3D Body Scanning With One Kinect

Yan Cui*, Didier Stricker*
DFKI, Augmented Vision, Kaiserslautern University
Germany

Abstract

In this paper we describe a method for 3D body scanning by aligning depth and color scans which were taken around a human body with a Kinect camera. The Kinect [18] camera is a “controller-free gaming and entertainment experience” by Microsoft for the Xbox 360 video game platform. It delivers depth and color scans at video rate. The proposed scanning solution makes 3D scanning technology more accessible to end-users, since it is easy-to-use and cost-effective. With this technique, 3D models could become a much more widely used asset, just as image and video data are today. This could open the door for many new applications, for instance in community web platforms or online shopping.

Keywords: body scanning, global alignment, ICP, non-rigid transformation, super-resolution, Kinect

1. Introduction

Nowadays, 3D geometry models of real world objects are essential for many application scenarios, such as design and virtual prototyping, quality assurance, or applications in visual media, such as games, virtual worlds and movie special effects. Existing 3D scanning technology are specialized and complex optical systems consisting of multiple cameras and structured light or lasers beamers. Even though they produce data of high quality, they are expensive and often require expert knowledge for their operation. They are reserved to the professional and semi-professional market and are not affordable by end-users. On the other hand, if easy-to-operate and cheap 3D scanners were more usable and low-cost, 3D models could turn into a much more widely used asset, just as image and video data are today. This could open the door to many new applications, for instance in community web platforms or online shopping.
In this paper, we propose a new easy-to-use 3D object scanning approach based on a single Kinect camera. The Kinect has a variety of advantages over existing 3D scanning technologies: It can measure depth and color at video rate simultaneously, what makes it very suitable for fast object scanning. The depth sensor does not interfere with the scene in the visual spectrum. Finally, its usage is not different from a video camera; end-users intuitively deal with it and are able to collect data with only a few explanations. However the development of a scanning system with a single Kinect is not straight forward and many algorithmic challenges must be overcome. The main two difficulties are the following:

1. The depth image has a low X/Y resolution and low accuracy, a strong random noise.
2. The single views must be aggregated and the model must be correctly closed after a 360° scan.

In this paper we show that an appropriate combination of methods for image enhancement and scan alignment allows the generation of a closed 3D shape model of a given object with reasonable quality. No specific capturing setting is required: the Kinect is hold in the hand and moved freely around the object.
The overall processing pipeline consists of three main steps, as illustrated in figure 2:

1. *Super-resolution*: the Kinect delivers synchronized RGB and depth images. Intermediate super-resolution color and depth frame are first created using the algorithm described in [2].
2. *Loop closing*: a probabilistic scan alignment method that extends [1] by explicitly performing 3D loop closure has been developed. This algorithm delivers closed 3D models while handling the inherent systematic bias on-the-fly.
3. *Non-rigid registration*: In order to correct residual errors due for example to small human’s arms and legs movements, a non-rigid registration is applied in the last step of the processing. Without assuming a systematic bias as in [1] we define different parameters for each pixel independently.

* {Yan.Cui / Didier.Stricker}@dfki.com; +49 (0)631 20 575 3620; http://av.dfki.de/

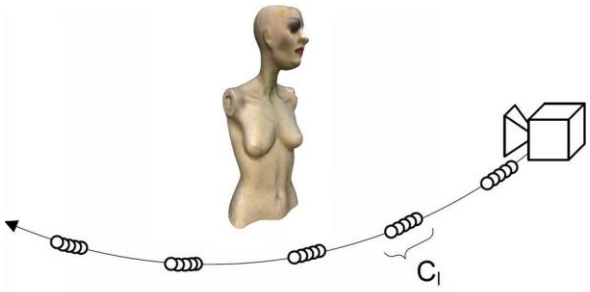


Fig1, A typical camera path: The dotted segments are the frame chunks C from which super-resolution depth scans are computed.

