Depth Recovery with Face Priors

Chongyu Chen[1,2], Hai Xuan Pham[3], Vladimir Pavlovic[3],
Jianfei Cai[4], and Guangming Shi[1]

[1] School of Electronic Engineering, Xidian University, China
[2] Institute for Media Innovation, Nanyang Technological University, Singapore
[3] Department of Computer Science, Rutgers University, USA
[4] School of Computer Engineering, Nanyang Technological University, Singapore

Abstract. Existing depth recovery methods for commodity RGB-D sensors primarily rely on low-level information for repairing the measured depth estimates. However, as the distance of the scene from the camera increases, the recovered depth estimates become increasingly unreliable. The human face is often a primary subject in the captured RGB-D data in applications such as the video conference. In this paper we propose to incorporate face priors extracted from a general sparse 3D face model into the depth recovery process. In particular, we propose a joint optimization framework that consists of two main steps: deforming the face model for better alignment and applying face priors for improved depth recovery. The two main steps are iteratively and alternatively operated so as to help each other. Evaluations on benchmark datasets demonstrate that the proposed method with face priors significantly outperforms the baseline method that does not use face priors, with up to 15.1% improvement in depth recovery quality and up to 22.3% in registration accuracy.

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 or 3D teleconferencing [2–4]. However, the depth measurements provided by commodity RGB-D sensors are far from perfect and often contain severe artifacts such as noise and holes. In order to combat these artifacts, several methods [5–10] have been proposed to recover the depth from commodity RGB-D sensors. The common idea of these methods is to make use of spatial consistency in the depth map, temporal consistency or the guidance from the aligned color image.

RGB-D sensors are often used in human-related applications such as teleconference, where the human face is the common focus of attention. Modeling the human face is also central to other application such as face detection, face recognition, and face tracking [11, 12]. Accurate face reconstruction critically depends on the quality of the measured depth and texture data. In the case of face modeling, the space of 3D face shapes is highly restrictive and can serve as

an important guidance to improve the depth reconstruction. To the best of our knowledge, we are not aware of any existing work that utilizes face priors (or in general high-level semantic prior) in the depth recovery process. Incorporating a face prior can play significant role for depth recovery, especially at large camera-object distance. This is because the depth quality rapidly deteriorates as the face-camera distance increases, which makes the depth-based face reconstruction challenging. Using face priors could therefore extend the domain of the depth-based face reconstruction and analysis beyond the current limited camera ranges.

In this work, we propose to incorporate a general sparse 3D face model for depth recovery. However, it is non-trivial to derive effective face prior information for depth recovery from the sparse 3D model. Several important challenges arise in this context. On one hand, the 3D model needs to be deformed to align it with the input RGB-D data. Nevertheless, accurate alignment is hard to achieve due to the heterogeneous, quantized noise in the input data. On the other hand, if the alignment is not accurate, the extracted face priors might provide wrong guidance to the depth recovery. In addition, the 3D model is often sparse to support tractable computation, while the depth recovery requires a dense guidance. To address all these issues, we propose a joint optimization framework to iteratively and alternatively refine the depth and the face alignment that will, while reinforcing each other, lead to improved depth recovery and model registration. Extensive results show that our method with face priors clearly outperforms the baseline method that does not utilize face priors.

The rest of this paper is organized as follows. Section 2 presents the existing works related to the proposed method, including the depth recovery framework and the 3D face model used in this paper. Section 3 describes the technical details of the proposed method. Finally, Section 4 shows the experimental results and Section 5 concludes this paper.

2 Background

In this section we present a baseline model for depth recovery based on a global energy minimization framework and also discuss a sparse 3D deformable shape prior model. These models form the basis of our joint sparse prior-guided depth recovery framework, which will be discussed in Section 3.

2.1 Depth recovery framework

For depth recovery, several global approaches based on convex optimizations [7, 9] have been proposed in recent years, which achieve improved recovery accuracy compared to the local approaches based on filtering techniques [13–15, 5]. In this paper, we use a simplified version of the depth recovery framework proposed by Chen et al. [9] as the baseline method due to its general form and practical effectiveness.

In [9], depth recovery is formulated as an energy minimization problem. Given a color image I and its corresponding (noisy) depth map Z, 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

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

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. To be consistent with the empirical accuracy model of Kinect depth measurements [16], 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 for Kinect [1]. The weight α_{ij} is designed according to the color and depth similarities between pixel i and pixel j (please refer to [9] for details).

