Surface UP-SR for an Improved Face Recognition Using Low Resolution Depth Cameras

Djamila Aouada Kassem Al Ismaeil Kedija Kedir Idris Björn Ottersten
⋆Interdisciplinary Centre for Security, Reliability and Trust
Universtity of Luxembourg
{djamila.aouada, kassem.alismaeil, bjorn.ottersten}@uni.lu, kedija.kedir@gmail.com

Abstract

We address the limitation of low resolution depth cameras in the context of face recognition. Considering a face as a surface in 3-D, we reformulate the recently proposed Upsampling for Precise Super–Resolution algorithm as a new approach on three dimensional points. This reformulation allows an efficient implementation, and leads to a largely enhanced 3-D face reconstruction. Moreover, combined with a dedicated face detection and representation pipeline, the proposed method provides an improved face recognition system using low resolution depth cameras. We show experimentally that this system increases the face recognition rate as compared to directly using the low resolution raw data. [1]

1. Introduction

In the past ten to fifteen years, research on automatic face recognition has actively moved from 2-D to 3-D data mostly acquired using high resolution (HR) laser scanners. Multiple approaches have been developed for this kind of data. Until recently the race was about designing sensors to capture data with higher levels of details and higher resolutions [1]. Today much more affordable and less bulky depth cameras, with 3-D capabilities, have become accessible. They are, however, of limited resolutions, and present a high level of noise. Some examples are the 3D MLI by IEE S.A. of resolution (56×64) [2], and the PMD camboard nano of resolution (120×165) [3]. Because of their low resolution (LR) and the noisy nature of the acquired data, previously defined 3-D face recognition algorithms are no longer ensured to be as effective [9].

The multi-frame super-resolution (SR) framework is an appropriate solution where it becomes possible to recover a higher resolution frame by fusing multiple LR ones. It has been successfully used in the case of 2-D face images [5, 6]. Similar efforts have been undertaken for 3-D facial data. In [11], a learning–based method has been proposed to directly find the mapping between an LR image and its corresponding HR image without using multiple frames. In [12], Peng et al. proposed to use facial features in a Maximum A Posteriori SR framework.

Depth facial data may also benefit from the SR framework. Recently, Berretti et al. proposed to use SR on facial depth images once back-projected in 3-D, and defined the *superfaces* approach [9]. The SR algorithm they deployed is similar in principle to the initial blurred estimate provided in the *enhanced Shift & Add* algorithm proposed by Al Ismaeil et al. in [7]. Later on, this work was extended to the dynamic case where the considered multiple realizations were ordered frames constituting a video sequence [8]. This approach is referred to as *Upsampling for Precise Super-Resolution* (*UP-SR*). Its key component is a prior upsampling of the observed data which is proven to enhance the registration of frames over time. In addition, it uses a bilateral total variation framework as a smoothness condition. In [16], a similar concept of temporal fusion was considered for 3-D facial data enhancement. However, the increase in resolution was induced from temporal data cumulation without a real SR formulation or upsampling. Moreover, smoothness was ensured by bilateral filtering as a post processing operation and not included in the optimization objective function.

The contribution of this paper is twofold; first, we reformulate *UP-SR* on 3-D point clouds constituting the facial surface similarly to the work in [9]. However, by performing the deblurring phase of *UP-SR*, 3-D face reconstruction results are maintained, if not enhanced. Second, we show experimentally that using these results for 3-D face recognition clearly improves the recognition rate as compared to using raw LR acquisitions. This second contribution requires a full dedicated pipeline for automatic face acquisition from depth cameras. Moreover, level curves equidistant from the nose tip and radially sampled are considered

[1]This work was supported by the National Research Fund, Luxembourg, under the CORE project C11/BM/1204105/FAVE/Ottersten.

as facial features for matching and comparison.

The remainder of the paper is organized as follows: Section 2 reviews the *UP-SR* algorithm. Its adaptation to facial depth data on a surface is given in Section 3. Our proposed face recognition pipeline is detailed in Section 4 which includes a description of the considered level curves. The experimental setup and results are summarized in Section 5. Finally, we conclude with Section 6.

