# KINECT-BASED REAL-TIME RGB-D IMAGE FUSION METHOD

Wei Guo [a*], Tangwu Du [a], Xinyan Zhu [a], Tao Hu[a]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University
129 Luoyu Road, Wuhan, Hubei, 430079, China– tangwudu@gmail.com

**KEY WORDS:** Vision、Kinect、RGB-D、Modelling、Fusion、Indoor Environment、Matching

## ABSTRACT:

3D reconstruction of indoor environments based on vision has been developed vigorously. However, the algorithm’s complexity and requirements of professional knowledge make it restricted in practical application. With the proposition of the concept of Volunteered Geographic Information (VGI), the traditional method is no longer suitable for VGI. So in this work we utilize consumer depth cameras - *Kinect* to enable non-expert users to reconstruct 3D model of indoor environment with RGB-D data. Considering the possibility of camera tracking failure we propose a method to perform automatic relocalization.

## 1. INTRODUCTION

So far, to reconstruct 3D model with fine geometric accuracy and rich visual information is the goal which is pursued unremittingly by researchers. Relative technologies have been developed vigorously by some communities including Photogrammetry, Computer Vision and Robotics. The applications of research results in the Digital City, Robot Navigation and other fields have demonstrated wide application prospect. However, due to the high complexity of real scene, to achieve this goal still seems far beyond our reach. In practice geometric fine 3D indoor models are still rarely used. On the other hand, with the development of Digital Earth technology, urban 3D modelling technology has been mature to some extent. But indoor models of large public places have not entered public’s view yet. However, because of the importance of indoor navigation, it is imperative for 3D reconstruction to shift from outdoor to indoor context. Now there have been some kinds of indoor 3D reconstruction methods. But these techniques are either expensive equipment relied or technically complex which makes it unsuitable for consumers. With the proposition and development of Volunteered Geographic Information (VGI), public’s involvement in geographic information creation, editing, management and maintenance has become an important trend. Therefore it becomes increasingly important for consumers to quickly and accurately obtain 3D information of the scene and rebuild its 3D model.

RGB-D camera is a new type of sensing device which can capture RGB images along with per-pixel depth information. RGB-D cameras rely on either active stereo or time-of-flight sensing to generate depth estimates at a large number of pixels. There has been a variety of depth sensing devices in which Kinect is a typical representative. Kinect is developed mainly to recognize human gesture. Its low cost has made it widely used in the field of game. In consideration of the ability of Kinect to quickly access RGB-D data in real-time, we believe that it is valuable to be used in the field of 3D reconstruction and related areas. Though there have been lots of research in 3D reconstruction using depth image, existing algorithms does not fully explore the potential Kinect offers.

## 2. RELATED WORK

Recently, image-based dense surface reconstruction has produced many compelling results. Microsoft's Photosynth and Photo Tourism and the University of Washington’s Bundler outperform among this kind of technology. But they are propitious to city-scale or outdoor environments. And it is extremely hard to extract dense depth information from color camera data alone, especially in indoor environments with very dark and sparsely textured areas. Besides, the complexity of indoor environments makes it very hard for data acquisition. In contrast, LiDAR is not sensitive to scene illumination and 3D point clouds are extremely well suited for frame-to-frame alignment and dense 3D reconstruction. But this equipment is very expensive and lack of texture information. Kinect is such a device generating color information and depth information simultaneously with a speed of 30 frames per second. However, Kinect provides depth only up to a limited distance (typically less than 5m) with 640 × 480 image size. Also, the depth data are very noisy and Kinect’s field of view (~60$^{o}$) with less depth precision (~3cm at 3m depth). So it is far more constrained than LiDAR. With the movement of camera there will be large areas of overlap between two adjacent frames. Therefore complete 3D model can be reconstructed using depth image registration. Nowadays there are two main types of 3D reconstruction from depth images, one is patch-based 3D reconstruction and the other is voxel-based.

The core idea behind patch-based method is data alignment. Different from the image-based approach data alignment rely on not visual features extraction but distance metric using depth data. To register two depth images Iterative Closest Point (ICP) algorithm (Chen, 91) (Besl, 92) is the most widely used technique. To speed up the convergence rate there have been many ICP variants proposed. When surface normal measurements are available point-to-plane (Chen, 92) (Rusinkiewicz, 2001) metric has been shown the most effective. In ICP to obtain the closest point correspondences is expensive. So in (Blais, 95) projective data association algorithm was proposed to drastic speed up this process. There is also extensive literature within the AR and robotics community on Simultaneous Localization and Mapping (SLAM). RGB-D SLAM (Henry, 2010), a project aiming to map large indoor environment, uses patch-based 3D reconstruction. It effectively utilizes visual and shape information from RGB-D camera to reconstruct 3D model. In the paper SIFT features are used to provide initial point pairs for ICP algorithm. The objective of RGB-D SLAM is not only alignment and registration but also building 3D models with both shape and appearance information. Using patch-based method can reconstruct a larger 3D model. But the speed of reconstruction is real-time but non-interactive. In addition due to the instability of human-computer interaction, too much noise in the data along with poor accuracy the system is not robust and difficult to eliminate “ghost image”. And as the scene continues to be explored, errors in alignment between a particular pair of frames, and noise and quantization in depth values, cause the estimation of camera position to drift over time.