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

2.2 Face Shape Model and Its Deformation

Statistical models such as Active Shape Models (ASMs) [17] and Active Appearance Models (AAMs) [18, 19] have become a common and effective approach to face modeling for the purpose of face pose, shape or deformation estimation and tracking. These methods were originally designed to work with monocular color image input, and there have been efforts to incorporate the depth data into these techniques, such as the work in [20] where the authors utilized the depth frame as an additional texture to the traditional color texture in their ASM framework. In [21], the author extend the AAM framework by fitting the 3D shape to the point cloud using the Iterative Closest Point (ICP) procedure separately after

each AAM optimization iteration. The biggest disadvantage of these methods is the fact that their performance depends heavily on the data which they learn the statistical models from.

Another approach for face modeling is to use 3D deformable models as in [22–27]. In these works, the 3D face model is controlled by a set of static shape deformation units (SUs) and action deformation units (AUs). SUs represent the face biometry of an individual, whereas facial expressions are modeled by action units, which are person-independent.

One such model is Candide-3, a generic wireframe model (WFM) developed by J. Ahlberg [28]. The Candide-3 WFM is a sparse model, consisting of 113 vertices and 184 triangles constructed from these vertices that define its surface, as shown in Fig. 2 (a). Every vertex $p(k) \in \Re^3$, $k \in \Omega_p$ (e.g., $\Omega_p = \{1, \ldots, 113\}$), of the 3D shape model is formed according to a low-dimensional subspace model:

$$p(k) = p_0(k) + S(k)\sigma + A(k)\alpha, \tag{5}$$

where $p_0(k)$ are the base coordinates of the vertex (corresponding to a reference neutral expression face), $S \in Re^{3\times K_S}$ and $A \in Re^{3\times K_A}$ are, respectively, the shape and action deformation bases (matrices) associated with the vertex. $\sigma \in \Re^{K_S}$ is the vector of shape deformation parameters and $\alpha \in \Re^{K_A}$ is the vector for action deformation parameters. For the Candide-3 model, $K_S = 14$ and $K_A = 7$. In this work, without loss of generality, we are only interested in the static shape deformation under the neutral face expression ($\alpha = 0$). Thus, the general transformation of a vertex given global rigid rotation R and translation t is defined as:

$$p(k) = R(p_0(k) + S(k)\sigma) + t. \tag{6}$$

The geometry of the model is therefore determined by the base (average) shape p_0, the models of deformation S, and is parameterized by the (rigid and non-rigid) deformation vector $u = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \sigma^T]^T$, where θ_x, θ_y, θ_z are three rotation angles of R, t_x, t_y, t_z are three translation values corresponding to three axes x, y and z. In the rest of this work, for brevity, we will use notation P to denote all vertices $p(k)$ in a shape $P = \{p(k), k \in \Omega_p\}$ and will often write $P = P(u) = R(P_0 + S\sigma) + t$ to indicate the full deformed 3D shape according to the model in (6).

A number of other 3D deformable face modeling approaches have refined this model, e.g., blendshape-based models using an interactive (bilinear) SU/AU composition [12]. Nevertheless, Candide-3 model has the benefit of being sparse and simple, yet general enough, thus lowering the computational burden while still being able to serve as a shape *prior* in the depth recovery process. For instance, a similar model was used in Pham et al. [29] who introduced an on-line tracking framework based on Candide-3, which operates on RGB-D streams. Their tracker performs acceptably even on low quality input, for examples when the point cloud is sparse, the texture and/or point cloud is noisy. We therefore restrict our attention to this spare family of 3D face shape models for our problem at hand.

In this work we extend the framework of [29] to include depth refinement in the initialization pipeline, thus improve the overall performance of both depth recovery and shape model fitting to a static neutral pose of test subject.

3 Proposed Method

Given a color image I and its corresponding (aligned) noisy depth map Z as input, our goal is to obtain a good depth map of the face region using the face priors derived from the general 3D deformable model. The pipeline of the proposed method is shown in Fig. 1. The first two components in Fig. 1 are

Fig. 1. The pipeline of the proposed method.

pre-processing steps to roughly clean up the depth data and roughly align the general face model to the input point cloud. The last two components in Fig. 1 are the core of our proposed framework. For 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.

3.1 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 by solving the following optimization problem:

$$\min_{U,u} E_r(U) + \lambda_d E_d(U) + \lambda_f E_f(U,u), \tag{7}$$

where u represents the parameters of the face model defined in Sec. 2.2, 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 defined in Eq. (3), we use the weights α_{ij}:

$$\alpha_{ij} = \frac{\beta_{ij}}{\sum_{j\in\Omega_i}\beta_{ij}}, \tag{8}$$

with

$$-\log\beta_{ij} \propto \frac{\|i-j\|^2}{2l_s^2} + \frac{\|I(i)-I(j)\|^2}{2l_I^2} + \frac{(Z(i)-Z(j))^2}{2l_z^2}, \tag{9}$$

which are essentially the weights used in joint trilateral filtering [15], with l_s, l_I, and l_z the lengthscale constants.