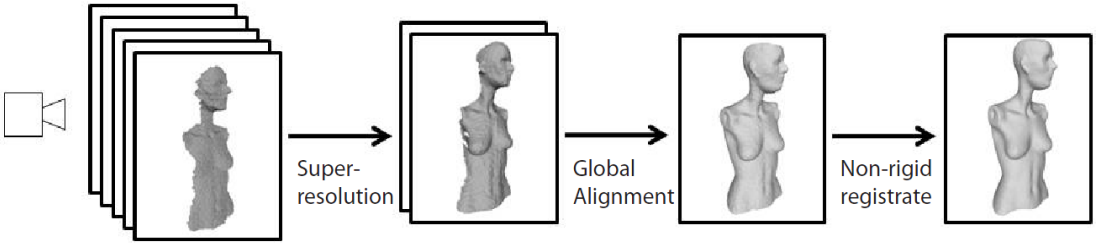


Fig2, Outline of our processing pipeline.

In section 2, we present the super-resolution processing pipeline in details; the global loop closure and non-rigid registration algorithms are described in section 3, and 4 respectively. Finally, reconstruction results are presented in section 5; the work is concluded in section 6.

2. Related work

Most commercial systems for 3D shape scanning are based on either active systems, such as structured light or laser stripes scanners, or passive image-based systems, which uses the images directly. A recent overview of the last approach can be found in [12][14] . The last category concerns the 3D or depth camera, which deliver depth images at video rate. In contrast to Time of Flight (ToF) cameras, the Kinect provides rather clean data of relatively low random noise and systematic error. With such data, local rigid alignment techniques, such as Iterative Closest Points (ICP) and its variants [5] or global rigid alignment techniques, [3][4][13] can be used to register the scans against each other. Then, a merging procedure [10] can be applied to build a single 3D mesh. Hand-held scanners have been proposed where the camera can be freely moved around an object (or vice versa), e.g. in [16]. Our work supports both hand-held scanning and scanning under controlled motion, with e.g. a turntable. A relatively simple 3D scanner has been proposed by Bouguet et al [6] who measures the 3D shape by recording a shadow casted by a rod moved over the object. However, this approach is suitable only for static set-up, and freely motion around an object is not possible. An alternative to active shape scanning are passive image-based approaches, such as stereo [17] or variants of shape-from-silhouette reconstruction [16]. Unfortunately, these algorithms often suffer from errors due to inaccurate image correspondences or the inability to capture concavities.
So far, Kinect cameras have not been used as sensors for 3D object scanning, even though they have a variety of advantages over the above technologies. This is mainly due to the high noise level which makes direct application of established filtering and alignment techniques infeasible. In this paper, we show that reliable shape capture can be achieved with the Kinect camera only. Related work to these methods is the method by Kil et al. [11]. Our approach however extends previous work on probabilistic non-rigid alignment of pairs of scans [15] into a global method. Suitable rigid and non-rigid scan alignment is achieved by explicitly incorporating specific noise characteristics, based on previous work [1].

3. Super-resolution

A super-resolution algorithm, similar the LidarBoost approach [2], is applied to each chunk of frames. The result is a high-resolution depth map aligned to the center frame of the chunk. In the following we briefly describe the core concepts of LidarBoost on which our new super-resolution approach relies, and refer the reader to [2] for more detail. We then extend the LidarBoost concepts with a new regularization framework that yields better results. First, all depth maps in the chunk are aligned to the center frame using 3D optical flow. This is sufficiently accurate since the maximum viewpoint displacement throughout the entire chunk is typically one to two depth pixels. LidarBoost method can extract a high-resolution denoised center depth map H_l. Based on LidarBoost, we add the color RGB information as a second data term to improve the depth super-resolution frame by solving an optimization problem of the form:

$$\min_{H_l} E_{depth\_data}\left(L_{\rho(l)},\ldots,L_{\rho(l)+\eta},H_l\right) + E_{reg}\left(H_l\right) \quad (1)$$

Here, $L_{\rho(l)},\ldots,L_{\rho(l)+\eta}$ are the raw depth maps of one chunk aligned to the center. E_{depth_data} measures the agreement of H_l with the aligned low resolution maps; unreliable depth pixels with low amplitude are discarded. E_{reg} is a feature-preserving smoothing regularization term tailored to raw data as in [2]. The proposed super-resolution approach is based on a similar energy and uses the same definition as LidarBoost:

$$E_{depth\_data} = \sum_{k=1}^{\eta} \left\| H_l - L_{\rho(l)+k} \right\|^2 \quad (2)$$

$$E_{reg}\left(H_l\right) = \sum_{u,v} \left\| \nabla H_{u,v} \right\|_2 = \sum_{u,v} \left\| \begin{pmatrix} G_{u,v}(0,1) \\ G_{u,v}(1,0) \\ \vdots \\ G_{u,v}(l,m) \end{pmatrix} \right\|_2 \quad (3)$$

