2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Facial Depth Map Enhancement via Neighbor Embedding

Shuai Yang[1], Sijie Song[1], Qikun Guo[2], Xiaoqing Lu[1] and Jiaying Liu[1]*
[1]Institute of Computer Science and Technology, Peking University, Beijing, P.R.China, 100871
[2]Department of Computer Science, Brown University, Providence Rhode Island, 02912

***Abstract*—The simple yet subtle structures of faces make it difficult to capture the fine differences between different facial regions in the depth map, especially for consumer devices like Kinect. To address this issue, we present a novel method to super-solve and recover the facial depth map nicely. The key idea of our approach is to exploit the learning-based method to obtain the reliable face priors from high quality facial depth map to further improve the depth image. Specifically, we utilize the neighbor embedding framework. First, face components are decomposed to train specialized dictionaries and reconstructed, respectively. Joint features, i.e. color, depth and position cues, are put forward for robust patch similarity measurement. The neighbor embedding results form high frequency cues of facial depth details and gradients. Finally, an optimization function is defined to combine these high frequency information to yield depth maps that fit the actual face structures better. Experimental results demonstrate the superiority of our method compared to state-of-the-art techniques in recovering both synthetic data and real world data from Kinect.**

I. Introduction

In recent years, with the development of consumer-level depth cameras such as Time-of-Flight (ToF) and Microsoft Kinect, the easy and real-time acquisition of depth images becomes available. Since depth map is insensitive to the environment and can provide spatial information, it has been widely used in 3D reconstruction, semantic scene analysis and object recognition, especially for human faces. However, the applications of depth information is significantly constrained by the limited resolution and noises of sampled depth maps. Many researchers try to solve this issue through super resolution (SR) approaches.

The methods for depth map super resolution can be divided into two categories: multiple depth map fusion [1], [2], [3] and single depth image super resolution [4], [5], [6]. Multiple depth map fusion techniques merge several unaligned low-quality depth maps to reconstruct a high-quality depth map. However, usually only one depth map is available in practice. Single depth image super resolution refers to recovering the information with a single low-quality input. For many depth cameras like Kinect, a corresponding high-quality color image is available and can be used as the guidance to improve the depth map recovery, which refers to RGB-D super resolution [7], [8], [9]. According to the super resolution strategies, RGB-D super resolution can be categorized as filter-based, optimization-based and learning-based methods.

Filter-based methods are widely adopted by early works. They usually filtered depth maps adaptively according to the structural information. In [10], the bilateral filter is used to consider both depth structures and color intensities. Inspired by successful stereo matching algorithms, Yang *et al.* [11] iteratively employed a bilateral filter to improve depth map super resolution. He *et al.* [7] proposed guided filtering as an edge-preserving smoothing operator like the bilateral filter. Liu *et al.* [12] proposed to use geodesic distance to calculate the filter weight and recover shaper edges. The performance of these methods relies on the high correlation between depth and color information. Their filter weights can be misled for facial depth maps because the color of human faces lacks changes.

More recently, optimization-based methods have been developed for RGB-D super resolution. MRF is widely used to model local priors in image enhancement [**?**], [13], [14]. Depth map refinement based on MRF optimization was first explored in [13]. Park *et al.* [14] add a non-local means term to their MRF formulation to preserve structures and remove outliers. Yang *et al.* [15] used the Auto-Regressive (AR) model to formulate the depth refinement problem. In [8], The upsampling is formulated as a global energy optimization problem using Total Generalized Variation (TGV) regularization. These methods can produce high-quality depth maps if the energy term which reflects image priors is well designed. However for facial depth maps, defining an universal face prior artificially is a tough task. This problem can be solved by adopting learning algorithms to automatically learn good face priors.

Learning-based methods attempt to model statistical dependencies between color and depth signals in RGB-D features through proper dictionaries. With edges extracted from a color image, Li *et al.* [16] trained a joint dictionary consisting of both the gradient of the depth map and the edge information of the color image. However, this method does not consider the discontinuities between color edges and depth edges. To tackle this problem, Tosic and Drewes [17] proposed a method based on a novel second order cone program for recovering signals from their common underlying 3D features. But this approach may yield depth distortions. Then Kwon *et al.* [18] refined the depth map by the normalized Absolute Gradient Dot Product (nAGDP), which resulted in good performances. Although the existing learning-based methods are capable to introduce extra high-quality depth information to yield better performance,

*Corresponding author.
This work was supported by National High-tech Technology R&D Program (863 Program) of China under Grant 2014AA015205 and National Natural Science Foundation of China under contract No. 61472011.

