Modeling 3D faces from samplings via compressive sensing

Qi Sun[1)], Yanlong Tang[1)], Ping Hu[2)]
{nowhereman.sq, yanlongtang, peggyhu0315}@gmail.com

1) Taishan College, Shandong University
2) School of Computer Science and Technology, Shandong University

ABSTRACT

3D data is easier to acquire for family entertainment purpose today because of the mass-production, cheapness and portability of domestic RGBD sensors, e.g., Microsoft Kinect. However, the accuracy of facial modeling is affected by the roughness and instability of the raw input data from such sensors. To overcome this problem, we introduce compressive sensing (CS) method to build a novel 3D super-resolution scheme to reconstruct high-resolution facial models from rough samples captured by Kinect. Unlike the simple frame fusion super-resolution method, this approach aims to acquire compressed samples for storage before a high-resolution image is produced. In this scheme, depth frames are firstly captured and then each of them is measured into compressed samples using sparse coding. Next, the samples are fused to produce an optimal one and finally a high-resolution image is recovered from the fused sample. This framework is able to recover 3D facial model of a given user from compressed simples and this can reducing storage space as well as measurement cost in future devices e.g., single-pixel depth cameras. Hence, this work can potentially be applied into future applications, such as access control system using face recognition, and smart phones with depth cameras, which need high resolution and little measure time.

Keywords: Kinect, super-resolution, compressive sensing, 3D face modeling, low-sampling rate requirement.

1. INTRODUCTION

3D domestic sensors have attracted great concentration both in academic and industrial field, as it can be used for 3D modeling, personal avatar generation, et al. However, the stability and robustness of such depth data captured using Kinect is limited because of the low resolution and dithering of IR projector. These problems consequently affect a series of related researches on computer vision, graphics and animation: The results of some works on modeling [1][2][3] are not quite acceptable, which results from the limitation of the device. Meanwhile, although some methods [4] on 3D resolution enhancement were implemented in Kinect [5], their experiments are computationally complex. Consequently, when it comes to large scale image, these algorithms can only be implemented in workstation computer.

As is reported that depth cameras, like Kinect, will probably be integrated into small mobile devices (for example smart phones), which have small storage space and limited computation ability, there is a demand that users may prefer to capture high quality depth face images (without losing facial details) with small storage devices and inaccurate depth sensors. And also, Kinect has potential application in face recognition, which has high requirement for the resolution of depth face images, measurement time (time cost of acquiring depth data) of human face and storage of the point clouds. Considering such potential demands, we aims to develop a facial-based depth sup-resolution method, which can generate high-resolution depth face while reducing potential measurement time and saving storage space.

Our work designs a novel scheme of facial modeling using compressive sensing (CS) theory. Compared with former facial modeling method, we make a series of contributions as follows:

(1)Our scheme produced high resolution depth human face with less noisy data than other methods.

(2)This method reduces measurements (which may potentially used to design single-pixel camera) and has comparable running time.

(3)We successfully reduced storage cost by saving the compressed data in prior of producing high-resolution depth face.

2. OUR SCHEME

Figure 1 shows our low storage cost super-resolution framework: Several frames of human frontal face are segmented from depth frames when a user slightly rotates head. These face frames with small difference are then temporarily enhanced to high resolution respectively. Next, They each will be compressed by compressive sensing (CS) and a optimal sample, which is the storage sample corresponded to the final optimal high resolution depth image, is obtained by combining these compressed samples. Finally the optimal high resolution depth image of human face is recovered from the optimal sample stored in device. And after smoothing the mesh, the lifelike result can be displayed.