And each $G_{u,v}(l,m)$ is a finite difference defined as follows:

$$G_{u,v}(l,m) = \frac{H(u,v) - H(u+l,v+m)}{\sqrt{l^2+m^2}} \quad (4)$$

Our implementation uses the Euler-Lagrange equation to transform the optimization problem into a linear equation system, which we solve using Gauss-Seidl method for linear systems. We have implemented this in C++ and Table 2 reports about the runtimes. Note the significant improvement in runtime of about 30 seconds compared to earlier LidarBoost implementations that took up to two hours for comparable scenes. The result of the super-resolution are shown in Figure 3, Mannequin and the woman sitting, left is the one raw frame data, right is the super-resolution result. Our algorithm can smooth the raw data, meanwhile, keep the detail structure.

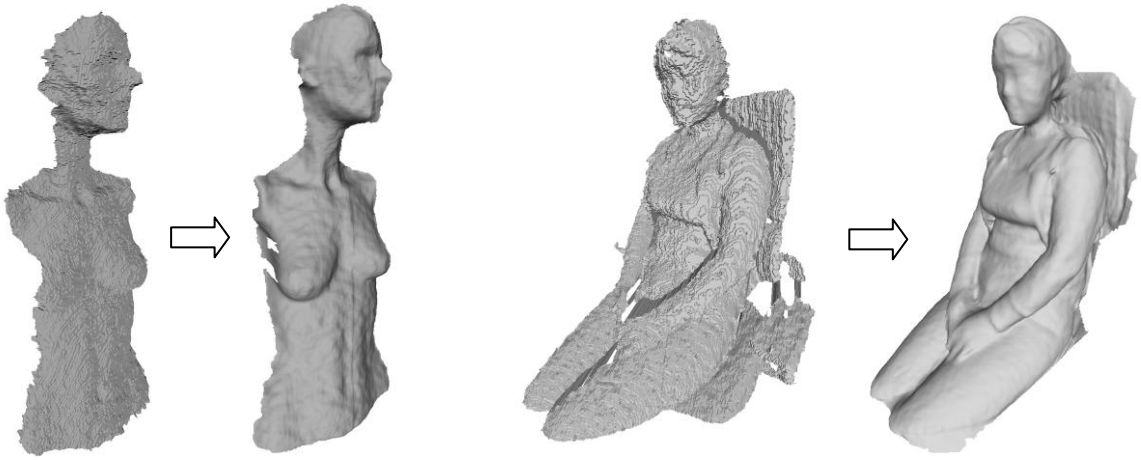


Fig3. *Raw data and super-resolution results for the Mannequin and sitting woman.*

4. Loop Closure Alignment based on the rigid and non-rigid transformation

In the scan alignment step, we don't use the 3×3 matrix R and 3×1 T, which parameterizes 12 degree of freedom. Instead we use the exponential map in conformal geometric algebra [8][9]. The idea is to get linear equations with respect to the generators of the motor. The exponential representation of motors is used, and the Taylor series expansion of first order is applied for approximation. This leads to a mapping of the above mentioned global motion transformation to a twist representation, which allows for incremental changes of pose. The Euclidean transformation of a point X in conformal space caused by the motor M is approximated as:

$$\begin{aligned} M\underline{X}\tilde{M} &= \exp\left(-\frac{\theta}{2}\left(l'+e_\infty m'\right)\right)\underline{X}\exp\left(\frac{\theta}{2}\left(l'+e_\infty m'\right)\right) \\ &\approx \left(1-\frac{\theta}{2}\left(l'+e_\infty m'\right)\right)\underline{X}\left(1+\frac{\theta}{2}\left(l'+e_\infty m'\right)\right) \\ &\approx E+e_\infty\left(x-\theta\left(l'\cdot x\right)-\theta m'\right) \end{aligned} \qquad (5)$$

Setting: $l \coloneqq \theta l'$ and $m \coloneqq \theta m'$, results in:

$$M\underline{X}\tilde{M} = E + e(x - l\cdot x - m) \qquad (6)$$

There are two parameters l and m, i.e. 6 degree of freedom to translate the 3D point. From the solution of the system of equations, the motion parameters R, T can easily be recovered.