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

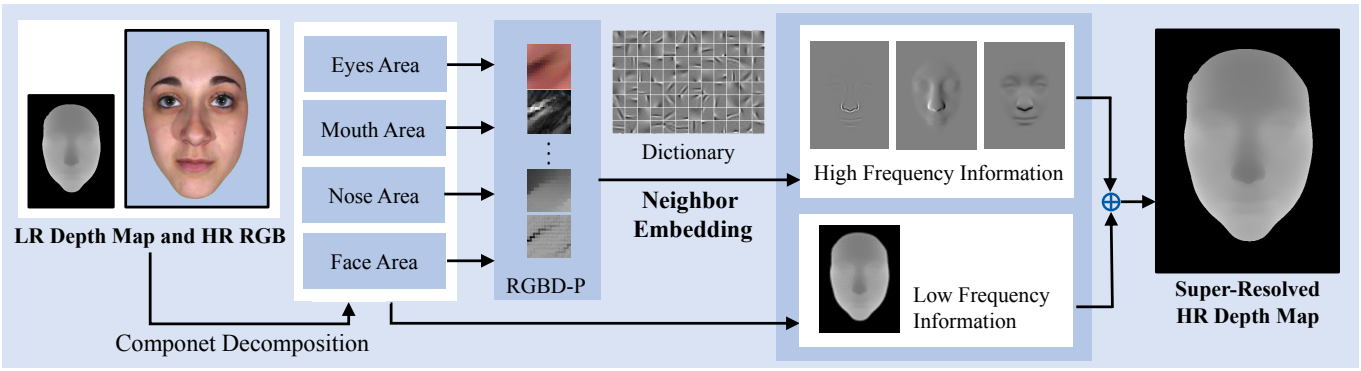


Fig. 1. Main framework of the proposed method.

seldom do they well study and utilize high-level human face prior knowledge. In our work, we focus on both high-level facial cues and low-level depth and intensity cues. These cues are carefully designed to jointly depict image patches of facial depth maps to learn reliable facial priors.

In this paper, we propose a learning-based approach for dealing with the problem of low-quality facial depth maps through neighbor embedding. Utilizing an external facial depth map dataset, our method can achieve high quality results from only a single noisy low resolution depth map and its corresponding color image. The target facial depth map first is decomposed into server facial component regions for reconstruction. Then, an RGBD-Position(RGBD-P) feature is calculated to combine high-level and low-level information to precisely measure patch similarity. Finally a global optimization function is used to perform specialized gradient constraints over the reconstructed depth map to further impose facial relative elevation priors.

The performance of our algorithm is evaluated with state-of-art depth map enhancement methods. Our approach demonstrates superior performance in both synthetic depth maps and Kinect depth data. In summary, the main contribution of our work are as follows:

- **Joint scale-independent RGBD-P feature**. We design robust RGBD-P features that take both high-level facial component position information and low-level intensity and depth information into account. This feature effectively solves the scale problem and the ambiguity problem for super resolution.
- **Face prior analysis and utilization in neighbor embedding model**. In our neighbor embedding depth map enhancement framework, we take full use of face prior knowledge from the external dataset. Thus the precise depth variation of face structures can be well recovered even the depth data is severely degraded.

The rest of this paper is organized as follows: Section II describes the proposed facial depth map super resolution approach. Experimental results are shown in Section III and concluding remarks are given in Section IV.

II. Proposed Method

In this section, the proposed facial depth map enhancement method is presented. Given a target Low Resolution (LR) facial depth map $\mathbf{X}_l$ and its corresponding High Resolution (HR) color image as input, we estimate the target HR depth map $\mathbf{X}_h$ with the help of the coupled LR and HR dictionaries $\mathcal{Y} = \{\mathcal{Y}_l, \mathcal{Y}_h\} = \{\mathbf{y}_l^i, \mathbf{y}_h^i\}_{i=1}^N$, where $\mathbf{y}_l^i/\mathbf{y}_h^i$ are paired LR/HR patches from external source depth maps and N is the dictionary size. Figure 1 shows the framework of the proposed method. To fully exploit the structural prior of human faces, we first decompose a whole face into facial components, based on the high-quality color image. Then, for each component, a neighbor embedding is performed in RGBD-P feature space, which takes intensity, depth and position information into accounts. Specifically, for each patch $\mathbf{x}_l$ in $\mathbf{X}_l$, we extract its RGBD-P feature and find its K nearest neighbors $\mathcal{N}_l \in \mathcal{Y}_l$. The corresponding HR neighbors $\mathcal{N}_h \in \mathcal{Y}_h$ are used to reconstruct the HR embedding $\mathbf{x}_h$, which provides the high frequency information of human faces. Finally, the low frequency part from the raw data and the learned high frequency part are fused to generate the final super-resolved and recovered results.