2. Background

In what follows, we review the *UP-SR* algorithm. We represent all images in lexicographic vector form. Let us consider an HR depth image $\mathbf{x}$ of size n, and N observed LR images $\mathbf{y}_k$, $k = 0, ..., (N-1)$, of size m, such that $n = r \cdot m$, where r is the SR factor. Every frame $\mathbf{y}_k$ is an LR noisy and deformed realization of $\mathbf{x}$ modeled as follows:

$$\mathbf{y}_k = \mathbf{DHW}_k\mathbf{x} + \mathbf{n}_k, \qquad k = 0, ..., (N-1), \qquad (1)$$

where $\mathbf{W}_k$ is an $(n \times n)$ invertible matrix corresponding to the geometric motion between $\mathbf{x}$ and $\mathbf{y}_k$. We assume that $\mathbf{y}_0$ is the reference frame for which $\mathbf{W}_0 = \mathbf{I}_n$. The point spread function of the depth camera is modeled by the $(n \times n)$ space and time invariant blurring matrix $\mathbf{H}$. The matrix $\mathbf{D}$ of dimension $(m \times n)$ represents the downsampling operator, and the vector $\mathbf{n}_k$ is an additive noise at k which follows a white multivariate Laplace distribution of mean zero and covariance $\mathbf{\Sigma} = \sigma^2\mathbf{I}_m$, with $\mathbf{I}_m$ being the identity matrix of size $(m \times m)$.

One of the key components of *UP-SR* is to upsample the observed LR images prior to any operation. We define the resulting r-times upsampled image as:

$$\mathbf{y}_k \uparrow = \mathbf{U} \cdot \mathbf{y}_k, \qquad (2)$$

where $\mathbf{U}$ is an $(n \times m)$ upsampling matrix. This allows to directly solve the problem of undefined pixels in the SR initialization phase. It also leads to a more accurate and robust estimation of the motion $\hat{\mathbf{W}}_k$ as it is now computed between $\mathbf{y}_k \uparrow$ and $\mathbf{y}_0 \uparrow$. The following registration of frames to the reference is consequently enhanced:

$$\overline{\mathbf{y}}_k \uparrow = \hat{\mathbf{W}}_k^{-1}\mathbf{y}_k \uparrow . \qquad (3)$$

Without loss of generality, both $\mathbf{H}$ and $\mathbf{W}_k$ are assumed to be block circulant matrices. Choosing the upsampling matrix $\mathbf{U}$ to be the transpose of $\mathbf{D}$, the product $\mathbf{UD} = \mathbf{A}$ defines a new block circulant blurring matrix $\mathbf{B} = \mathbf{AH}$. We have, therefore, $\mathbf{BW}_k = \mathbf{W}_k\mathbf{B}$. As a result, the estimation of $\mathbf{x}$ may be decomposed into two steps; estimation of a blurred HR image $\mathbf{z} = \mathbf{Bx}$, followed by a deblurring step. The data model in (1) becomes

$$\overline{\mathbf{y}}_k \uparrow = \mathbf{z} + \boldsymbol{\nu}_k, \qquad k = 0, ..., (N-1), \qquad (4)$$

Algorithm 1: ***UP-SR***

1. Choose the reference frame $\mathbf{y}_0$.
for k, s.t., $k = 1, \cdots, N$,
do
2. Compute $\mathbf{y}_k \uparrow$ using (2).
3. Find $\hat{\mathbf{W}}_k$ by optical flow estimation.
4. Compute $\overline{\mathbf{y}}_k \uparrow$ using (3).
end do
end for
5. Find $\hat{\mathbf{z}}$ by applying a median estimator (5).
6. Deduce $\hat{\mathbf{x}}$ by deblurring using (6).
end for

Table 1. Classical Upsampling for Precise Super-Resolution

where $\boldsymbol{\nu}_k = \hat{\mathbf{W}}_k^{-1}\mathbf{U} \cdot \mathbf{n}_k$ is an additive noise vector of length n. Using an L_1–norm $\| \cdot \|_1$, the estimate of $\mathbf{z}$ using the corresponding Maximum Likelihood is

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \sum_{k=0}^{N-1} \|\mathbf{z} - \overline{\mathbf{y}}_k \uparrow \|_1. \qquad (5)$$

The result in (5) is, by definition, the pixel-wise temporal median estimator $\hat{\mathbf{z}} = \text{med}_k\{\overline{\mathbf{y}}_k \uparrow\}$.