Next, we will present the loop closure alignment problem, based on the probabilistic simultaneous scan alignment.

There are K high-resolution 3D point cloud frames captured for one model object. For each corresponding frame f, $Y_f = \{y_{f,j} \mid j = 1,\ldots,N_f\}$ and frame g, $Y_g = \{y_{g,j} \mid g = 1,\ldots,N_g\}$ we give the new energy function sum all of the corresponding frames, which is similar the energy function in [1].

$$\begin{aligned} &E\ \ M_1,\ldots M_K, d_{1,1},\ldots d_{1,O},\ldots d_{K,O}\ \ = \\ &\sum_{f,g}\left(-\sum_{n=1}^{N_f}\log\sum_{m=1}^{N_g}\exp\left(\frac{\left\|M_f\ \ y_{f,n}+V_{f,n}d_{f,n}\ \ \tilde{M}_f - M_g\ \ y_{g,m}+V_{g,m}d_{g,m}\ \ \tilde{M}_g\right\|^2}{-2{\sigma_{f,g}}^2}\right)\right) \\ &+\lambda\sum_{c=1}^{K}\left\|d_{c,1}-d_{c,2},\ldots,d_{c,O}-d_{c,O}\right\|^2 \end{aligned} \qquad (7)$$

In this equation, d is the distribution of non-rigid transformation distance; d is different for each frame; $y_{f,n}$ stands for the 3D point position in frame f; M_f stands for the transformation for the frame f. Experimentally, we could verify that a simultaneous optimization of all alignment parameters often fails to converge to a suitable minimum. Instead, we propose to alternate between optimizing for $M_1,\ldots M_K$ with fixed $d_{1,1},\ldots d_{1,O},\ldots d_{K,O}$ (case I) as Sect.3.1, and optimizing for $d_{1,1},\ldots d_{1,O},\ldots d_{K,O}$ with fixed $M_1,\ldots M_K$ as Sect.3.2 below.

The complete working pipeline for aligning all 3D point clouds frames Y is given in Table 1 as pseudo-code. The optimizer terminates if there is no further improvement or the maximum number of iterations has been reached. The result is the set of rigid alignment parameters M_ℓ, $\ell = 1,\ldots,K$, as well as the systematic bias values $d_{1,1},\ldots d_{1,O},\ldots d_{K,O}$, $O = max(N_{y_1},\ldots,N_{y_K})$.

Table1. Loop closure scan alignment

- Initialize: Local rigid processing, pair-wise corresponding 3,1, get the initial position for each frame and the corresponding points for each pair.
- Repeat until no further improvement or max. iterations :
  - Rigid registration, for all pairs frame f, g (case I):
    1. σ^2 update: according to Eq. (9)
    2. E-step: Compute P for Eq. (11)
    3. M-step: Solve the transformation for each frame $M_1,\dots M_N$ by minimizing Eq.(8) as Eq.(10)
  - Non-rigid registration, for all pairs frame f, g (case II):
    1. σ^2 update: according to Eq. (14)
    2. E-step: Compute P for Eq. (16)
    3. M-step: Solve the transformation for each frame $d_{1,1},\dots d_{1,O},\dots d_{K,O}$ by minimizing Eq.(12) as Eq.(15)
- Poisson Reconstruction for the final model.

4.**1 Rigid part**

There are two steps in the rigid transformation: Local rigid part and global rigid part. The local rigid transformation part is calculated with the ICP algorithm to find an initial corresponding frame. In this step, registration is done pair-wise [7].

- Step 1: Calculate ICP for one frame to the other frames.
- Step2: Justification for labeling matches 'correct' or 'incorrect'.