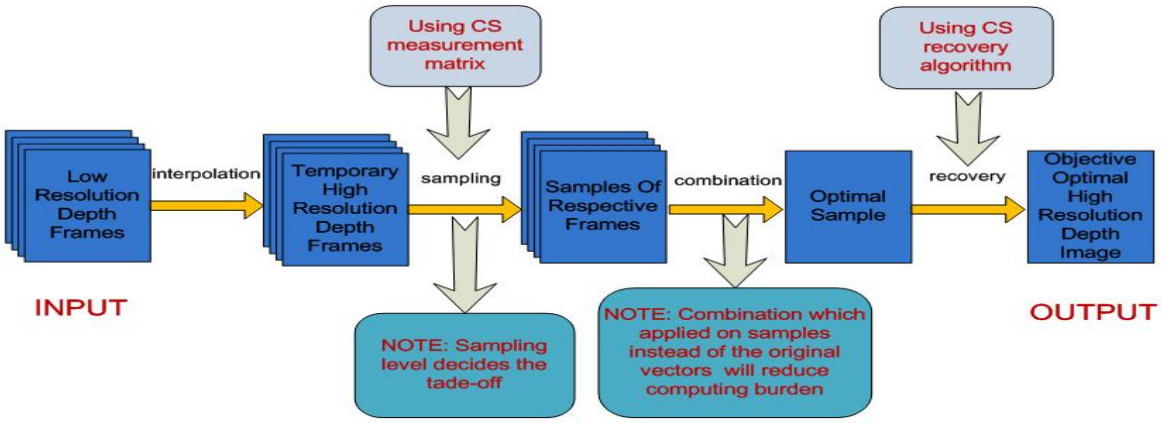


Fig.1. Our proposed scheme.

2.1 Raw data pre-processing

We introduce BIWI dataset on 3D heads with different poses captured by Kinect built by Fanelli et al. [6][7]. We first apply a part of Sun et al.'s pre-processing [8] steps including iterative closest point (ICP) algorithm to align each frame to the central one and then use the nearest neighbor interpolation (NNI) method to temporarily enhance each frame' s resolution at the same scale.

2.2 Compress each frame via sparse coding

In this step, all the rough high-resolution frames are compressed using CS method. Actually, why here CS is introduced is that future devices will capture these compressed data directly with this method, which will significantly reduce measuring time.

2.2.1 Compression theory (Compressive Sensing)

Compressive Sensing overcomes traditional compression methods' limitation of taking extra needless measurements by directly acquiring the needed compressed measurements. Here is a brief introduction to the CS theory. An nature signal X (such as sound wave, image) usually has a sparse representation $\mathbf{S}$ (all the element but several ones are zero or very close to zero) in a basis or dictionary $\boldsymbol{\Psi}$ (frequency domain or time/space domain). This can be formulated as $\mathbf{X} = \boldsymbol{\Psi}\,\mathbf{S}$. For example, here $\mathbf{X}$ and $\mathbf{S}$ are both N*1 vectors, $\boldsymbol{\Psi}$ is an N*N matrix and $\mathbf{S}$ is K-sparse (only K out of N elements of $\mathbf{S}$ are not zero). Then X can be directly projected into a down sample vector $\mathbf{Y}$ in a particular way, with the projection matrix $\boldsymbol{\Phi}$ specially designed. This is given by $\mathbf{Y} = \boldsymbol{\Phi}\,\mathbf{X}$, where $\boldsymbol{\Phi}$ is an M*N matrix and Y is an M*1 vector (M<<N, this is why it can reduce measuring time). It should be point out that measurements are reduced using down-sampling like method. Thus we can recover $\mathbf{X}$ with the down sampling vector $\mathbf{Y}$ by solving $\mathbf{arg\,min}\,||\mathbf{S}||_{\mathbf{0}}$, subject to $\mathbf{Y} = \boldsymbol{\Phi}\boldsymbol{\Psi}\,\mathbf{S}$. This $\mathbf{L_0}$ form solution has been proved to be very slow or NP hard. However, the work [9] has proved that if $\boldsymbol{\Phi}$ satisfied Restricted Isometry Property (RIP), the problem can be equivalently solved by using $\mathbf{L_1}$ form, which is successful as well as efficient. It goes like: $\mathbf{arg\,min}\,||\mathbf{S}||_{\mathbf{1}}$, subject to $\mathbf{Y} = \boldsymbol{\Phi}\boldsymbol{\Psi}\,\mathbf{S}$. And the work [10] also indicates that if $\boldsymbol{\Phi}$ is chosen as a Random Gaussian Matrix and $\mathbf{M} = \mathbf{O}\left(\mathbf{K}\,\mathbf{log}\left(\frac{\mathbf{N}}{\mathbf{K}}\right)\right) \ll \mathbf{N}$, there is high probability that $\boldsymbol{\Phi}$ satisfy RIP condition. Then some $\mathbf{L_1}$ programming algorithms can be proposed to recover sparse representation $\mathbf{S}$. Finally, original signal $\mathbf{X}$ is recovered from$\mathbf{X} = \boldsymbol{\Psi}\,\mathbf{S}$.