To recover $\hat{\mathbf{x}}$ from $\hat{\mathbf{z}}$, an iterative optimization is performed as a deblurring step. Considering a regularization term $\Gamma(\mathbf{x})$, chosen to be the bilateral total variation (Bilateral TV) given in [13], we find

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\text{argmin}}\Big(\|\mathbf{Bx} - \hat{\mathbf{z}}\|_1 + \lambda\Gamma(\mathbf{x})\Big), \qquad (6)$$

where λ is the regularization parameter. The *UP-SR* algorithm is given in Table 1, and summarized in Figure 1.

3. Surface Upsampling for Precise Super-Resolution

The different steps in *UP-SR* as described in Section 2 may be directly applied on LR depth images $\mathbf{y}_k$ of faces as those illustrated in Figure 2 (a). The resulting reconstructed face $\mathbf{x}$ is shown in Figure 2 (c). While it is of higher resolution, it presents artifacts that we argue are caused by applying UP-SR on gridded depth data. To remedy these artifacts, we propose in what follows to back–project the $\mathbf{y}_k$ frames, $k = 1, \cdots, N$, to $\mathbb{R}^3$ using the intrinsic parameters of the camera used for the acquisition. We end up with N corresponding point clouds $\mathcal{Y}_k = \{\mathbf{p}_i^k = (x_i^k, y_i^k, z_i^k) \in \mathbb{R}^3, i = 1, \cdots, m\}$ as shown in Figure 2 (b). The objective is now to reconstruct an HR point cloud $\mathcal{X} = \{\mathbf{q}_i^k = (x_i^k, y_i^k, z_i^k) \in \mathbb{R}^3, i = 1, \cdots, n\}$ belonging to the surface $\mathcal{S}$ of the original face, i.e., $\mathcal{X} \subset \mathcal{S}$. We adapt the algorithm in Table 1 to point clouds, and define a modified version of the *UP-SR* algorithm that we refer to as *SurfUP-SR*. The two main

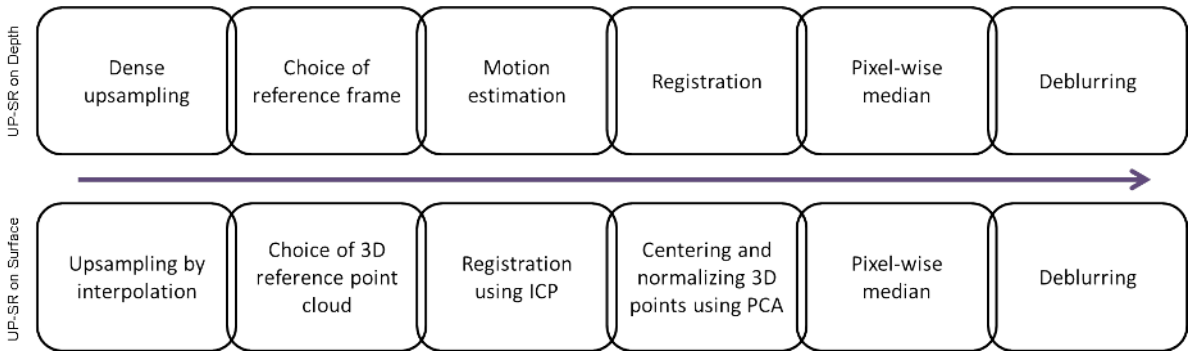


Figure 1. *UP-SR* steps on depth data and on a surface in 3-D.