From the local rigid processing, we can get the pair-wise registration, (e.g. for frame f , there are w frames $g_1,\dots g_w$, corresponding to frame f), the corresponding points for each pair frame and the initial position for each frame. The results are shown in Fig.4 (brown color object). The object is not closed because there is a strong random noise of raw data for each frame. We need the global rigid processing and the non-rigid processing to get a correctly closed 360° model. For the global rigid part, optimizing for $M_1,\dots M_K$ with fixed $d_{1,1},\dots d_{1,O},\dots d_{K,O}$ (case I) in Eq. (7), the energy function for the global rigid part is:

$$E_{rigid}\ \ M_1,\dots M_K\ \ = \sum_{f,g}\left(-\sum_{n=1}^{N_f}\log\sum_{m=1}^{N_g}\exp\left(\frac{\left\| M_f y_{f,n}\tilde{M}_f - M_g y_{g,m}\tilde{M}_g \right\|^2}{-2\sigma_{f,g}}\right)\right) \quad (8)$$

For each corresponding pair frame (get it from the local rigid part, frame f, g are the corresponding frames), $y_{f,n}$ is the 3D point in frame f . $y_{g,m}$ is the 3D point in frame g . The variance $\sigma_{f,g}$ of the mixture components is estimated using:

$$\sigma_{f,g}{}^2 = \frac{1}{N_f N_g}\sum_{n=1}^{N_f}\sum_{m=1}^{N_g}\left\| M_f y_{f,n}\tilde{M}_f - M_g y_{g,m}\tilde{M}_g \right\|^2 \quad (9)$$

Using the EM algorithm to minimize the Energy function to get the exponential transform for each frame $M_1,\dots M_K$:

$$Q_{rigid}\ \ M_1,\dots M_K\ \ = \sum_{f,g}\left(\sum_{n=1}^{N_f}\sum_{m=1}^{N_g} P_{f,g}^{old}(m \mid y_{f,n})\frac{\left\| M_f y_{f,n}\tilde{M}_f - M_g y_{g,m}\tilde{M}_g \right\|^2}{2\sigma_{f,g}{}^2}\right) \quad (10)$$

$$P_{f,g}^{old}(m \mid y_{f,n}) = \frac{\exp\left(\frac{\left\| M_f y_{f,n} \tilde{M}_f - M_g y_{g,m} \tilde{M}_g \right\|^2}{-2\sigma_{f,g}{}^2}\right)}{\sum_{k=1}^{N_g} \exp\left(\frac{\left\| M_f y_{f,n} \tilde{M}_f - M_g y_{g,k} \tilde{M}_g \right\|^2}{-2\sigma_{f,g}{}^2}\right)} \qquad (11)$$

Two issues which should be considered carefully: 1) Because we use the first order of Taylor series expansion to express the M_f and M_g, we need three iterations to get the final motor M. 2) N_f is the point number of the frame f, N_g is the point number of the frame g. In order to speed-up the algorithm, we can just select Z nearest points in frame g corresponding one point in frame f; in the real experiments, we define $N_g = Z = 20$ in Eq. (10) and Eq. (11).

The results are shown in the Fig.4, the blue color object. The point cloud is correctly closed. Matching error from frame to frame remains because the processing is done assuming rigid registration. We need to use the non-rigid transformation to correct the error between frames.

4.2 Non-rigid part

We know the rigid transformation $M_1,\ldots M_K$ from the global rigid part. Next we need to estimate the non-rigid transformation part $d_{1,1},\ldots d_{1,O},\ldots d_{N,O}$ for each frame. The Energy function for the non-rigid part is:

$$E_{non-rigid}\ \ d_{1,1},\ldots d_{1,O},\ldots d_{K,O} = \sum_{f,g}\left(-\sum_{n=1}^{N_f}\log\sum_{m=1}^{N_g}\exp\left(\frac{y'_{f,n} + V'_{f,n}\, d_{f,n} - y'_{g,m} - V'_{g,m}\, d_{g,m}}{-2\sigma_{f,g}^2}\right)\right) + \lambda\sum_{c=1}^{K}\left\| d_{c,1} - d_{c,2},\ldots, d_{c,O} - d_{c,O}\right\|^2 \qquad (12)$$

Here, $y'_{f,n}$ and $V'_{f,n}$ are translated by the M_f, $y'_{g,m}$ and $V'_{g,m}$ is translated by the M_g calculated in Sec. 3.1.

$$\begin{aligned} y'_{f,n} &= M_f y_{f,n} \tilde{M}_f & V'_{f,n} &= M_f V_{f,n} \tilde{M}_f \\ y'_{g,m} &= M_g y_{g,m} \tilde{M}_g & V'_{g,m} &= M_g V_{g,m} \tilde{M}_g \end{aligned} \qquad (13)$$