A. Joint Scale-Independent RGBD-P Feature

For RGB-D super resolution, the noise problem, scale problem and ambiguity problem between LR/HR pairs and RGB/D pairs are three main issues. To tackle these issues, the RGBD-P feature is proposed for similarity measurement and depth map reconstruction. We combine high-level cues of human faces with low-level RGBD cues, and all theses cues are carefully designed to be scale-independent.

We start with a face detection and a landmark localization [19]. Each face is annotated by landmark points that locate facial components of interest. As shown in Figure 2, we concentrate on the eyes, nose and mouth regions. The patch features of four component regions are then extracted to form specialized dictionaries $\mathcal{Y}^i, i \in \{1, 2, 3, 4\}$. And each region of the testing image is reconstructed using the corresponding dictionaries. After facial component decomposition, each patch is implicitly classified and with the help of this high-level

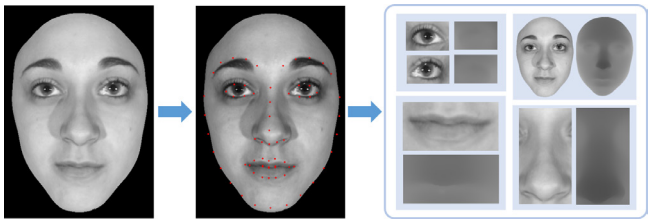


Fig. 2. Facial component decomposition. From left to right: a face image, detected landmarks and facial component regions

classification cue, the obtained neighbors are more reliable, as shown in Figure 3(b) and (d). In addition, we enrich source patches by including the mirror symmetry of the facial component regions based on the symmetry of human faces. For simplicity, in the following sections, we use $\mathcal{Y}$ to refer to the four training sets $\mathcal{Y}^i$.

Then, in the training phase, we extract RGBD-P features of source patches as $\{\mathbf{y}_l, \mathbf{y}_c, \mathbf{y}_p, \mathbf{y}_h\}$, where $\mathbf{y}_l$ and $\mathbf{y}_h$ are the low and high frequency depth features, respectively. $\mathbf{y}_c$ describes the intensity feature of the color image. Furthermore, $\mathbf{y}_p$ depicts the position feature.

Low Frequency Depth Features: Let $\bar{\mathbf{y}}$ denote the low frequency component of $\mathbf{y}$ and it is defined by $\bar{\mathbf{y}} = D^T(D\mathbf{y})$ where D is bicubic downsampling and D^T is bicubic upsampling. Then the low frequency depth feature is given by: $\mathbf{y}_l = [\nabla\bar{\mathbf{y}}; \nabla^2\bar{\mathbf{y}}; w_d\bar{\mathbf{y}}_{norm}]$, where ∇ and ∇^2 are the first and second derivatives, respectively. $\bar{\mathbf{y}}_{norm}$ is the normalized depth with zero means to cope with the scale problem.

Intensity Features: $\mathbf{y}_c = [I_{norm}; I_{edge}]$ contains intensity and edge information. I_{edge} is obtained by calculating the maximum gradient magnitude among the RGB channels and normalizing to have 1 as the maximum element.

Position Features: $\mathbf{y}_p = [x/W; y/H]$, where (x, y) are the coordinates of the patch center and $[W, H]$ are the width and height of the facial component region. Each component is aligned implicitly after facial component decomposition and the localization of a patch in its corresponding region can be an important high-level cue for neighborhood searching. To determine its reliably, we perform an experiment on the impact of the position features. As shown in Figure 3(b) and (d), the ambiguity problem is effective resolved.