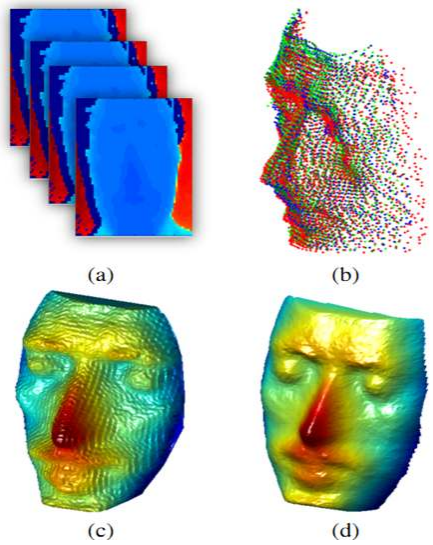


Figure 2. Face reconstruction with *UP-SR* using (a) depth images, (b) point clouds. The corresponding results are shown in (c) and (d), respectively.

phases are maintained: 1) estimation of $\mathcal{Z}$, a blurred version of $\mathcal{X}$; 2) deblurring by optimization as in (6). The steps of upsamling and registration need to be adapted as described in the following sections. An illustration of differences between *UP-SR* and *SurfUP-SR* is given in Figure 1.

3.1. Surface Upsampling

Assuming that the point cloud $\mathcal{Y}_k$ is a sampling of a surface $\mathcal{S}_k$, the upsampling of $\mathcal{Y}_k$ may be reformulated as a problem of interpolating the surface $\mathcal{S}_k$ from scattered points. The surface $\mathcal{S}_k$ may be defined implicitly by a function f as: $f(x, y, z) = 0, \quad \forall\ \mathbf{p} = (x, y, z) \in \mathcal{S}_k$, or equivalently by using the interpolant $\mathcal{P}_f$ as:

$$\mathcal{P}_f\,(x, y) = z, \qquad \forall\ \mathbf{p} = (x, y, z) \in \mathcal{S}_k. \tag{7}$$

The m points in $\mathcal{Y}_k$ verify (7), hence they form a system of m equations, from which $\mathcal{P}_f$ may be defined. A solution using kernel regression has been proposed in [14]. An efficient GPU implementation has been given in [15]. We used the Matlab `scatteredInterpolant` function in our implementation. Once $\mathcal{P}_f$ is found, it is used to define $(r-1)\cdot m$ additional points belonging to $\mathcal{S}_k$ for chosen (x, y)-positions. As a result, a denser point cloud $\mathcal{Y}_k\uparrow$ containing a total of n points is found such that

$$\mathcal{Y}_k\uparrow = \mathcal{Y}_k \cup \{\mathbf{p}_i^k = (x_i^k, y_i^k, z_i^k) \in \mathbb{R}^3, i = m+1, \cdots, n\}, \tag{8}$$

and $(x_i^k, y_i^k) \in [-1, 1] \times [-1, 1]$.

3.2. Surface Registration

The motion estimation and registration steps in *UP-SR* are replaced by directly using classical 3-D point cloud registration techniques. We use iterative closest points (ICP) to rigidly register each point cloud $\mathcal{Y}_k\uparrow$ to the reference $\mathcal{Y}_0\uparrow$. This is done by estimating the optimal transformation parameters, namely, 3-D rotation $\hat{\mathbf{R}}_k$, translation $\hat{\mathbf{t}}_k$, and global scaling factor $\hat{\alpha}_k$ that minimize the distance $Err(\cdot)$ between the transformed and the reference point clouds such that

$$[\hat{\mathbf{R}}_k, \hat{\mathbf{t}}_k, \hat{\alpha}_k] = \underset{\mathbf{R},\mathbf{t},\alpha}{\operatorname{argmin}} Err\left(\alpha\mathbf{R}\mathcal{Y}_k\uparrow + \mathbf{t}, \mathcal{Y}_0\uparrow\right). \tag{9}$$

The registered point cloud $\overline{\mathcal{Y}}_k\uparrow$ is then computed as:

$$\overline{\mathcal{Y}}_k\uparrow = \hat{\alpha}_k\hat{\mathbf{R}}_k\mathcal{Y}_k\uparrow + \hat{\mathbf{t}}_k. \tag{10}$$

With these modifications, the new *SurfUP-SR* algorithm is given in Table 2. Its visual impact is shown in the example of Figure 2 (d).

