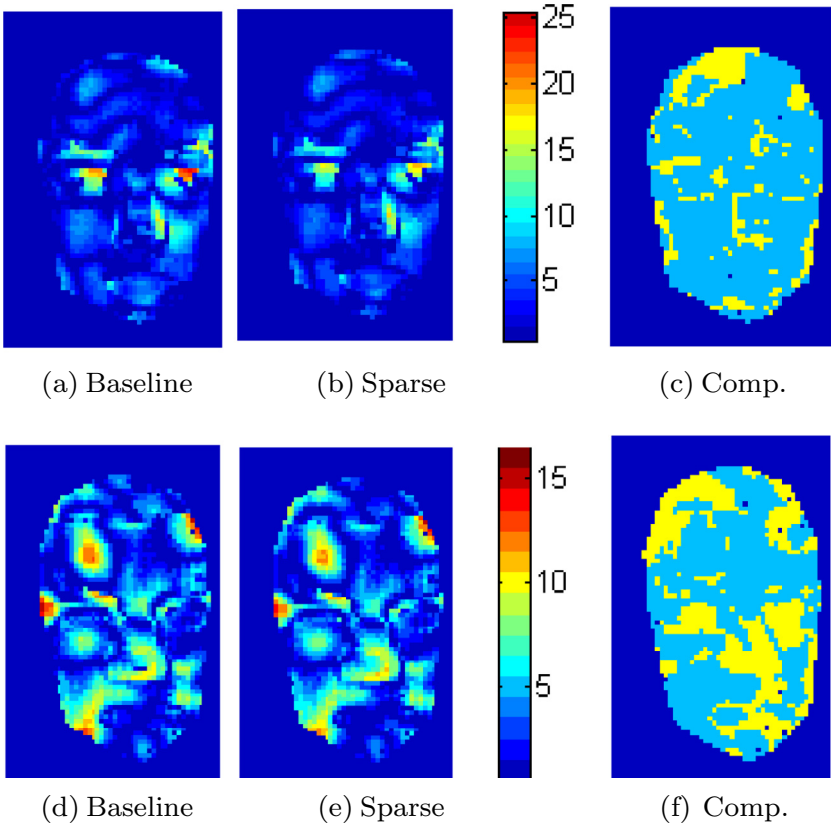


Fig. 9. Representative comparisons between the baseline method and the proposed method with sparse face priors. (a) and (d) are the difference maps for the baseline method, in which the errors are encoded by different colors. (b) and (e) are the difference maps for the proposed method with sparse face priors. (c) and (f) show the comparisons between the baseline and the proposed methods. The light blue color indicates the region where the proposed method achieves lower recovery error and the yellow color indicates the region where the baseline method is better. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

around 4000 vertices for describing the face region, which is more than enough for good depth recovery.

5.2.1. Qualitative comparisons

First, we present several qualitative (visual) comparisons between the baseline method and the proposed method with sparse face priors. The first comparison is on data sets "F005" and "M004" rendered at 1.75 m. As shown in Fig. 9(a), (b), (d) and (e), we color the differences between the recovered depth maps and the ground truths, where dark blue indicates small differences and red represents larger differences. Our results suggest that the baseline method fails to handle the case of rich textures. Some large errors around the eyes' region are specific examples for this case. In contrast, the face priors used in the proposed method can reduce such artifact and thus lead to higher recovery quality. In Fig. 9(c) and (f), we use the light blue color to represent the region where the proposed method achieves higher recovery quality, and yellow color to represent the region where the baseline method is better. The cases shown in Fig. 9 are representative for most data sets, suggesting that the proposed method generally