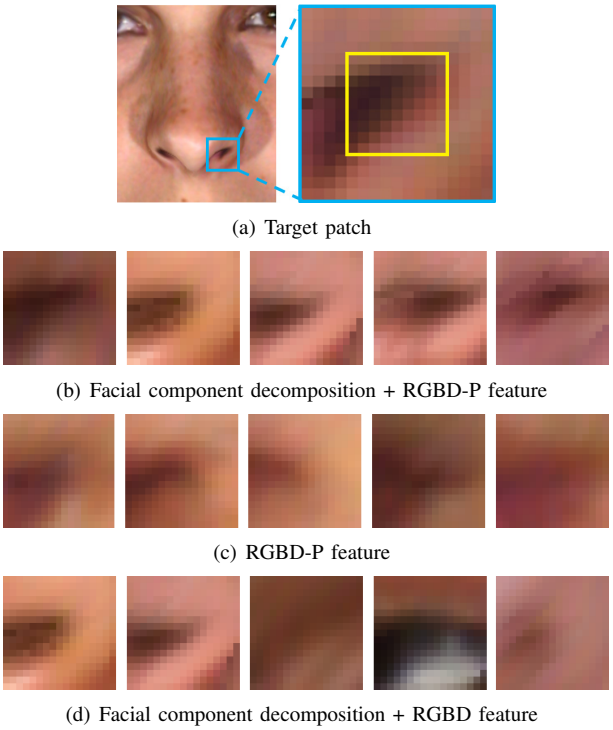
High Frequency Depth Features: $\mathbf{y}_h = [\delta\mathbf{y}; \nabla\mathbf{y}]$, where $\delta\mathbf{y} = \mathbf{y} - \bar{\mathbf{y}}$. To overcome the noise problem in the degraded LR depth map and restore precise facial structures, we put forward the gradients $\nabla\mathbf{y}$ of noise-free high-quality depth map, which forms the relative elevation of the face priors.

Meanwhile, in the testing phase, target image patch features are extracted in similar ways: $\{\mathbf{x}_l, \mathbf{x}_c, \mathbf{x}_p\}$, where the low frequency component of $\mathbf{x}$ is calculated by $\bar{\mathbf{x}} = D^T\mathbf{x}$.

Given the joint scale-independent RGBD-P features, we formulate a measure for two patches:

$$\text{dist}(\mathbf{x}_l, \mathbf{y}_l) = ||F(\mathbf{x}_l) - F(\mathbf{y}_l)||_2^2, \tag{1}$$

where $F(\mathbf{x}_l) = [\mathbf{x}_l; w_c\mathbf{x}_c; w_p\mathbf{x}_p]$ and w_c, w_p are the weights to combine the depth, intensity and position information.



(a) Target patch

(b) Facial component decomposition + RGBD-P feature

(c) RGBD-P feature

(d) Facial component decomposition + RGBD feature

Fig. 3. Influence factors for target patch neighborhood searching. (a) The target patch is shown in the yellow rectangle. To better recognize the content of an patch, expanded boundaries are added to all the patches as shown in the blue rectangle. For space saving, we do not show the corresponding depth maps. (b) The five most similar patches found based on RGBD-P feature in the nose component dataset. (c) The nostril is matched to the corners of the mouth if using the general face dataset. (d) Removing position features (by setting $w_p = 0$) leads to failure search. See the fourth patch (the corner of the eye).

B. Depth Map Reconstruction via Neighborhood Regression

By jointly considering the proposed RGBD-P features, we are capable to find reliable similar patches to form K neighborhoods $\mathcal{N}_l^i = [\mathbf{y}_l^{i_1}, \mathbf{y}_l^{i_2}, ..., \mathbf{y}_l^{i_K}]$ for each patch $\mathbf{x}_l^i$. Following the standard neighbor embedding procedure, the regression weight $\alpha_i \in \mathbb{R}^K$ is calculated by:

$$\min_{\alpha_i} ||\mathbf{x}_l^i - \mathcal{N}_l^i\alpha_i||_2^2 + \mu||\alpha_i||_2^2, \tag{2}$$

where μ is the sparse coefficient. Next, the corresponding HR neighbors $\mathcal{N}_h^i = [\mathbf{y}_h^{i_1}, \mathbf{y}_h^{i_2}, ..., \mathbf{y}_h^{i_K}]$ are used to reconstruct the high frequency information:

$$\mathbf{x}_h^i = \mathcal{N}_h^i\alpha_i. \tag{3}$$

After that, we average the feature values in overlapped regions between adjacent patches and patch features $\mathbf{x}_h$ are merged into the image space, resulting in the high frequency details and facial structures $\{\delta\mathbf{X}, \nabla\mathbf{X}\}$.