We define the novel face prior E_f term as

$$E_f(U,u) = \sum_{i \in \Omega_f} \eta_i \left(U(i) - T_f(P(u),i)\right)^2, \tag{10}$$

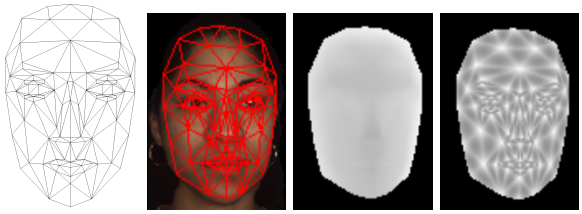
where Ω_f is the set of pixels with the face prior, and T_f is a function that transforms the sparse face model P defined through a latent deformation u to a dense depth map compatible with U. This term is critical to the recovery process and we will describe it in detail in the next section.

3.2 Shape Prior for Depth Recovery

Considering that the guidance from the sparse vertices of the Candide model may be too weak to serve as the prior for the full (dense) depth map U, we need to generate a dense synthetic depth map Y from the aligned face prior $P(u)$ 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 surface smoothness. Here, for the purpose of a shape prior we choose a simple piece-wise linear interpolation. This process is denoted as

$$Y = \text{lerp}(P(u)). \tag{11}$$

Fig. 2 shows an example of the generated dense depth map from the sparse shape P.



(a) Candide (b) $P(u)$ (c) Y (d) η_i

Fig. 2. The synthetic depth map Y and its weights $\eta_i (i \in \Omega_f)$. (a) The base shape of Candide-3 Wireframe Model. (b) The 3D wireframe model $P(u)$ drawn upon the texture frame. (c) The synthetic depth map Y generated from the 3D wireframe model. (d) The weights distribution associated with the synthetic dense depth map, where brighter means a larger weight.

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 (10).

In particular, for each pixel $Y(i)$, we use a normalized weight that is 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

$$\eta_i = \sqrt{a_i^2 + b_i^2 + c_i^2},\ i \in \Omega_f. \tag{12}$$

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. 2(c), where bright pixels represent large weights.

3.3 Energy Optimization

From the definitions of the energy functions in Eq. (7), it can be seen that the overall optimization of U remains a convex task, for a given fixed prior P. However, the optimization of the face model parameter set u might not be convex since it involves rigid and non-rigid deformation. Therefore, to tackle the global optimization task which includes both the depth U and the deformation u recovery, we resort to a standard recursive alternate optimization process. In other words, we will first optimize u while keeping U fixed, and then optimize U for the fixed deformation u.

Specifically, we divide problem (7) into three well studied subproblems: depth recovery, rigid registration, and non-rigid deformation. The subproblem of depth recovery is solved with fixed u,

$$\hat{U} = \arg\min_{U} E_r(U) + \lambda_d E_d(U) + \lambda_f E_f(U). \tag{13}$$

With U now fixed, the rigid registration between the shape prior P and the point cloud U is solved by an ICP approach, while the non-rigid deformation of the face model is found by solving

$$\hat{\sigma} = \arg\min_{\sigma} E_f(\sigma), \tag{14}$$

where σ represents the shape unit (SU) parameters of Candide model.

We here assume that the solution to (14) is only related to the sparse face vertices P and is therefore independent of the interpolation process or the pixel-wise weights. Therefore, the optimal σ can be obtained by solving

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{k \in \Omega_p} \left(R(p_0(k) + S(k)\sigma) + t - V(k)\right)^2, \tag{15}$$

where V represents the points in the input point cloud that correspond to the model vertices. The correspondences are found by a point-to-point ICP. The overall optimization procedure for solving (7) is summarized in Algorithm 1.