4. Proposed Face Recognition Pipeline

Our proposed pipeline is composed of three main stages: preprocessing of raw data, feature extraction and matching.

4.1. Preprocessing

The preprocessing step is an essential step in the design of a face recognition system as it affects the performance of

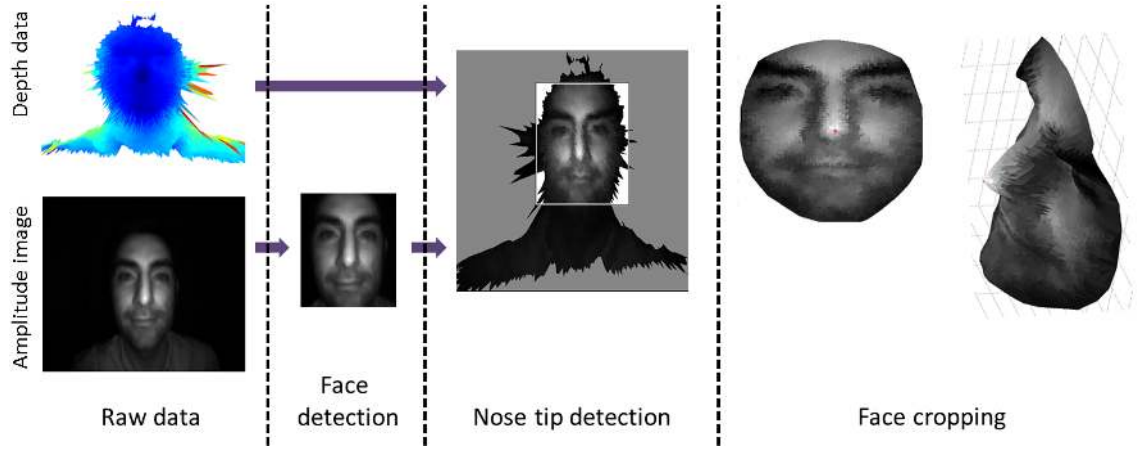


Figure 3. Preprocessing step of the facial acquisition pipeline using a depth camera.

Algorithm 2: ***SurfUP-SR***

1. Choose the reference frame $\mathcal{Y}_0$.

for k, s.t., $k = 1, \cdots, N$,
do

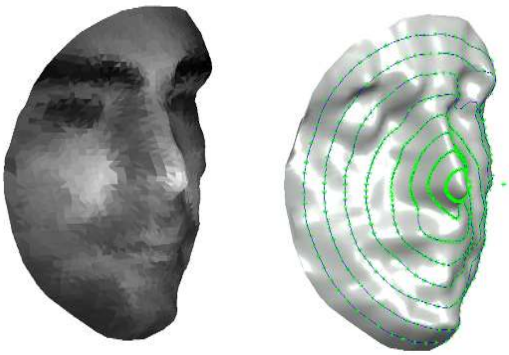
2. Compute $\mathcal{Y}_k \uparrow$ using (8).
3. Estimate $\hat{\mathbf{R}}_k$, $\hat{\mathbf{t}}_k$, and $\hat{\alpha}_k$ using ICP as in (9).
4. Compute $\overline{\mathcal{Y}}_k \uparrow$ using (10).

end do
end for

5. Find $\hat{\mathcal{Z}}$ by applying a median estimator (5).
6. Deduce $\hat{\mathcal{X}}$ by deblurring using (6).

end for

Table 2. Surface Upsampling for Precise Super-Resolution



(a) (b)

Figure 4. Feature extraction step using: (a) Observed LR 3-D face with texture from amplitude or 2-D images. (b) Extracted level curves.

the system significantly. We implement fast and efficient techniques to detect the face region and the nose tip for an effective segmentation and alignment. We apply a face detection algorithm on the amplitude or 2-D image only, then we map the face region with the corresponding depth image to obtain the corresponding 3-D facial region. In this work, the Viola-Jones [19] face detection algorithm is used for its computational efficiency and high detection rate. Once we detect the depth face region, we detect the nose tip represented by the point with the smallest depth value. The nose tip is used as a basic feature for our segmentation and alignment. Using a spherical cropping centered at the nose tip, we discard the ear, hair and part of the neck areas. Finally, the ICP registration is used for alignment.