Finally, the HR depth map $\mathbf{X}_h$ is estimated by combining low-frequency information $\bar{\mathbf{X}} = D^T\mathbf{X}_l$, high-frequency details $\delta\mathbf{X}$ and facial structure priors $\nabla\mathbf{X}$:

$$\begin{aligned}\mathbf{X}_h = \arg\min_{\mathbf{u}} &||\nabla_x\mathbf{u} - \nabla_x\mathbf{X}||_2^2 + ||\nabla_y\mathbf{u} - \nabla_y\mathbf{X}||_2^2 \\ &+ \lambda||\mathbf{u} - \bar{\mathbf{X}} - \delta\mathbf{X}||_2^2,\end{aligned} \tag{4}$$

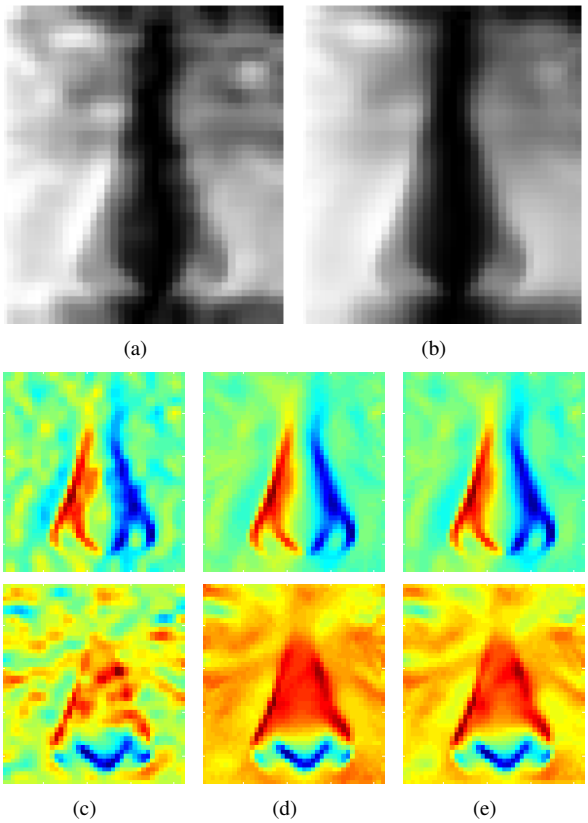


Fig. 4. The impact of gradient terms. (a) The reconstruction result without gradient terms ($\lambda = 0$). (b) The reconstruction result with $\lambda = 1$. (c) and (e) are the gradients of (a) and (b) respectively in the x and y directions. (d) are $\nabla_x \mathbf{X}$ and $\nabla_y \mathbf{X}$, which provide reliable facial structure information.

where $\nabla_x \mathbf{X}$ and $\nabla_y \mathbf{X}$ are the x component and y component of $\nabla \mathbf{X}$ and λ balances three terms. As shown in Figure 4, $\nabla \mathbf{X}$ imposes high-quality facial structure priors on the noisy depth map and creates smoother results.

C. Multi-Scale Solution

Since the ambiguity gets severer when the scale different gets greater, we take the multi-scale strategy to address this issue. Specifically, for the scale l, the training set is obtained using the downsampled source color images and depth maps with factor $1/2^{l-1}$, and the reconstruction result at scale l forms the LR target depth map at scale $l-1$.

III. Experimental Results And Analysis

The proposed method is implemented on Matlab R2014a platform. We first compare our method with state-of-the-art depth map enhancement methods on the synthetic depth maps derived from 3D face models in the BU-3DFE dateset [20]. Both noise-free and noisy cases are considered. Beyond these simulations, we evaluate our method on the real world data collected by Kinect cameras. More experimental results can be found in the supplemental material. In the experiments, we set patch size $n = 9 \times 9$. The dictionary size N is $100,000$, and for each patch $K = 9$ nearest neighbors are searched. For joint scale-independent RGBD-P feature extractions, the weights between different terms are set to $w_d = 0.1$, $w_c = 3$ and $w_c = 81$ for all experiments, which demonstrates the robustness of the proposed feature. The sparse coefficient μ in the neighborhood regression is 0.15. Meanwhile, the factor λ to control the depth map smoothness is set to 4 in the noise-free case. For noisy depth map, it is empirically chosen 0.25, 1 and 1 for $2\times$, $4\times$ and $8\times$ upsampling, respectively.

A. Super Resolution of Synthetic Facial Depth Data

The BU-3DFE dateset is used to construct the dictionaries and testing data. Figure 5 shows an example of the synthetic RGB-D data. We take 70 3D face models and synthesize depth maps using the z-axis data. The corresponding color image is obtained from the textured models. The rest of the face models in the dataset form our testing data in the following experiments.

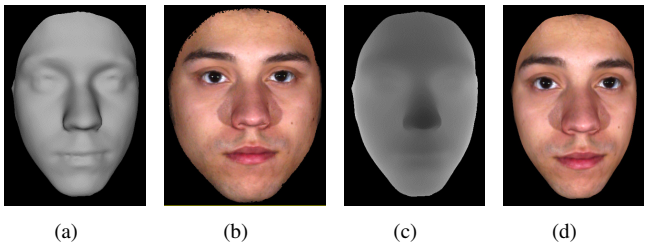


Fig. 5. RGB-D data construction for experiments. (a) and (b) 3D face mesh and the corresponding texture from BU-3DFE dataset. (c) and (d) synthetic facial depth map and color image.