The variance $\sigma_{f,g}$ of the mixture components is estimated using:

$$\sigma_{f,g}{}^2 = \frac{1}{N_f N_g}\sum_{n=1}^{N_f}\sum_{m=1}^{N_g}\left\| y'_{f,n} + V'_{f,n}\, d_{f,n} - y'_{g,m} - V'_{g,m}\, d_{g,m}\right\|^2 \qquad (14)$$

Minimize the energy function with the EM algorithm:

$$Q_{non-rigid}\ \ d_{1,1}\ldots d_{N,O} = \sum_{f,g}\left(\sum_{n=1}^{N_f}\sum_{m=1}^{N_g} P_{f,g}^{old}(m \mid y'_{f,n})\frac{\left\| y'_{f,n} + V'_{f,n}\, d_{f,n} - y'_{g,m} - V'_{g,m}\, d_{g,m}\right\|^2}{2\sigma_{f,g}{}^2}\right) \qquad (15)$$

$$P_{f,g}^{old}(m \mid y'_{f,n}) = \frac{\exp\left(\frac{\left\| y'_{f,n} + V'_{f,n}\, d_{f,n} - y'_{g,m} - V'_{g,m}\, d_{g,m}\right\|^2}{2\sigma_{f,g}{}^2}\right)}{\sum_{k=1}^{N_g}\exp\left(\frac{\left\| y'_{f,n} + V'_{f,n}\, d_{f,n} - y'_{g,k} - V'_{g,k}\, d_{g,k}\right\|^2}{2\sigma_{f,g}{}^2}\right)} \qquad (16)$$

Please note that for efficiency reasons, we don't evaluate $Q_{non-rigid}$ for all 3D points, but only for a subset of samples from frame f and the frame g. We can just select Z nearest points in frame g corresponding one point in frame f, in the real experiments, we define $N_g = Z = 100$ in Eq. (15) and Eq. (16). Also, for camera paths covering a larger viewpoint range, we perform several global alignments to several reference scans, such that sufficient overlap is guaranteed.

The results of non-rigid registration are shown in Figure 5, Mannequin and the woman sitting, left is without non-rigid registration, right is with non-rigid registration. Our algorithm can remove most of the outliers and noises by drafting the 3d points frame by frame. We have implemented the rigid and non-rigid algorithm in C++ - Table 2 reports runtimes.

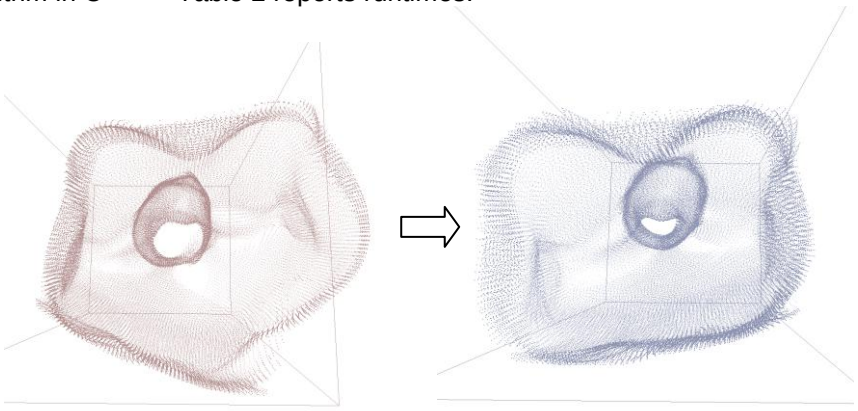
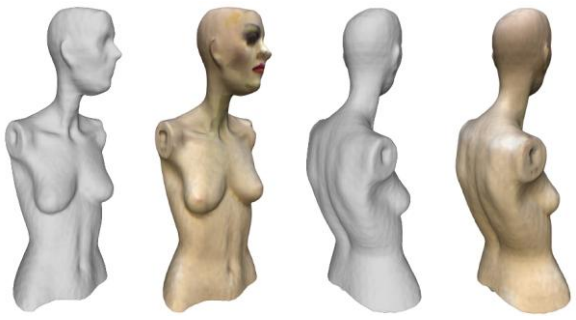