2.2.2 Compression using CS theories

The temporally interpolated high-resolution depth frames (totally we selected k frames) are represented in the one column vector (N*1) forms, $\mathbf{X_i}$ (i =1… k). As all the frames $\mathbf{X_i}$ are very similar, we measure each of the frames with the same

measurements matrix $\mathbf{\Phi}$ and basis matrix $\mathbf{\Psi}$, where $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are M*N and N*N dimensions respectively. Thus the corresponding down sampling measurements $\mathbf{Y_i}$ of each frame can be obtained in the form: $\mathbf{Y_i} = \mathbf{\Phi}\,\mathbf{X_i}$. In this way, each frame is compressed directly without discarding the redundant measurements that carry little information.

2.3 Fuse compressed frames

As each compressed frame $\mathbf{Y_i}$ carries slightly different information, by fusing these samples, an optimal sample $\mathbf{Y}$, with most information, can be obtained for storage to recover a high-resolution depth face image later. The measurements vectors combination procedure works like this [11]: Single-level 1D Discrete wavelet transform is applied on each sample vector $\mathbf{Y_i}$, and each vector can be separated into approximation coefficients $\mathbf{cA_i}$(i=1,...,k) and $\mathbf{cD_i}$(i=1,...,k). As larger coefficient carries more information, we respectively combine approximation coefficients and detail coefficients using weight mean. Thus, the combined approximation coefficient **cA** and detail coefficient **cD** can be obtained by solving:

$$\mathrm{cA} = \sum_{i=1}^{k} w_i * \mathrm{cA}_i \quad \text{and} \quad \mathrm{cD} = \sum_{i=1}^{k} v_i * \mathrm{cD}_i$$

Where

$$w_i = \frac{|\mathrm{cA}_i|}{\sum_{j=1}^{k} |\mathrm{cA}_j|} \quad \text{and} \quad v_i = \frac{|\mathrm{cD}_i|}{\sum_{j=1}^{k} |\mathrm{cD}_j|}$$

Then the fused optimal sample **Y** can be obtained using discrete wavelet transform from **cA** and **cD**.

2.4 Recover HR depth image

As **Y** is an optimal measurements vector, which carries fused information of all the initial measurements vectors Y_i (i=1… k), it will recover a optimal High-Resolution depth image X, which outperforms the images recovered from single measurement simple and average of simples, respectively. SAMP algorithm [12], based on L_0 form pursuit, is introduced to solve the ill-posed recovery problem $\mathbf{arg\,min}\,||\mathbf{S}||_0$, subject to $\mathbf{Y} = \mathbf{\Phi\Psi\,S}$. Then **X** is recovered from $\mathbf{X} = \mathbf{\Psi\,S}$. SAMP method has advantages that the only input are $\mathbf{Y}, \mathbf{\Phi}, \mathbf{\Psi}$ and two threhold parameters, while the sparse level K needs not to be known in prior and sometimes it has better performance than L_1 programming.

Finally we apply the classical smooth method, Laplacian smoothing to **X** and get a smooth and acceptable result.

3. EXPERIMENTS

In this section, we perform comparison experiments with a particular sampling level to test our scheme. We implemented our scheme using Biwi Kinect Head Pose Database [6][7] as the input data.

(1) From the whole database, we choose four frames of the same human face as the input data to test out scheme. Each input frame is of resolution 128*128, and they are captured from multiple views as the user slightly rotates her head (See Figure 2).

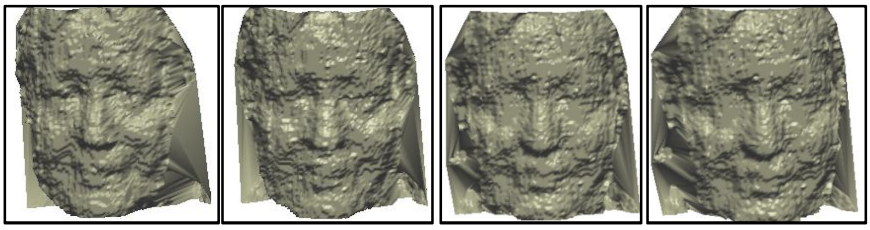
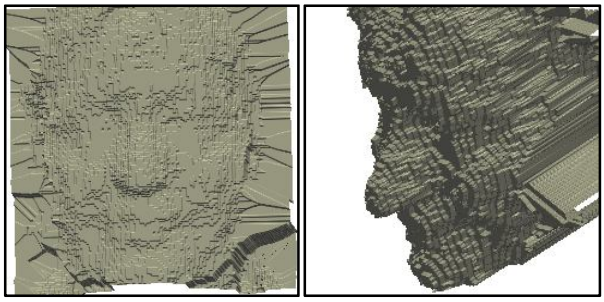