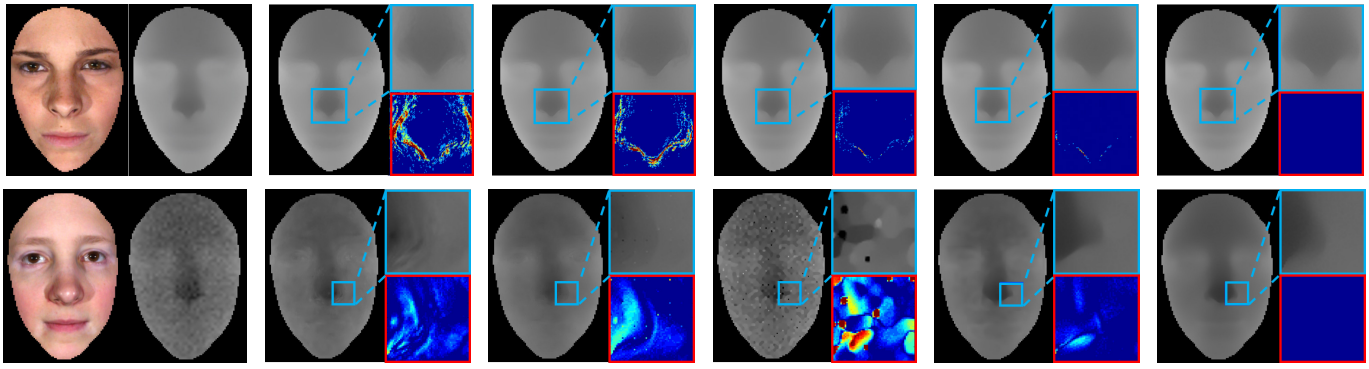
1) Noise-Free Testing Data: We apply the proposed method to noise-free facial depth maps. We compare with methods from [7][1], [8][2] and [9][3], which can be representative for filter-based, optimization-based and learning-based techniques, respectively. The training data used for [9] is identical to that used in our method. Table I reports the comparison of $2\times$, $4\times$ and $8\times$ upsampling in terms of Root-Mean-Square Error (RMSE) and the proposed method obtains lowest RMSE. Kiechle's method [9] also achieves low RMSE, indicating that the depth reconstruction can be well improved through learning from high-quality depth maps. The first row of Figure 6 gives the detailed qualitative comparison of $4\times$ upsampling for the Woman1 example. Judging from the difference map, the result of the proposed method is the closest to the ground truth.

2) Noisy Testing Data: The experiments for nosiy facial depth map super resolution is conducted. The quantitative and qualitative comparisons are given in Table II and the second row of Figure 6. Without sufficient reliable depth information, the filter-based method [7] produces distinct texture copying artifacts. Ferstl's method [8] successfully suppresses the noise while importing noticeable false reconstruction result. For instance, the sudden change around the nose and eyeball,

[1]Code from http://research.microsoft.com/en-us/um/people/kahe/eccv10/

[2]Code from http://rvlab.icg.tugraz.at/project_page/project_tofusion/project_tofsuperresolution.html

[3]Code from http://www.gol.ei.tum.de/index.php?id=6&L=1



(a) Input RGBD pairs (b) He [7] (c) Ferstl [8] (d) Kiechle [9] (e) Ours (f) Groundtruth

Fig. 6. Comparisons on the BU-3DFE dataset. The first row shows an example of $4\times$ upsampling on noise-free data and the second row shows an example of $8\times$ upsampling on noisy data.

TABLE I
UPSAMPLING OF NOISE-FREE BU-3DFE DATA

	Woman1			Woman2			Man1			Man2		
	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$
He *et al.*[7]	0.659	0.904	1.637	0.607	0.807	1.363	0.639	0.926	1.632	0.640	0.795	1.284
Ferstl *et al.*[8]	0.677	0.796	1.052	0.630	0.730	1.059	0.689	0.841	1.294	0.656	0.720	0.850
Kiechle *et al.*[9]	0.604	0.624	0.801	0.575	0.596	0.694	0.574	0.596	0.863	0.617	0.639	0.746
Ours	**0.535**	**0.598**	**0.633**	**0.505**	**0.567**	**0.603**	**0.506**	**0.571**	**0.636**	**0.541**	**0.609**	**0.622**