Fig4.*Result of the local pair wise alignment and global loop closure alignment.*

Fig5*. Without non-rigid and with non-rigid results*

5. Result

The human body is turned in front of a static Kinect continuously about 30 sec. For each body, we capture about 1200 frames (For each second, there are 30-50 frames captured by Kinect). Select 360 frames as the useful original data, get one super-resolution frame for each 10 raw frames, then align with global and non-rigid approach for these 36 super-resolution frames, calculate mesh with Poisson method last. The final results are complete model with smooth surfaces, and still a lot of detail structures, e.g. the fold of clothes, the hair, the face and so on, as figure 6 shows. The whole processing time for each human body is about 5 min under C++, the PC setting is Intel(R) CPU 2.67GHz, 12GB RAM memory, Windows 7 system. The running times for each object are shown in the Table 1.



(a)Mannequin: Height 79cm

(b)Snowman: Height 82cm

(c)Standing woman: Height 160cm

(d)Sitting woman: Height 78cm

Fig 6. Kinect 3D reconstruction results

Table 2. Running time for each processing ste.

	Key Frames	Super-resolution	Rigid Alignment	Non-rigid Alignment	Poisson Recon	All
Mannequin	36	32sec	132sec	72sec	79sec	315sec
Snowman	35	31sec	121sec	61sec	68sec	281sec
Human standing	30	28sec	118sec	66sec	66sec	278sec
Human sitting	32	29sec	119sec	71sec	77sec	296sec

6. Conclusion

In this paper we demonstrated that 3D shape models of static objects can also be acquired with a Kinect sensor that, at first glance, seems completely inappropriate for the task. The key in making this possible is the effective combination of 3D super-resolution with a new probabilistic multi-scan alignment algorithm tailored to Kinect cameras. In future, we plan to investigate approaches for real-time shape scanning, as well as incorporation of more sophisticated noise models into the reconstruction framework.

References

1. Y. Cui, S. Schuon, C. Derek, S. Thrun, and C. Theobalt, 3d shape scanning with a time-of-flight camera. In Proc. of IEEE CVPR 2010, 2010.
2. S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. Proc. CVPR, 2009.
3. R. Benjemaa and F. Schmitt. A solution for the registration of multiple 3d point sets using unit quaternions. . In Proc. ECCV '98 II, pages 34–50, 1998
4. R. Bergevin, M. Soucy, H. Gagnon, and D. Laurendeau. Towards general multi-view registration technique. IEEE PAMI, 18(5):540–547, 1996
5. P. J. Besl and N. D. McKay. A method for registration of 3-d shapes. IEEE PAMI., 14(2):239–256, 1992
6. J.-Y. Bouguet and P. Perona. 3d photography on your desk. In Proc. ICCV, page 43. IEEE, 1998
7. D. F. Huber. Automatic three-dimensional modeling from reality. Technical report, 2002
8. Y. Cui and D. Hildenbrand. Pose estimation based on geometric algebra. 2009
9. B. Rosenhahn. Pose estimation revisited. PhD thesis, Universitat Kiel, September 2003.
10. B. Curless and M. Levoy. A volumetric method for building complex models from range images. In Proc. SIGGRAPH, pages 303–312. ACM, 1996
11. Y. J. Kil, B. Medereos, and N. Amenta. Laser scanner superresolution. In Point-based Graphics, 2006
12. Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In CVPR '07, 2007
13. Q.-X. Huang, S. Fl¨ory, N. Gelfand, M. Hofer, and H. Pottmann. Reassembling fractured objects by geometric matching. ACM TOG., 25(3):569–578, 2006
14. D. Lanman and G. Taubin. Build your own 3d scanner: 3d photograhy for beginners. In SIGGRAPH courses, pages 1– 87. ACM, 2009
15. Myronenko, X. Song, and M. Carreira-Perpinan. Non-rigid point set registration: Coherent Point Drift. NIPS, 19:1009, 2007
16. S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3d model acquisition. In Proc. SIGGRAPH, pages 438–446. ACM, 2002.
17. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV, 47(1):7–42, 2002
18. KINECT FOR XBOX 360, C. 2010. http://en.wikipedia.org/wiki/kinect.