ALGORITHM 1: The proposed solving procedures

Input: Color image I and its corresponding depth map Z of the user's face, 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 u which consists of rotation angles θ, translation vector t, and SU parameters σ.
Preparation:
Estimate the initial model parameters θ_0, t_0, and σ_0 from I and Z;
Compute the weights ω_i and η_i for each pixel i;
$u_0 \leftarrow [\theta_0, t_0, \sigma_0^T]^T$, $\sigma_1 \leftarrow \mathbf{0}$, $U_0 \leftarrow Z$, $U_1 \leftarrow \mathbf{0}$, $n \leftarrow 1$;
while *not* $(\|\sigma_n - \sigma_{n-1})\|_2^2 \leq \epsilon_1$ *and* $\|U_n - U_{n-1}\|_2^2 \leq \epsilon_2)$ **do**
  $U_n = \arg\min_U E_r(U) + \lambda_d E_d(U) + \lambda_f E_f(U, u_{n-1})$;
  Construct a point cloud from U_n;
  Solve ICP for θ_n (i.e. R_n), t_n and the correspondences V;
  $\sigma_n = \arg\min_\sigma \|R_n(P_0 + S\sigma) + t_n - V\|^2$;
  $u_n \leftarrow [\theta_n, t_n, \sigma_n^T]^T$;
  $n \leftarrow n + 1$;
end

3.4 Implementation

The proposed guided depth recovery assumes starting with a roughly aligned face model. To get this rough registration, several pre-processing steps are added before solving the optimization problem (7), as shown in Fig. 1. For depth denoising, we use the baseline method [9] to reduce the noise of the input depth map.

The preparation step shown in Algorithm 1 is a coarse-to-fine procedure. In the coarse alignment, we use a classical face detector [30] to detect the face and an ASM alignment algorithm [31] to extract 2D landmark points. After we convert these 2D landmark points to 3D points according to the pre-processed depth map, the SVD-based registration method [32] is used to roughly align the Candide-3 model to the RGB-D data. An example of such coarse alignment is shown in Fig. 3 (a). Then, in the refining alignment, the small set of correspondences from the coarse alignment is used as regularization to the standard point-to-plane ICP optimization [33]. In particular, we solve

$$\min_{R,t} \sum_{i \in \Omega_p} \Big((Rp_0(i) + t - d(i))^T n(i) \Big)^2 + w_a \sum_{j=1}^{6} \|Rp_0(j) + t - d(j)\|^2 \qquad (16)$$

to update rotation R and translation t, where $d(i)$ is the correspondence from the data point cloud for vertex $p(i)$ and $n(i)$ is the normal vector at $d(i)$. The first term in (16) is the point-to-plane distance function of all vertices of the 3D shape; minimizing this energy function has the effect of sliding the wireframe model over the surface of the data point cloud. The second term in (16) is the point-to-point distance function of six anchor point pairs as in [29] with weight w_a, where the six anchor points are the six eye and mouth corner vertices of the

Candide model, and $d(j)$ are kept fixed as the six correspondences of eye and mouth corners used in the previous SVD-based registration step. Minimizing the second term helps prevent the shape model from moving away too much. At the end, some heuristics on search for the corresponding nose and chin tips are performed to estimate initial values of shape deformation parameters σ_0 before entering the main optimization loop. Fig. 3 gives an intuitive illustration for this coarse-to-fine alignment.

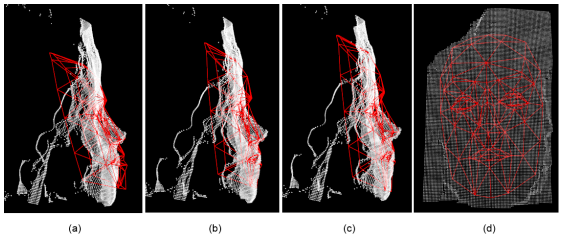


Fig. 3. The coarse-to-fine face alignment. (a) The alignment after SVD-based pose estimation. (b) The alignment refined by the point-to-plane ICP with regularization. (c) and (d) The alignment after estimating initial shape deformation parameters.

4 Experiments

In this section, we conduct experiments to evaluate the performance of the proposed method. We first use the BU4D Facial Expression Database [34] for quantitative evaluation. Considering that Kinect is the most popular commodity RGB-D sensor, we add some Kinect-like artifacts according to [16] to the depth maps generated from the BU4D database. By using synthetic data, we are able to obtain the ground truth for quantitative evaluation. We also show qualitative results on a real-world data captured by Kinect sensor [5].

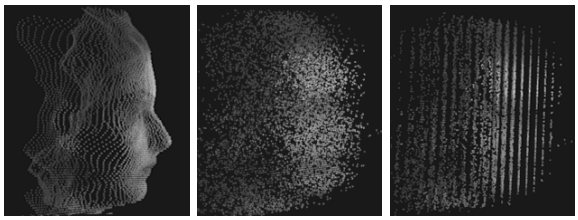
4.1 Generating data sets for quantitative evaluations

According to [16], 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 computed by

$$Z'(i) = Z_0(i) + n(i), \tag{17}$$

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 [16]. The quantization artifact is

[5] More results can be found at http://www.ntu.edu.sg/home/asjfcai/



(a) (b) (c)

Fig. 4. 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.