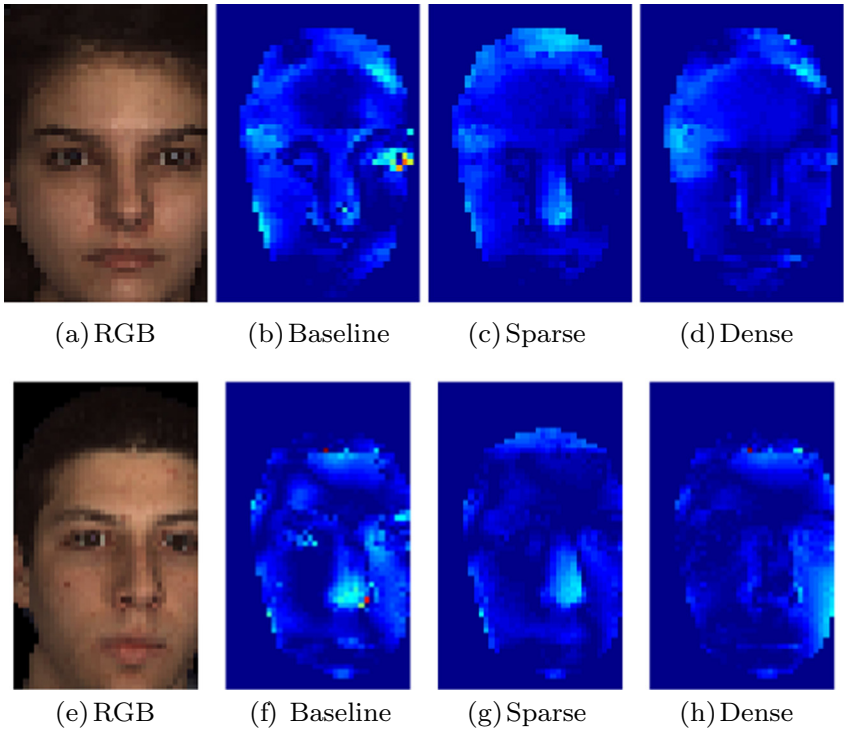


Fig. 10. Visual examples of depth recovery on the data sets "F001" and "M009" rendered at 2.00 m. (a) and (e) are the input color images of subject "F001" and "M009", respectively. (b) and (f) are the difference maps for the baseline method, where dark blue represents small errors and light blue represents large errors. (c) and (g) are the difference maps for our method with sparse face priors. (d) and (h) are the difference map for our method with dense face priors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

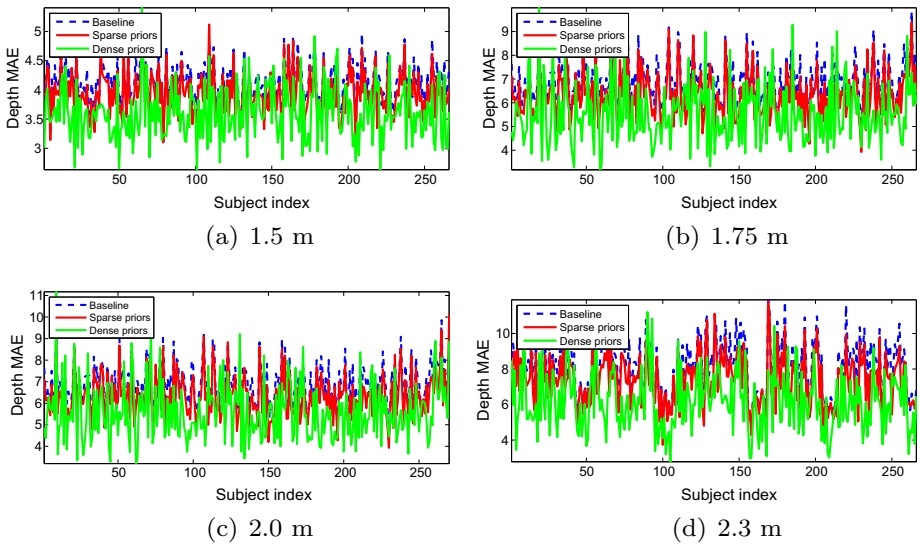


Fig. 11. The depth MAE results on data sets rendered at different camera distances, (a) 1.5 m, (b) 1.75 m, (c) 2.0 m and (d) 2.3 m.

Table 1
Quantitative comparisons among the baseline method, the proposed methods with sparse and dense face priors (in mm). The improvements of the proposed methods over the baseline method are shown in percentage points.