TABLE II
UPSAMPLING OF NOISY BU-3DFE DATA

	Woman1			Woman2			Man1			Man2		
	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$	$\times 2$	$\times 4$	$\times 8$
He *et al.*[7]	1.897	2.175	2.624	1.834	2.045	2.603	1.910	2.133	2.705	1.856	2.079	2.476
Ferstl *et al.*[8]	1.293	2.352	2.516	1.199	1.840	2.563	1.310	2.032	2.618	1.245	1.965	2.330
Kiechle *et al.*[9]	5.446	6.836	8.340	5.314	6.875	8.211	5.558	6.777	8.144	5.495	6.901	8.101
Ours	**1.185**	**1.800**	**2.325**	**1.031**	**1.717**	**2.280**	**1.191**	**1.790**	**2.191**	**1.156**	**1.749**	**2.149**

which is clearly misled by the color image. The learning-based method in [9] is severely affected by the noise and produces some impulsive noise. By comparison, thanks to the proposed RGBD-P features, our method is capable of using the low-level and high-level cues to obtain reliable high-quality depth gradients and produces ideal results.

B. Super Resolution of Real World Kinect Facial Depth Data

In the end, we apply the proposed method to real world Kinect facial depth data. In the experiment, the facial depth maps are upsampled by a factor of 8. Figure 7 illustrates that the proposed method preserves most of the facial components, and the boundaries in the side view are quite smooth. The sudden change in depth map reconstructed by the method in [9] leads to the stepping artifacts on 3D surfaces. Although Ferstl's method [8] generates smoother results, it cannot recover noses and lips that matches the normal facial physical structures. Please enlarge and view these figures on the screen for better comparison.

IV. CONCLUSION

We present a novel facial depth map enhancement method via neighbor embedding. We decompose the facial depth map into four facial regions to seek the high-level position cues. We combine low-level RGB-D cues and these high-level cues to form a joint RGBD-P feature for better similarity measurement. The proposed neighbor embedding framework can learn high-quality facial details and structures to significantly improve facial depth reconstruction. We validate the superiority of our method by comparisons with state-of-the-art technologies.

REFERENCES

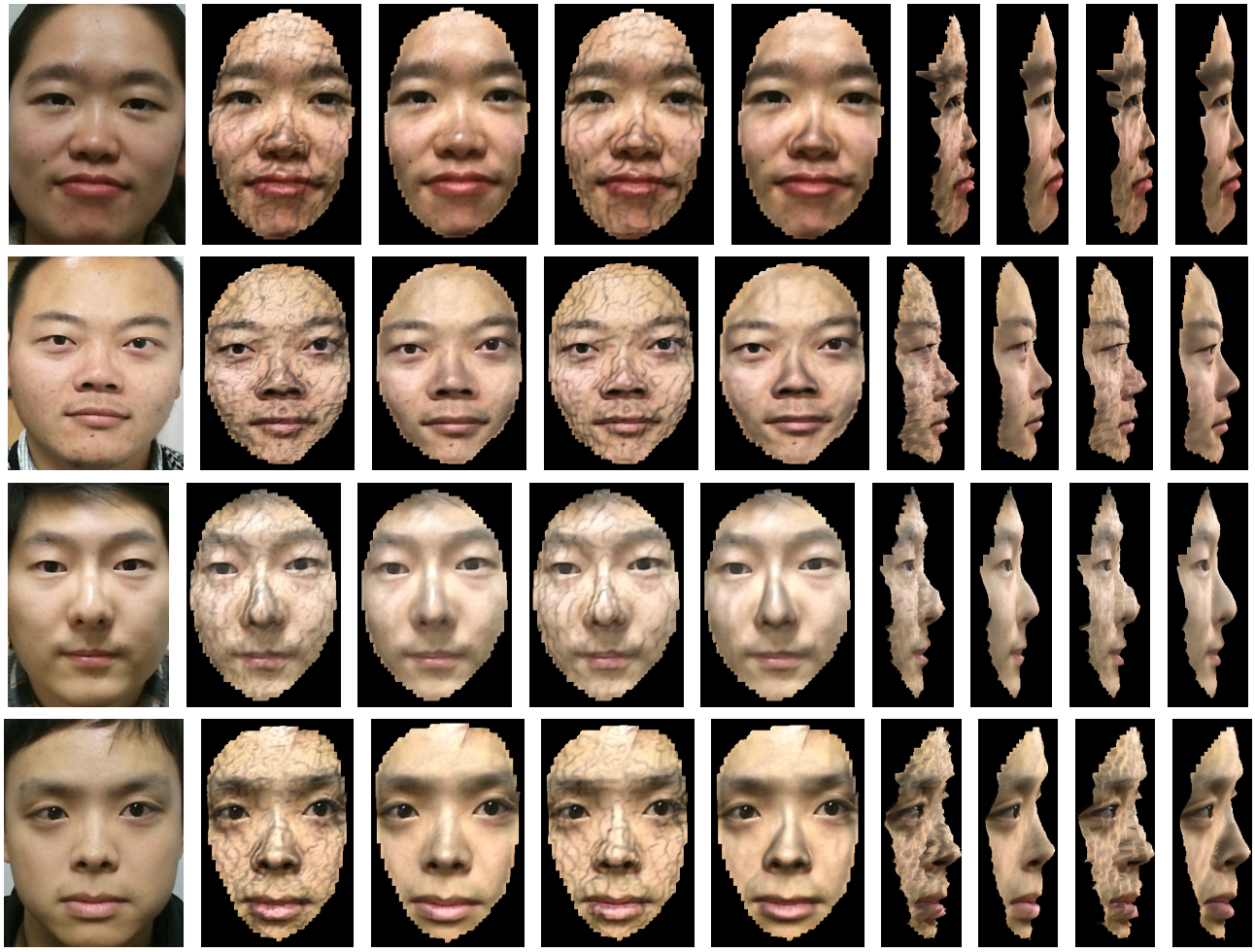
[1] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "Lidarboost: Depth superresolution for tof 3d shape scanning," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2009, pp. 343–350.