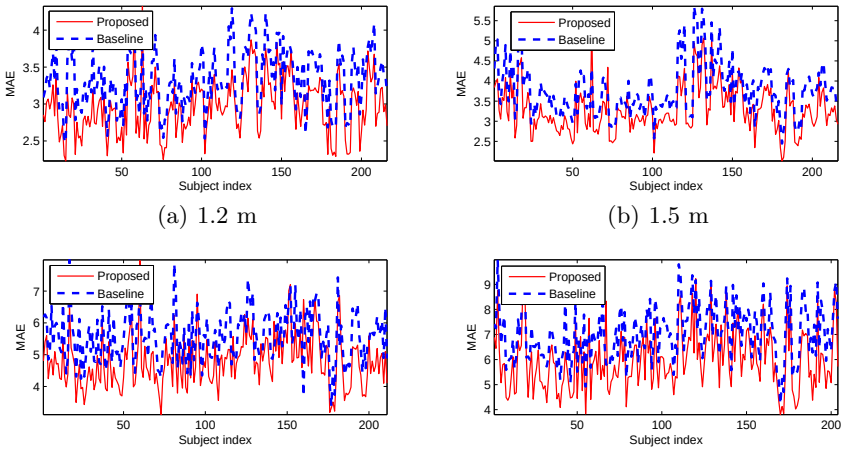
simulated by quantizing the noisy depth map using quantization steps computed from the camera parameters of Kinect. An example of adding Kinect-like artifacts is shown in Fig. 4.

4.2 Quantitative evaluations

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_I, l_z, and λ_d are set according to [9] for both the proposed and the baseline methods. For the proposed method, we empirically set $\epsilon_1 = 0.5$ and $\epsilon_2 = 3$. It should be noted that the proposed method is not sensitive to these parameters because its performance will be 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. According to our experience, $[0.1, 0.5]$ is a reasonable range of λ_f for the distance between 1.2 m and 2.0 m.

BU4D database contains more than 600 3D face expression sequences. For the evaluation of depth recovery, we re-render them as RGB-D data sets and only use the frame of neutral expression because the action units are not considered in our depth reconstruction task. Each depth map is rendered at four different camera distances: 1.2 m, 1.5 m, 1.75 m, and 2.0 m. There are 220 sets of RGB-D data with a neutral expression as the 1st frame, which are used in our experiments.

Fig. 5 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 method achieves higher recovery accuracy compared to the baseline method, which suggests that the face prior is helping the depth recovery. Fig. 6 gives a representative comparison between the baseline and the proposed methods. In Fig. 6 (a) and (b), we color the differences between the recovered depth maps and the ground truth, where dark blue represents small difference and other colors represent large differences. It is shown that the baseline method



(a) 1.2 m (b) 1.5 m

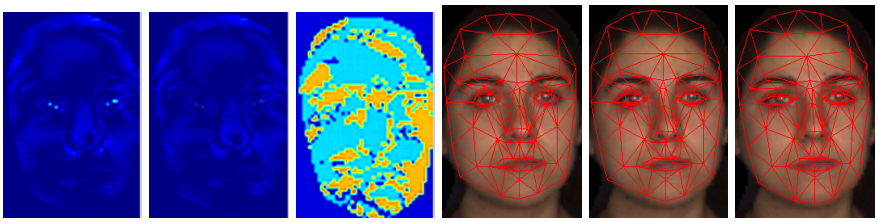
(c) 1.75 m (d) 2.0 m

Fig. 5. The depth MAE results on data sets rendered at different camera distances, (a) 1.2 m, (b) 1.5 m, (c) 1.75 m, and (d) 2.0 m.

cannot well handle the case of rich texture. Some large errors around the eyes' region are good examples for this case. In contrast, the face priors used in the proposed method can reduce such artifact and thus leads to higher recovery quality. In Fig. 6 (c), 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 case shown in Fig. 6 (c) is representative for most data sets. Therefore, the proposed method generally achieves higher recovery quality compared to the baseline method.