Distance (m)	Noise level (mm)	Baseline	Sparse priors	Dense priors
1.50	32.18	4.12	3.95 (4.2%)	3.61 (12.6%)
1.75	43.79	5.36	5.00 (6.6%)	4.50 (16.0%)
2.00	57.20	6.84	6.34 (7.3%)	5.54 (19.0%)
2.30	75.65	8.09	7.39 (8.6%)	6.16 (23.8%)

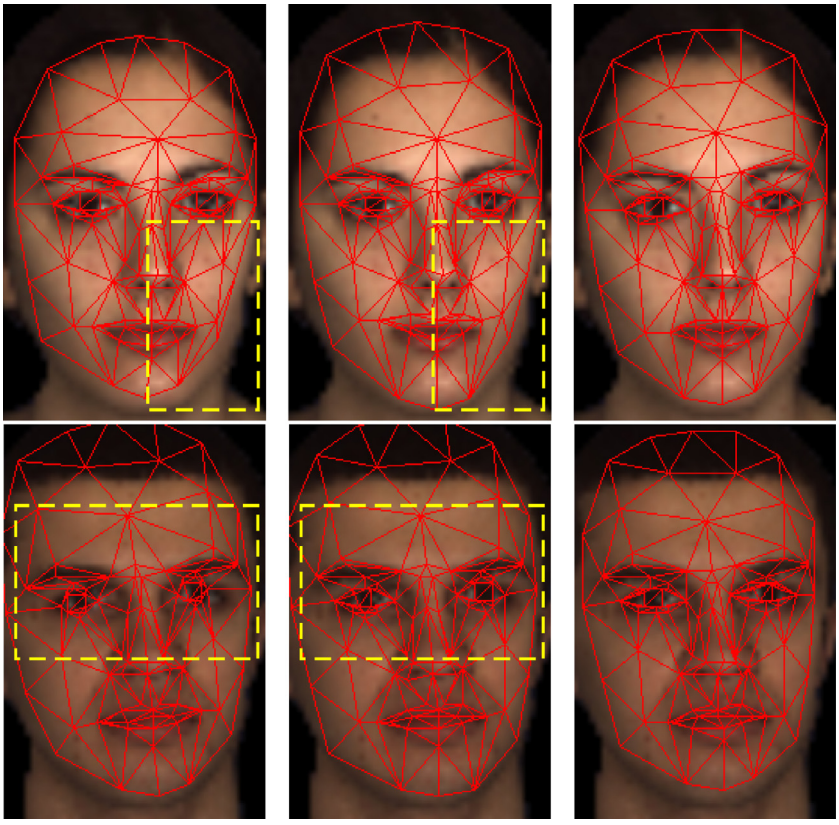


Fig. 12. Representative results of face registration on the data sets "F005" and "M004" rendered at 1.75 m. The proposed method produces a better fitting around the face boundary and the eyes' region compared to the baseline method.

achieves higher recovery quality compared to the baseline approach.

Next, we consider the impact of the dense face priors on the refinement approach. We construct difference maps to show the pixel-wise MAE of the results on two data sets "F001" and "M009" (rendered at 2.0 m), respectively. Fig. 10(b) and (f) shows that the baseline method fails to reduce the noise around the eyes and nose areas due to the complex texture and high noise level. Fig. 10(c) and (g) indicates that our method with sparse face priors can effectively reduce the noise in the region of complex texture. However, due to the piece-wise planar nature of the guidance depth map, some face details, such as the nose, are not recovered very well. Using a more geometrically detailed depth guidance, our method can recover additional facial details, including the nose structure.

5.2.2. Quantitative evaluations

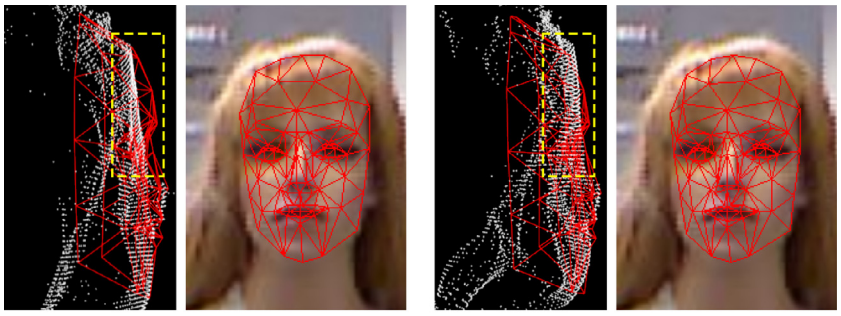
Fig. 11 shows the mean absolute error (MAE) of the recovered depth map for each data set. Since we focus on face fidelity, the MAE is computed only using the depth values inside the face region. It can be seen that, in most cases, the proposed methods achieve higher recovery accuracy compared to the baseline method, which suggests that the face prior is helping the depth recovery. In addition, these results suggest that dense face priors improve depth recovery over the sparse counterparts.