[2] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon, "Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 2011, pp. 559–568.

[3] Sigurjon Arni Gudmundsson, Henrik Aanaes, and Rasmus Larsen, "Fusion of stereo vision and time-of-flight imaging for improved 3d estimation," *Int. J. Intell. Syst. Technol. Appl.*, vol. 5, no. 3/4, pp. 425–433, November 2008.

[4] David Ferstl, Matthias Ruther, and Horst Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2015, pp. 513–521.

[5] Jun Xie, R.S. Feris, Shiaw-Shian Yu, and Ming-Ting Sun, "Joint super resolution and denoising from a single depth image," *IEEE Transactions on Image Processing*, vol. 17, no. 9, pp. 1525–1537, Sep 2015.



(a) RGB (b) Raw (c) Ferstl [8] (d) Kiechle [9] (e) Ours

Fig. 7. Comparison on Kinect face data.

[6] J. Xie, R Feris, and M. T. Sun, "Edge-guided single depth image super resolution," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 428 – 438, Nov 2015.

[7] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[8] D. Ferstl, C. Reinbacher, R. Ranftl, M. Ruether, and H. Bischof, "Image guided depth upsampling using anisotropic total generalized variation," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, Dec 2013, pp. 993–1000.

[9] M. Kiechle, S. Hawe, and M. Kleinsteuber, "A joint intensity and depth co-sparse analysis model for depth map super-resolution," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, Dec 2013, pp. 1545–1552.

[10] Johannes Kopf, Michael F. Cohen, Dani Lischinski, and Matt Uyttendaele, "Joint bilateral upsampling," *Acm Transactions on Graphics*, vol. 26, no. 3, pp. 96, 2007.

[11] Qingxiong Yang, Ruigang Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2007, pp. 1–8.

[12] Ming-Yu Liu, O. Tuzel, and Y. Taguchi, "Joint geodesic upsampling of depth images," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 169–176.

[13] James Diebel and Sebastian Thrun, "An application of markov random fields to range sensing," *Advances in Neural Information Processing Systems*, pp. 291–298, 2005.

[14] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, M.S. Brown, and Inso Kweon, "High quality depth map upsampling for 3d-tof cameras," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, Nov 2011, pp. 1623–1630.

[15] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang, "Color-guided depth recovery from rgb-d data using an adaptive autoregressive model," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3443–3458, Aug 2014.

[16] Yanjie Li, Tianfan Xue, Lifeng Sun, and Jianzhuang Liu, "Joint example-based depth map super-resolution," in *Proc. IEEE Int'l Conf. Multimedia and Expo*, July 2012, pp. 152–157.

[17] I. Tosic and S. Drewes, "Learning joint intensity-depth sparse representations," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2122–2132, May 2014.

[18] HyeokHyen Kwon, Yu-Wing Tai, and S. Lin, "Data-driven depth map refinement via multi-scale sparse representation," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 159–167.

[19] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 2879–2886.

[20] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *International Conference on Automatic Face and Gesture Recognition*, April 2006, pp. 211–216.