4.2. Feature extraction

We use spherical curves and their radial discretization as features to represent each face. A spherical curve is obtained by intersecting the facial surface with a sphere. In order to have smoother and continuous curves, we apply the interpolation technique proposed in [18]. Spherical curves are discretized radially by slicing the spherical intersection curves using a plane that is parallel to the face normal and that intersects the spherical curves radially at uniform angles. Each face is represented by an indexed collection of $M \times L$ points in 3-D, where M denotes the number of curves per sample face and L is the number of points in each curve. We end up with a feature vector of size $M \times L \times 3$ for each face. An example of the extracted feature curves is shown in Figure 4.

4.3. Matching

The matching step aims to associate each probe 3-D face to the the closest 3-D face in the database by comparing their extracted features. The comparison is carried out by an appropriate distance measure on the space of the extracted feature curves. We choose the cosine distance in our experiments as we found it to be the best performing one. This is confirmed by the survey of Smeets et al. [20].

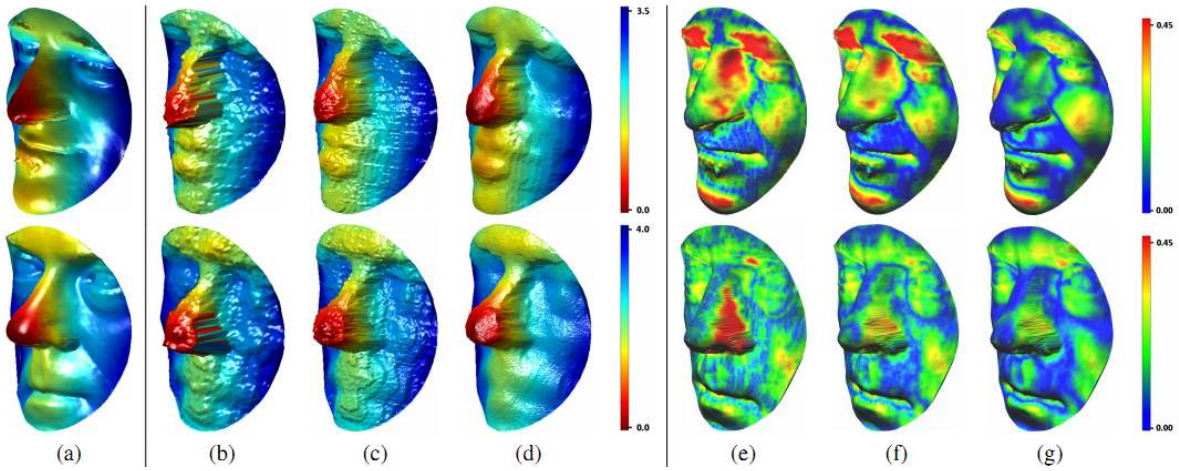


Figure 5. 3-D face reconstruction results. (a) 3-D laser scan ground truth. (b) One of the LR 3D faces. (c) Results of the *superfaces* algorithm. (d) Results of the proposed *SurfUP-SR* algorithm. (e) 3-D error map corresponding to the 3-D LR face. (f) 3-D error map corresponding to the *superfaces* results. (g) 3-D error map corresponding to the proposed *SurfUP-SR*.

.

5. Experimental Part

We evaluate the performance of the proposed system for both 3-D face reconstruction and recognition. First, to evaluate the quality of the reconstructed 3D faces, we use the publicly available *superfaces* dataset [17]. It has been acquired using the well known Kinect camera [4]. A sequence of 2-D and depth images for 20 different subjects are provided. Moreover, an HR scanned version for each subject is available as ground truth. The dataset has only one realization for each subject which makes it not appropriate for recognition purposes. Thus, we built our real dataset using 10 subjects with two different realizations for each subject. The dataset is acquired using the PMD camboard nano time of flight camera with a resolution of (120×165) pixels [3].