Table 1 summarizes the average MAE results, in which we also show the actual noise levels of the data sets. It can be seen that, with face priors, the improvement of recovery accuracy exhibits a generally increasing trend with the distance. This is because the quality of the input depth map keeps decreasing with the increase of the distance and the baseline method only uses the input depth map as the data term. These results demonstrate that our proposed method with dense priors can improve the recovery accuracy up to 23.8%. The improvement difference between with dense and with sparse priors is up to 15.2%.

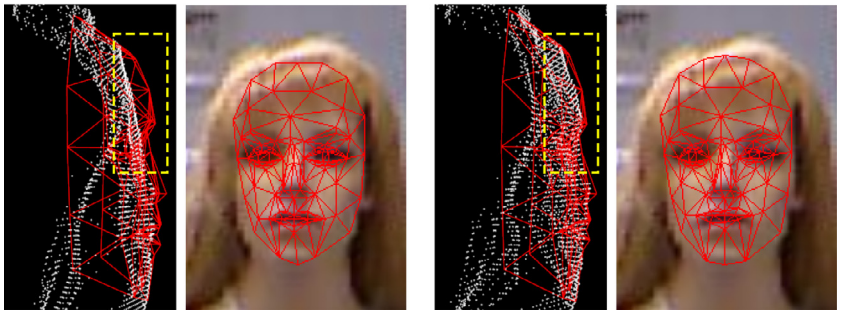
Besides the recovery error, we also evaluate the registration accuracy by comparing the baseline method with the proposed method with sparse face priors. To get the reference registration and shapes, we fit Candide-3 model to noise-free data. The face model is also fitted to the depth maps obtained by different methods. We then compare the fitting result with the reference registration. Fig. 12 gives two visual comparisons of the registra-

Table 2
Quantitative evaluations of the proposed method using 4 metrics of registration accuracy. The results obtained by the baseline and proposed methods are separated by "/". The improvement of the proposed method over the baseline method is shown in percentage points.

Distance (m)	Overall 2D error (px)	2D translation error (px)	2D landmark error (px)	3D shape error (mm)
1.50	1.04/ 0.89 (13.9%)	0.59/ 0.51 (12.5%)	0.86/ 0.74 (14.0%)	2.06/ 1.79 (13.1%)
1.75	1.01/ 0.87 (13.7%)	0.58/ 0.49 (15.3%)	0.82/ 0.72 (12.6%)	2.26/ 1.99 (12.0%)
2.00	1.12/ 0.93 (17.4%)	0.56/ 0.45 (19.5%)	0.96/ 0.80 (16.8%)	2.89/ 2.42 (16.5%)
2.30	1.14/ 0.99 (13.3%)	0.61/ 0.53 (12.2%)	0.95/ 0.82 (14.0%)	3.08/ 2.54 (17.5%)