Fig.2. Four depth frames of low resolution128*128.

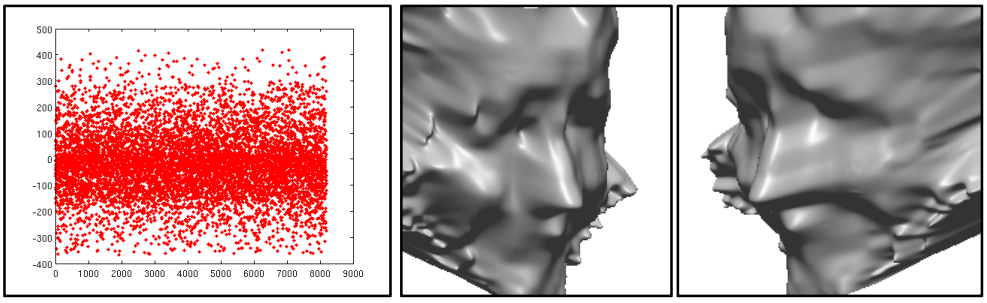
(2) Then after ICP alignment, the temporarily resolution enhancement result is shown in Figure 3(a) (b). Each aligned frame's resolution has been enhanced to 512*512. It can be observed that facial details lose after interpolation and it doesn't matter as some details will be recovered after fusing their compressed simples in the coming steps. The data before compression are these 4 frames represented in the one column vector form $\mathbf{X_i}(\,i = 1,2,3,4\,)$ and $\mathbf{N} = \mathbf{512^2}$.



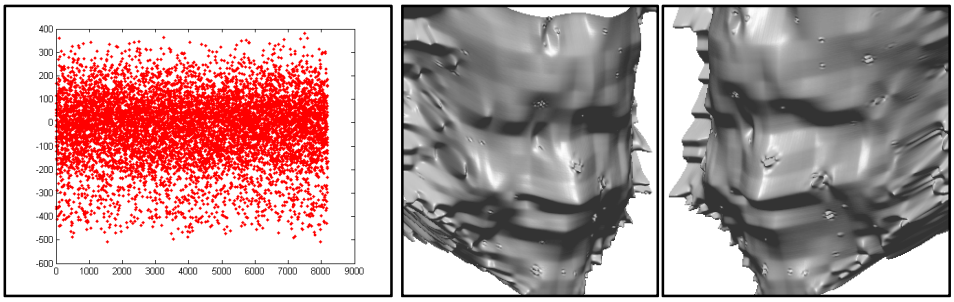
(a) (b)

Fig.3. One single frame (512*512) after resolution enhancement (a) front view (b) side view.

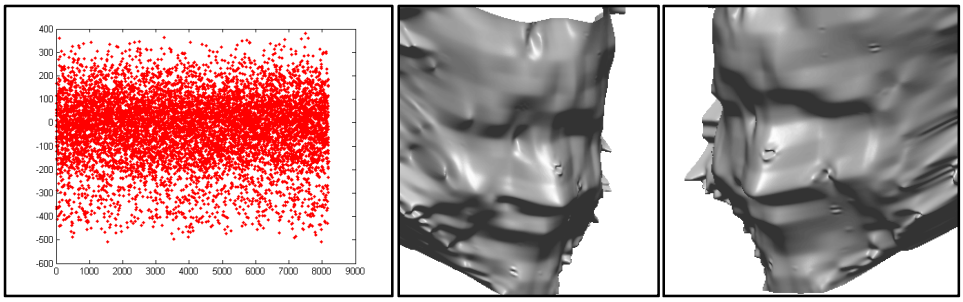
(3) During the compression procedure, the dimension of each sample **M** is assigned with value such that the sampling level M/N = 0.3125. And the measurement matrix **Φ** is Structurally Random Matrices for sensing operator and the sparsity matrix **Ψ** is wavelet transform matrix. To evaluate the sample-fusion effectiveness, we perform the recovery procedure with fused sample, single sample and average of samples (mean of four compressed simples) respectively and the comparison of the results is shown in Figure 4 (a) (b) (c). The visual results indicate that image recovered from fused sample is of better quality than the other two.