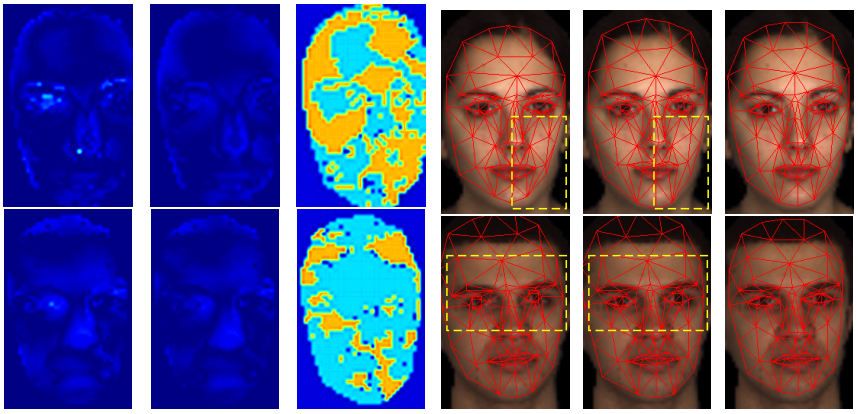
Besides the recovery error, we also evaluate the registration accuracy. To get the reference registration and shapes, we fit the 3D face model to noise-free data. The face model is also fitted to the depth maps obtained by different methods. We then compare the fitting result with the reference registration. A visual comparison between the registration results is shown in Fig. 6 (d)-(f). Fig. 7 gives more visual comparisons. 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 depth quality and registration accuracy are shown in Table 1. Besides the mean MAE for the recovery error, three metrics are used to compute the registration error. The 2D translation error is the 2D Euclid distance between the center of the fitted face model and the center of the reference model in the image plane. After aligning two 2D face models by aligning their centers, we compute the mean distance between the 2D landmarks of the fitted model and that 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., $err_{3D} = \|P_{\text{fit}} - P_{\text{ref}}\|/N_M$, where P_{fit} is the model that fits to the recovered depth



(a) Baseline (b) Ours (c) Comp. (d) Baseline (e) Ours (f) Ref.

Fig. 6. Representative results of depth recovery and face registration. (a) The difference map for the result of the baseline method, where dark blue represents small errors and light blue represents large errors. (b) The difference map for the result of the proposed method. (c) The comparison 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. (d) The registration result of fitting the model to the depth map recovered by the baseline method. (e) The registration result of the proposed method. (f) The registration result of fitting the model to the noise-free depth map, which is used as the reference.



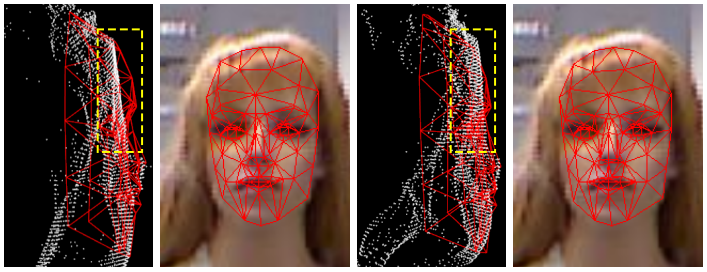
(a) Baseline (b) Ours (c) Comp. (d) Baseline (e) Ours (f) Reference

Fig. 7. Representative results of depth recovery and face registration on the datasets rendered at 1.75 m. For the depth recovery results, the baseline method produces some severe recovery error around the eyes and the nose, while the proposed method can effectively reduce the noise in these regions. For the face registration results, the proposed method produces a better fitting around the face boundary compared to the baseline method.

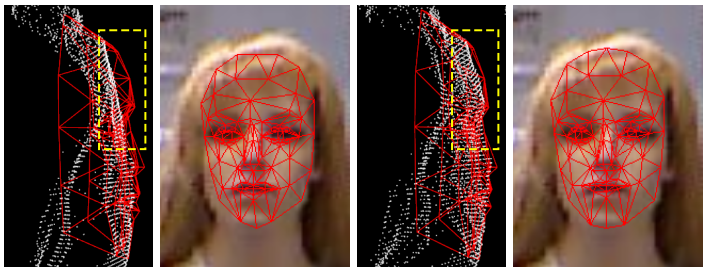
map and P_{ref} is the reference model that fits to the noise-free depth map. It can be seen that the proposed method achieves an improvement of recovery accuracy up to 15.1%, and the improvement of recovery accuracy exhibits a generally increasing trend with the distance. This is because the quality 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. The improvement of

Table 1. Quantitative evaluations of the proposed method using 4 metrics. The results obtain by the baseline and the proposed methods are separated by "/". The improvement of the proposed method over the baseline method is shown in percentages.