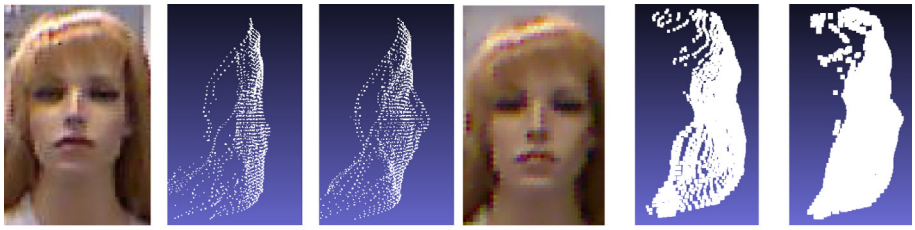


(a) Mannequin (1.50 m): baseline/ours

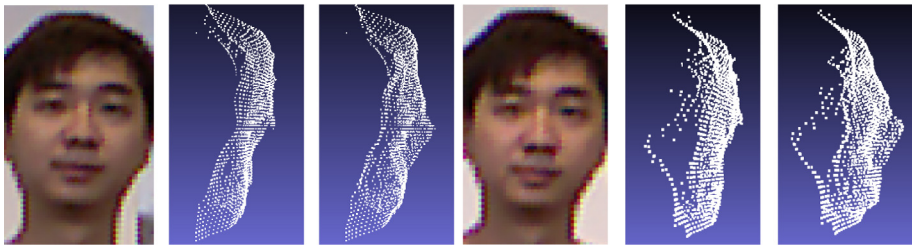


(b) Mannequin (1.68 m): baseline/ours

Fig. 13. The results on real Kinect data of a mannequin, which are shown in both 2D and 3D. In each figure, the result of the baseline method is on the left-hand side, while the result of the proposed method with Candide-3 model is on the right-hand side.



(a) Mannequin (1.6 m) (b) Mannequin (2.0 m)

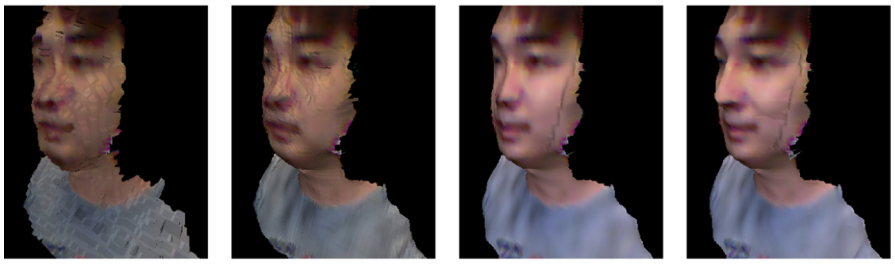


(c) Male (1.6 m) (d) Male (2.0 m)

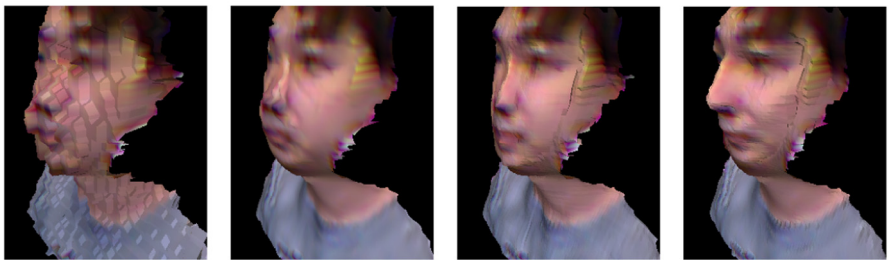
Fig. 14. The results of our methods with sparse and dense priors on real Kinect data sets. For each comparison, the color image, point cloud recovered by utilizing sparse priors, and point cloud recovered by utilizing dense priors are placed from left to right.

(a) Mannequin (1.6 m): raw/baseline/sparse/dense

(b) Mannequin (2.0 m): raw/baseline/sparse/dense



(c) Male (1.6 m): raw/baseline/sparse/dense



(d) Male (2.0 m): raw/baseline/sparse/dense

Fig. 15. The results of different methods on real Kinect data sets.

tion results on data sets. We can see that the proposed method produces a more accurate face registration compared to the baseline method, especially in the eyes' region and around the face boundary.

Quantitative evaluations of the registration accuracy are shown in Table 2. Four metrics are used to compute the registration error. Given two sets of 2D points, i.e., the estimated points $P_{\text{fit}}^{\text{2D}}$ and the ground-truth points $P_{\text{ref}}^{\text{2D}}$, the overall 2D error is computed as $\|P_{\text{fit}}^{\text{2D}} - P_{\text{ref}}^{\text{2D}}\|/N_v$, where $N_v = 113$ is the number of vertices in Candide-3 model. Considering that there may be a global misalignment between these two point sets, the 2D landmark error is computed after aligning their center. The Euclid distance between these two centers is denoted as 2D translation error and shown in Table 2. After aligning the centers of the two 2D face models, we compute the mean squared error between the 2D landmarks of the fitted model and those of the reference model and denote it as 2D landmark error. The 3D shape error is computed by scaling the difference between 3D models, i.e., $\|P_{\text{fit}}^{\text{3D}} - P_{\text{ref}}^{\text{3D}}\|/N_v$, where $P_{\text{fit}}^{\text{3D}}$ represents the 3D points of Candide-3 that fits to the recovered depth map and $P_{\text{ref}}^{\text{3D}}$ represents that fits to the noise-free depth map. It can be seen that the improvement of the proposed method in registration accuracy is significant, up to 17.5%. It indicates that a better recovered depth map is helpful for the face alignment.