5.1. Reconstruction

In order to evaluate the quality of the reconstructed faces, we use the above mentioned real dataset [17]. The faces in the depth frames are of low resolution due to the object distance from the camera. To improve its quality, we conduct the following test. We apply *SurfUP-SR*, and show the results for two subjects (01 and 19) using 5 LR frames. An LR frame for each subject is shown in Figure 5.(b), first and second rows, respectively. Obtained results show that the proposed algorithm provides a visually improved HR 3-D faces as seen in Fig. 5.(d) as compared to the LR captured data Figure 5.(b). Moreover, our algorithm provides better visual results than the recently proposed *superfaces* algorithm [9], Figure 5. (c). This is due to the fact that *SurfUP-SR* includes an additional deblurring step. Our results are of sufficient quality for many applications such as 3-D face recognition. In order to provide a quantitative evaluation, we measure the reconstruction error of *SurfUP-SR* and *superfaces* against the laser scanned ground truth. In Figure 5. (f) and (g), we may see the color-coded reconstruction error of the *superfaces* method [9] and *SurfUP-SR*, respectively. As expected, obtained results show that *SurfUP-SR* is at least as good as *superfaces* and sometimes better. Moreover, by taking a look to the error range bar in Figure 5, we note that in most areas the errors are below 0.5 cm.

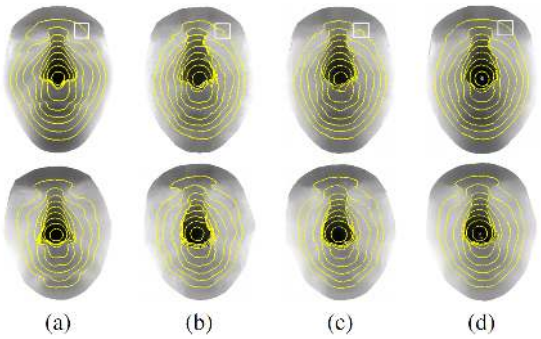
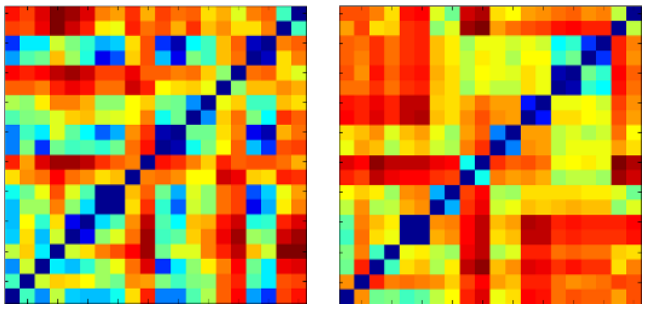


Figure 6. Extracted level curves from 3-D faces for: (a) Ground truth. (b) LR. (c) *superfaces*. (d) *SurfUP-SR*.

5.2. Recognition

In order to test the impact of *SurfUP-SR* on a face recognition algorithm, we evaluate the performance of the pipeline presented in Section 4 on the raw LR faces in our database. We then run the same pipeline on the superresolved faces of our database. We may see in Figure 6 the enhancement incurred by *SurfUP-SR* on the quality of the extracted feature curves. Indeed, their extraction from LR faces leads to noisy curves. For the same subject, these curves become



(a) (b)

Figure 7. Confusion matrices. (a) Using the LR 3-D observed faces. (b) Using the super-resolved 3-D faces by the proposed *SurfUP-SR*.

smoother and less noisy if extracted from superresolved data. The quality of these curves directly affects the final result of the face recognition algorithm. The corresponding confusion matrices are given in Figure 7(a) and in Figure 7(b). We notice an improved recognition rate from 50% to 80% when super-resolving. This confirms the importance of having a higher resolution for an increased recognition rate and the effectiveness of the proposed *SurfaceUP-SR*.