(a) Recovery results from fused sample (proposed sampling method).



(b) Recovery results from single sample.



(c) Recovery results from average of samples.

Fig.4. Recovery results (from left to right: samples, recovered image (left view) and recovered image (right view) respectively). By comparing the recovery quality (the more scar on the face, the lower quality), our fused sample recovers better quality face than that of the other two methods.

4. CONCLUSIONS

In this paper, we present a novel scheme to model high-resolution 3D face while saving storage space using Kinect as a sensor. Traditional super-resolution algorithm Lidarboost [4], can also produce high-resolution depth image, but it needs huge computational cost, can only deal with low resolution depth images, and cannot reduce measuring time and storage cost of the point clouds. Our facial modeling scheme overcomes these limitations with a few steps. First, low-resolution depth frames are simply interpolated into temporary high-resolution frames which may be not lifelike. Then these pro-processed frames are transformed into rough samples using compressive sensing. Next, the samples are combined to form a new sample vector with better quality for storage in disk. Finally, from the optimal fused sample vector, we get the recovered high-resolution depth face image using state-of-the-art SAMP recovery algorithm. Our method can recover high-resolution face image with less noisy data from stored compressed sample with comparable computational cost. And most importantly, by using compressive sensing technique, the compressed sample may be acquired during the measuring procedure in future devices and this means reducing measuring time as well as storage space. Considering these features of our scheme, our work can be applied to Kinect's potential future applications (such as depth cameras in smart phones, access control system using Kinect for face recognition) which require higher resolution, shorter measuring time in data acquisition procedure, and smaller storage space.

REFERENCES

[1] Tong, J., Zhou, J., Liu, L., Pan, Z., and Yan, H., "Scanning 3D full human bodies using Kinects", IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE Virtual Reality), 18(4), 643-650 (2012).

[2] Zollh¨ofer, M., Martinek, M., Greiner, G., Stamminger,M., and S¨ussmuth, J., " Automatic reconstruction of personalized avatars from 3D face scans", Journal of Visualization and Computer Animation 22, 2-3, 195–202(2011).

[3] Blanz, V., and Vetter, T., "A morphable model for the synthesis of 3D faces", In SIGGRAPH, 187–194 (1999).

[4] Schuon, S., Theobalt, C., Davis, J., and Thrun, S., "Lidarboost: Depth super resolution for toF 3D shape scanning", In CVPR, 343–350 (2009).

[5] Cui, Y., and Stricker, D., "3D shape scanning with a Kinect", In SIGGRAPH Posters, ACM, 57 (2011).

[6] Fanelli. G., Dantone, M., Fossati, A., Gall, J., and Van Gool, L., "Random forests for real time 3D face analysis", International Journal of Coputer Vision (2012).

[7] Fanelli, D., Weise, Y., Gall, J., and Van Gool, L., "Real time head pose estimation from consumer depth cameras", 33rd Annual Symposium of the German Association for Pattern Recognition (2011).

[8] Sun, Q., Tang, Y., Hu, P., Peng, J., "Kinect-based automatic 3D high-resolution face modeling", International Conference on Image Analysis and Signal Processing (2012).

[9] Cand`es, E. J., Romberg, J. K., and Tao, T., "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information", IEEE Transactions on Information Theory 52, 2, 489–509 (2006).

[10] Donoho, D. L., "Compressed sensing", IEEE Transactions on Information Theory 52, 4, 1289–1306 (2006).

[11] Han, J., Loffeld, O., Hartmann, K., and Wang, R., "Multi image fusion based on Compressive sensing", Proc. Int. Con] on Image Processing, pp. 1463-1469 (2010).

[12] Do, T., Gan, L., Nguyen, N., Tran, T., "Sparsity adaptive matching pursuit algorithm for practical compressed sensing", Asilomar Conference on Signals, Systems & Computers - ASILOMAR , pp. 581-587 (2008).