5.3. Experiments on real data

For the experiments on real data, we use Kinect-1 to capture several RGB-D frames of a mannequin and a male subject at distances ranging from 1.0 m to 2.0 m. The results of the proposed and the baseline methods are then visually compared because we do not have the ground truth.

5.3.1. Depth recovery with sparse face prior

Fig. 13 shows the registration (red[1] wireframe) and depth recovery (white cloud) results of the mannequin at distances 1.50 m and

[1] For interpretation of color in 'Fig. 13', the reader is referred to the web version of this article.

1.68 m, where Candide-3 model is used. It is shown that the proposed method clearly outperforms the baseline method. The depth maps in these tests were captured at a relatively large distance from the sensor and, as a consequence, using the baseline alone is insufficient to reconstruct the depth maps properly. Specifically, the depth maps recovered by the baseline method are flat on the upper half of the face in Fig. 13, mainly at the forehead and noseline areas. On the other hand, the proposed method guided by the sparse face priors is able to reconstruct more reasonable depth maps in those areas, e.g., following the natural shape of the forehead. This also affects the final registration quality, although to a somewhat lesser extent than in the BU4DFE synthetic data. We attribute this to the discrepancy between the depth noise model used in BU4DFE experiments and that in the real data, as well as the additional noise in the color channels.

5.3.2. 3D reconstruction with face priors

To compare our methods using sparse and dense face priors, we reconstruct the 3D models of both the mannequin and the male subject. The largest camera-face distance in this set is 2.0 m, both for the mannequin and the male subject. Fig. 14 shows the reconstruction results in the form of point clouds. The results clearly indicate that the dense priors lead to improved fidelity of the reconstruction. At large camera-face distances, the depth maps recovered with the sparse face priors become piece-wise flat, while the depth maps recovered with the dense face priors preserve the face details, especially around the nose area. This is because the deformation of Candide-3 model is constrained only by a set of sparse vertices, which is relatively loose. The blendshape model follows stricter constraints and thus preserves sufficient facial details to serve as a good guidance.

3D models can be easily reconstructed from the recovered depth maps by projecting 2D pixels to 3D points and then connecting neighboring points as triangles. Therefore, we additionally render the depth maps and color images as textured meshes, depicted in Fig. 15. These results indicate that our method using the dense face priors produces improved rendering images at novel views while preserving the face geometry. These results also inspires us that the proposed method can be used to generated a user-like 3D model with deformation parameters identical to pre-defined 3D models, which sheds lights on non-rigid 3D model understanding and facial expression transfer researches.

It should be noted that the presented experiments only contain the simple case of neutral face. When there are dynamic face expressions, a more powerful face tracker should be used.

6. Conclusion

In this work, we review our preliminary idea of employing face priors in depth recovery [19] from the aspect of depth recovery for non-rigid objects, and present a detailed study on it w.r.t. model density, parameter selection, and thorough evaluations on benchmark data sets. In particular, two representative face models are utilized and compared to better understand the model density to the recovery accuracy. Experimental results on a benchmark dataset show that, even for the coarse and sparse face model, properly taking into account the face priors brings in up to 8.6% of improvement in depth quality. Such improvement is further enlarged to 23.8% when a dense face model is used, which can be essential for applications such as 3D telepresence and teleconference. Moreover, the proposed method also leads to better registration accuracy, with up to 17.5% of improvement, suggesting its possible role in helping other high-level face analysis tasks.

As the starting point of non-rigid prior guided depth recovery, there are many possible avenues for broadening this work. Utilizing RGB images of higher resolution, dividing the face into different parts, and extensions to other types of high-level priors would be nice choices. Note that this work still has limitations, two of which may be the lack of considerations of temporal consistency and dynamic face priors. For face depth recovery on RGB-D videos, considering these two factors would be more helpful.