6. Conclusion

In this paper we proposed a new multi-frame super-resolution algorithm *SurfUP-SR* which improves 3-D face recognition rate using low resolution, and cost-effective depth cameras. We reformulated the *UP-SR* algorithm on a 3-D point cloud instead of its original formulation on a depth image. In addition, we provided a full automatic 3-D face acquisition from depth cameras. Experimental evaluation of *SurfUP-SR* using a real low resolution 3-D face dataset has been carried out. Obtained results show an efficient enhancement in the resolution and the quality of the captured low resolution 3-D faces. Moreover, we showed the impact of the proposed algorithm in decreasing the 3-D reconstruction error, and most importantly in increasing the 3-D face recognition rate.

References

[1] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, M. Gross, "High-Quality Single-Shot Capture of Facial Geometry," ACM Trans. on Graphic, vol. 29, pp. 40:1–40:9, 2010.

[2] http://www.iee.lu/technologies

[3] http://www.pmdtec.com

[4] http://www.primesense.com/

[5] F. Lin, C. Fookes, V. Chandran, S. Sridharan, "Super-resolved faces for improved face recognition from surveillance video," In Advances in Biometrics, pp. 1-10. Springer Berlin Heidelberg, 2007.

[6] C. Fookes, F. Lin, V. Chandran, S. Sridharan, "Evaluation of image resolution and super-resolution on face recognition performance," Journal of Visual Communication and Image Representation, vol. 23, no. 1, pp. 75-93, 2012.

[7] K. Al Ismaeil, D. Aouada, B. Mirbach, B. Ottersten, "Depth super-resolution by enhanced shift & add," In Proceedings of the 15th International Conference on Computer Analysis of Images and Patterns, pp. 27-29, 2013.

[8] K. Al Ismaeil, D. Aouada, B. Mirbach, B. Ottersten, "Dynamic super-resolution of depth sequences with non-rigid motions," In Proceedings of the IEEE International Conferebce on Image Processing, pp. 15-18, 2013.

[9] S. Berretti, A. Del Bimbo, P. Pala, "Superfaces: A Super-resolution Model for 3D Faces," 5th Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment, vol. 7583/2012, pp.73-82, 2012.

[10] M. Ebrahimi, E. Vrscay, "Multi-frame super-resolution with no explicit motion estimation," In Proceedings of the IEEE International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), pp. 455459, 2008.

[11] S. Peng, G. Pan, Z. Wu, "Learning-based super-resolution of 3D face model," In Proceedings of the IEEE International Conferebce on Image Processing (ICIP), vol. 2, pp. 382385, 2005.

[12] G. Pan, S. Han, Z. Wu, Y. Wang, "Super-Resolution of 3D Face," IEEE European Conference on Computer Vision, vol. 3952, pp. 389401, 2006,

[13] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Fast and Robust Multi-Frame Super-Resolution," IEEE Transactions on Image Processing, vol. 13, pp. 1327-1344, 2004.

[14] H. Takeda, S. Farsiu, P. Milanfar, "Kernel regression form image processing and reconstruction," IEEE Transactions on Image Processing, vol. 16, no. 2, pp. 349-365, 2007.

[15] S. Cuomo, A. Galletti, G. Giunta, A. Starace, "Surface reconstruction from scattered points via RBF interpolation on GPU," Federated Conference on Computer Science and Information Systems, vol. abs/1305.5179, pp. 433-440, 2013.

[16] M. Hernandez, J. Choi and Grard Medioni, "Laser Scan Quality 3-D Face Modeling Using a Low-Cost Depth Camera," In Proceedings of the IEEE European Signal Processing Conference (EUSIPCO), pp. 1995-1999, 2012.

[17] http://www.micc.unifi.it/vim/datasets/4d-faces/

[18] D. Aouada, H. Krim, "Squigraphs for fine and fompact modeling of 3-D shapes," IEEE Transactions on Image Processing, vol. 19, pp. 306-321, 2010.

[19] P. Viola, M. Jones, " Rapid object detection using a boosted cascade of Simple features," In Proceeding of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 511-518, 2001.

[20] D. Smeets, P. Claes, J. Hermans, D. Vandermeulen, P. Suetens, "A Comparative Study of 3-D Face Recognition Under Expression Variations," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.42, no.5, pp.710,727, Sept. 2012.