Dataset	Mean depth MAE	2D translation error	2D landmark error	3D shape error
1.20 m	3.33 / 2.97 (10.8%)	0.75 / 0.59 (21.3%)	1.17 / 1.02 (12.8%)	2.24 / 1.87 (16.5%)
1.50 m	3.76 / 3.29 (12.5%)	0.52 / 0.44 (15.4%)	0.99 / 0.85 (14.1%)	2.30 / 1.93 (16.1%)
1.75 m	5.58 / 4.79 (14.2%)	0.54 / 0.44 (18.5%)	1.07 / 0.89 (16.8%)	2.60 / 2.02 (22.3%)
2.00 m	7.04 / 5.98 (15.1%)	0.57 / 0.46 (19.3%)	1.09 / 0.93 (14.7%)	3.03 / 2.46 (18.8%)



(a) 1.50 m



(b) 1.68 m

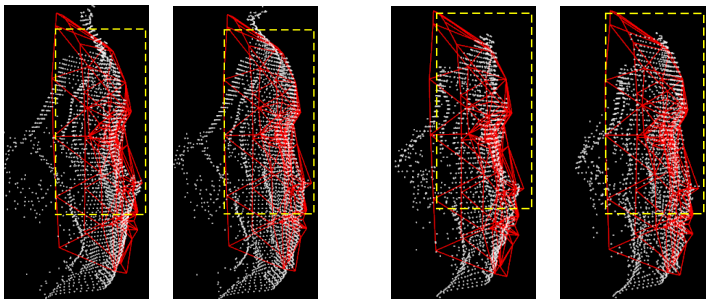
Fig. 8. The results on real Kinect data of a mannequin, which are shown in both 2D and 3D. In each figure, the result of the baseline method is on the left-hand side, while the result of the proposed method is on the right-hand side.

the proposed method in registration accuracy is also significant, up to 22.3%. It indicates that a better recovered depth map is helpful for the face alignment.

4.3 Experiments on real data

For the experiments on real data, we use Kinect 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 true face geometry.

Fig. 8 shows the registration (red wireframe) and depth recovery (white cloud) results of the mannequin at distances 1.50 m and 1.68 m. Fig. 9 shows some similar results for the male subject. It is shown that the proposed method



(a) Baseline / Ours (1.6 m) (b) Baseline / Ours (2.0 m)

Fig. 9. The results on real Kinect data of a male subject.

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.8, mainly at the forehead and noseline areas. On the other hand, the proposed method guided by the face prior is able to reconstruct more reasonable depth maps in those areas, which, e.g., follows the natural shape of a forehead. This also affects the final registration quality, although to a somewhat lesser extent than in the BU4D synthetic data. We attribute this to the discrepancy between the depth noise model used in BU4D experiments and that in the real data, as well as the additional noise in the color/texture channels.

5 Conclusion

The major contributions of this paper are twofold. First, we introduce the idea of using face priors in depth recovery, which has not been studied in literature before. Second, we formulate the problem as a joint optimization and develop an effective solution for it. Experimental results on a benchmark dataset show that, despite the coarse and sparse face prior model, properly taking into account the face priors brings in up to 15.1% of improvement in depth quality, which can be essential for applications such as 3D telepresence and teleconference. It can be expected that more accurate face priors will bring in more improvements. Moreover, the proposed method also leads to better registration accuracy, up to 22.3% of improvement, suggesting that the proposed method can also help other RGB-D face analysis tasks such as face tracking.

Acknowledgement. This research, which is carried out at BeingThere Centre, is mainly supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. This research is also partially supported by the 111 Project (No. B07048), China.