Acknowledgement

This research is partially by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office and the NSF Project CNS-1229628, USA. This work is also partially supported by the National Natural Science Foundation of China under Grant 61602533, The Fundamental Research Funds for the Central Universities, the 111 Project (No. B07048), and Science and Technology Planning Project of Guangdong Province (No. 2016A020210086), China.

References

[1] C. Mutto, P. Zanuttigh, G. Cortelazzo, Microsoft Kinect™ range camera, in: Time-of-Flight Cameras and Microsoft Kinect™, SpringerBriefs in Electrical and Computer Engineering, Springer, US, 2012, pp. 33–47.
[2] A. Maimone, H. Fuchs, Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras, in: Int'l Symposium Mixed Augmented Reality (ISMAR), IEEE, Basel, Switzerland, 2011, pp. 137–146.
[3] C. Kuster, T. Popa, C. Zach, C. Gotsman, M. Gross, Freecam: a hybrid camera system for interactive free-viewpoint video, in: Proc. Vision, Modeling, and Vis. (VMV), Berlin, Germany, 2011, pp. 17–24.
[4] C. Zhang, Q. Cai, P. Chou, Z. Zhang, R. Martin-Brualla, Viewport: a distributed, immersive teleconferencing system with infrared dot pattern, IEEE Multimedia 20 (1) (2013) 17–27, http://dx.doi.org/10.1109/MMUL.2013.12.
[5] D. Min, J. Lu, M. Do, Depth video enhancement based on weighted mode filtering, IEEE Trans. Image Process. 21 (3) (2012) 1176–1190.
[6] C. Richardt, C. Stoll, N.A. Dodgson, H.-P. Seidel, C. Theobalt, Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos, Comp. Graph. Forum 31 (2pt1) (2012) 247–256.
[7] F. Qi, J. Han, P. Wang, G. Shi, F. Li, Structure guided fusion for depth map inpainting, Pattern Recogn. Lett. 34 (1) (2013) 70–76.
[8] J. Yang, X. Ye, K. Li, C. Hou, Color-guided depth recovery from RGB-D data using an adaptive autoregressive model, IEEE Trans. Image Process. 23 (8) (2014) 3443–3458.
[9] C. Chen, J. Cai, J. Zheng, T.J. Cham, G. Shi, Kinect depth recovery using a color-guided, region-adaptive, and depth-selective framework, ACM Trans. Intell. Syst. Technol. 6 (2) (2015) 1–19.
[10] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, A.M. Bruckstein, RGBD-fusion: real-time high precision depth recovery, in: Computer Vision and Pattern Recognition, 2015, pp. 5407–5416.
[11] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1063–1074.
[12] F.B. ter Haar, R.C. Veltkamp, 3D face model fitting for recognition, in: Computer Vision–ECCV 2008, Springer, 2008, pp. 652–664.
[13] C. Ding, C. Xu, D. Tao, Multi-task pose-invariant face recognition, IEEE Trans. Image Process. 24 (3) (2015) 980.
[14] T. Weise, S. Bouaziz, H. Li, M. Pauly, Realtime performance-based facial animation, in: SIGGRAPH, 2011.
[15] H. Li, J. Yu, Y. Ye, C. Bregler, Realtime facial animation with on-the-fly correctives, ACM Trans. Graph. 32 (4) (2013) 42:1–42:10.
[16] O. Aldrian, W.A.P. Smith, Inverse rendering of faces with a 3D morphable model, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1080–1093.
[17] J. Heo, M. Savvides, Gender and ethnicity specific generic elastic models from a single 2D image for novel 2D pose face synthesis and recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (12) (2012) 2341–2350.
[18] Z.H. Feng, G. Hu, J. Kittler, W. Christmas, X.J. Wu, Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting, IEEE Trans. Image Process. 24 (11) (2015) 3425–3440, http://dx.doi.org/10.1109/TIP.2015.2446944.
[19] C. Chen, H.X. Pham, V. Pavlovic, J. Cai, G. Shi, Depth recovery with face priors, in: Computer Vision – ACCV 2014, Lecture Notes in Computer Science, vol. 9006, Springer International Publishing, Singapore, 2015, pp. 336–351.
[20] C. Tomasi, R. Manduchi, Bilateral filtering for gray and color images, in: Int'l Conf. Comput. Vision (ICCV), IEEE, Bombay, India, 1998, pp. 839–846.
[21] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, K. Toyama, Digital photography with flash and no-flash image pairs, ACM Trans. Graph. 23 (3) (2004) 664–672.
[22] S.-W. Jung, Enhancement of image and depth map using adaptive joint trilateral filter, IEEE Trans. Circ. Syst. Video Technol. 23 (2) (2013) 258–269, http://dx.doi.org/10.1109/TCSVT.2012.2203734.