References

1. Mutto, C., Zanuttigh, P., Cortelazzo, G.: Microsoft Kinect^{TM} range camera. In: Time-of-Flight Cameras and Microsoft Kinect^{TM}. SpringerBriefs in Electrical and Computer Engineering. Springer US (2012) 33–47
2. Maimone, A., Fuchs, H.: Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In: Int'l Symposium Mixed Augmented Reality (ISMAR), Basel, Switzerland, IEEE (2011) 137–146
3. Kuster, C., Popa, T., Zach, C., Gotsman, C., Gross, M.: Freecam: A hybrid camera system for interactive free-viewpoint video. In: Proc. Vision, Modeling, and Vis. (VMV), Berlin, Germany (2011) 17–24
4. Zhang, C., Cai, Q., Chou, P., Zhang, Z., Martin-Brualla, R.: Viewport: A distributed, immersive teleconferencing system with infrared dot pattern. IEEE Multimedia **20** (2013) 17–27
5. Min, D., Lu, J., Do, M.: Depth video enhancement based on weighted mode filtering. IEEE Trans. Image Process. **21** (2012) 1176–1190
6. Richardt, C., Stoll, C., Dodgson, N.A., Seidel, H.P., Theobalt, C.: Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. Comp. Graph. Forum **31** (2012) 247–256
7. Yang, J., Ye, X., Li, K., Hou, C.: Depth recovery using an adaptive color-guided auto-regressive model. In: Europ. Conf. Comput. Vision (ECCV), Florence, Italy, Springer-Verlag (2012) 158–171
8. Zhao, M., Tan, F., Fu, C.W., Tang, C.K., Cai, J., Cham, T.J.: High-quality Kinect depth filtering for real-time 3d telepresence. In: IEEE International Conference on Multimedia and Expo (ICME). (2013) 1–6
9. Chen, C., Cai, J., Zheng, J., Cham, T.J., Shi, G.: A color-guided, region-adaptive and depth-selective unified framework for Kinect depth recovery. In: Int'l Workshop Multimedia Signal Process. (MMSP), Pula, Italy, IEEE (2013) 8–12
10. Qi, F., Han, J., Wang, P., Shi, G., Li, F.: Structure guided fusion for depth map inpainting. Pattern Recognition Letters **34** (2013) 70–76
11. Li, H., Yu, J., Ye, Y., Bregler, C.: Realtime facial animation with on-the-fly correctives. ACM Trans. Graph. **32** (2013) 42:1–42:10
12. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: FaceWarehouse: A 3D Facial Expression Database for Visual Computing. IEEE Transactions on Visualization and Computer Graphics **20** (2014) 413–425
13. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Int'l Conf. Comput. Vision (ICCV), Bombay, India, IEEE (1998) 839–846
14. Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., Toyama, K.: Digital photography with flash and no-flash image pairs. ACM Trans. Graph. **23** (2004) 664–672
15. Lai, P., Tian, D., Lopez, P.: Depth map processing with iterative joint multilateral filtering. In: Picture Coding Symposium (PCS), Nagoya, Japan, IEEE (2010) 9–12
16. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of Kinect depth data for indoor mapping applications. Sensors **12** (2012) 1437–1454
17. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models - their tranining and applications. Comput. Vis. Image Underst. (1995) 39–59
18. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. (2001) 681–684
19. Matthews, I., Baker, S.: Active appearance models revisited. Int. J. Comput. Vis. **60** (2004) 135–164

20. Baltruaitis, T., Robinson, P., Matthews, I., Morency, L.P.: 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking. In: CVPR. (2012) 2610–2617
21. Wang, H., Dopfer, A., Wang, C.: 3D AAM Based Face Alignment under Wide Angular Variations using 2D and 3D Data. In: ICRA. (2012)
22. Cai, Q., Gallup, D., Zhang, C., Zhang, Z.: 3d deformable face tracking with a commodity depth camera. In: Europ. Conf. Comput. Vision (ECCV). (2010)
23. Ahlberg, J.: Face and facial feature tracking using the active appearance algorithm. In: 2nd European Workshop on Advanced Video-Based Surveillance Systems (AVBS), London, UK (2001) 89–93
24. DeCarlo, D., Metaxas, D.: Optical flow constraints on deformable models with applications to face tracking. Int. J. Comput. Vis. **38** (2000) 99–127
25. Dornaika, F., Ahlberg, J.: Fast and reliable active appearance model search for 3d face tracking. IEEE Trans. Syst., Man, Cybern. **34** (2004) 1838–1853
26. Dornaika, F., Orozco, J.: Real-time 3d face and facial feature tracking. J. Real-time Image Proc. **2** (2007) 35–44
27. J. Orozco, O. Rudovic, J.G., Pantic, M.: Hierarchical on-line appearance-based tracking for 3D head pose, eyebrows, lips, eyelids and irises. Image and Vis. Comput. **31** (2013) 322–340
28. Ahlberg, J.: An updated parameterized face. Technical report, Image Coding Group, Dept. of Electrical Engineering, Linkoping University (2001)
29. Pham, H.X., Pavlovic, V.: Hybrid On-line 3D Face and Facial Actions Tracking in RGBD Video Sequences. In: Proc. International Conference on Pattern Recognition (ICPR). (2014)
30. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR. (2001) I–511 – I–518
31. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision **91** (2011) 200–215
32. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3d point sets. IEEE Trans. Pattern Anal. Mach. Intell. **9** (1987) 698–700
33. Low, K.: Linear least-squares optimization for point-to-plane ICP surface registration. Technical Report TR04-004, Department of Computer Science, University of North Carolina at Chapel Hill (2004)
34. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: 8th IEEE International Conference on Automatic Face Gesture Recognition, IEEE (2008) 1–6