[23] K.A. Ismaeil, D. Aouada, T. Solignac, B. Mirbach, B. Ottersten, Real-time enhancement of dynamic depth videos with non-rigid deformations, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2016), http://dx.doi.org/10.1109/TPAMI.2016.2622698, pp. 1–1.
[24] T. Cootes, C. Taylor, D. Cooper, J. Graham, Active shape models – their training and applications, Comput. Vis. Image Underst. (61) (1995) 39–59.
[25] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. (23) (2001) 681–684.
[26] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.
[27] T. Baltruaitis, P. Robinson, I. Matthews, L.-P. Morency, 3D constrained local model for rigid and non-rigid facial tracking, in: CVPR, 2012, pp. 2610–2617.
[28] H. Wang, A. Dopfer, C. Wang, 3D AAM based face alignment under wide angular variations using 2D and 3D data, in: ICRA, 2012.
[29] Z. Zhang, Iterative point matching for registration of free-form curves and surfaces, Int. J. Comput. Vis. 13 (2) (1994) 119–152.
[30] B. Welsh, Model-based Coding of Images Ph.D. thesis, British Telecom Research Lab, January 1991.
[31] J. Ahlberg, An Updated Parameterized Face Tech. Rep., Image Coding Group, Dept. of Electrical Engineering, Linkoping University, 2001.
[32] Q. Cai, D. Gallup, C. Zhang, Z. Zhang, 3D deformable face tracking with a commodity depth camera, in: Europ. Conf. Comput. Vision (ECCV), 2010.
[33] C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou, FaceWarehouse: a 3D facial expression database for visual computing, IEEE Trans. Visual. Comput. Graph. 20 (3) (2014) 413–425.
[34] C. Ding, D. Tao, A comprehensive survey on pose-invariant face recognition, ACM Trans. Intell. Syst. Technol. 7 (3) (2016) 37.
[35] J. Heo, 3D Generic Elastic Models for 2D Pose Synthesis and Face Recognition Ph.D. thesis, Carnegie Mellon University, 2010.
[36] S. Bouaziz, Y. Wang, M. Pauly, Realtime Facial Animation with On-the-fly Correctives, in: SIGGRAPH, 2013.
[37] K. Khoshelham, S.O. Elberink, Accuracy and resolution of Kinect depth data for indoor mapping applications, Sensors 12 (2) (2012) 1437–1454.
[38] C. Cao, Y. Weng, S. Lin, K. Zhou, 3D shape regression for real-time facial animation, in: SIGGRAPH, 2013.
[39] H.X. Pham, V. Pavlovic, Hybrid on-line 3D face and facial actions tracking in RGBD video sequences, in: Proc. International Conference on Pattern Recognition (ICPR), 2014.
[40] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215, http://dx.doi.org/10.1007/s11263-010-0380-4.
[41] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: CVPR, 2001, pp. I-511–I-518.
[42] K.S. Arun, T.S. Huang, S.D. Blostein, Least-squares fitting of two 3D point sets, IEEE Trans. Pattern Anal. Mach. Intell. 9 (5) (1987) 698–700.
[43] K. Low, Linear Least-squares Optimization for Point-to-plane ICP Surface Registration Tech. Rep. TR04-004, Department of Computer Science, University of North Carolina at Chapel Hill, 2004.
[44] L. Yin, X. Chen, Y. Sun, T. Worm, M. Reale, A high-resolution 3D dynamic facial expression database, in: 8th IEEE International Conference on Automatic Face Gesture Recognition, IEEE, 2008, pp. 1–6, http://dx.doi.org/10.1109/AFGR.2008.4813324.