As the patch-based approach has many shortcomings, researchers proposed voxel-based data fusion to reconstruct 3D model (Curless, 1996). Different from the data alignment the core idea behind it is data fusion. Compared to patch-based method the geometric accuracy of reconstructed 3D model is higher. In voxel-based data fusion a predefined 3D volume with fixed resolution maps to a 3D physical space. The volume is subdivided uniformly into a 3D grid of voxels. Each voxel stores a weighted average of its distance to the assumed position of a physical surface. Microsoft KinectFusion (Izadi, 2011) (Newcombe, 2011) apply voxel-based RGB-D data fusion to reconstruct indoor 3D model, which take full advantage of modern GPU's acceleration capacity. Their work only uses ICP algorithm to track camera position so that visual feature extraction and matching are no longer needed. KinectFusion can achieve real-time interactive rates for both camera tracking and 3D reconstruction. But voxel-based method is too much memory consuming leading to its limited application in large-scale scenarios. At the same time because that depth data are the only information used to track camera position when camera tracking fails it is difficult to perform automatic relocalization.

In summary, the patch-based technology can reconstruct larger 3D model, but the accuracy of the model is very not satisfying. Voxel-based reconstruction can generate a 3D model with high accuracy but valuable information contained in RGB images is ignored. And camera tracking may fail while the camera moves too fast or the scene contains too much flat areas. Under this circumstance KinectFusion has to rely on human-computer interaction to relocate camera position and orientation which is not efficient. In this paper, in guarantee that the 3D model of the scene has been finely rebuilt, we will try to solve the problem of automatic relocalization when the tracking has failed. The main idea behind this paper is to organize RGB-D data which has been fused into the 3D model in graph structure. To reduce the data volume stored in the graph sample frame based on geometric constraints is defined. When camera tracking failure has happened, RGB-D data alignment based on SIFT and ICP along with color similarity measurement is introduced to re-initialize camera pose.

## 3. SYSTEM OVERVIEW

To completely reconstruct a 3D model with high geometric fidelity in this paper we will employ voxel-based reconstruction method. This part has been described concretely in (Izadi, 2011) (Newcombe, 2001). At the same time to enhance the stability of the system we will emphatically resolve the problem of automatic relocalization when the tracking has failed. On the whole our system is comprised of the following components. First, a pre-processing stage, the live depth image acquired by Kinect is converted from image coordinates into 3D points and normals in the coordinate space of the camera. Then a rigid 6DOF transform is computed to closely align the current point cloud with the previous frame, using point-to-plane ICP variant. In order to improve the registration efficiency a GPU-base Iterative Closest Point (ICP) algorithm is implemented. Relative transforms are incrementally applied to a single transform that defines the global pose of Kinect. Then voxel-based data fusion is applied to incrementally reconstruct 3D model of the scene. However in practice due to environmental impact there may be some mismatch or the accuracy is not satisfying. In this paper we use colour similarity measurement to evaluate the accuracy of registering results. If the evaluation results within a pre-defined threshold the point clouds can be fused to the 3D model and the frame is possible to be added into a graph structure. The graph structure is constructed to solve the problem of camera relocalization. When the camera tracking has failed it may be caused by too fast moving or the quality of the data is poor all of which will lead the ICP algorithm converge into local minimum. As such, combining with the generated map we have to re-initialize camera position. In camera relocalization the first step is to make use of geometric relationship in the map to reduce date volume. Then feature matching is used to find 2 to 3 sample frames which best match current frame. Finally color similarity measurement is reused to evaluate the results of data registration. And one frame which scores the highest will be chosen to relocate the camera position. The primary process is shown as below,

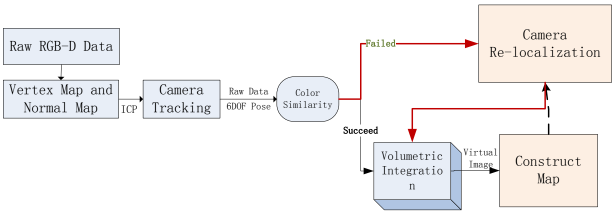

Figure 1: the primary components of our system

## 4. METHODOLOGY

Volumetric surface representation based on (Curless, 96) is the main method we employed to reconstruct 3D model from RGB-D data. In this process a huge mass of mutually independent RGB-D data are acquired. When camera tracking has failed automatic and precise relocalization is very important to maintain the consistency of reconstructed model. In relocalization the core primary mission is to find a sample frame best matches to current frame. Obviously, if every frame is matched one by one in brute force it will be time consuming. So it is necessary to efficiently organize those data. In this paper the overall strategy is representing constraints between frames with a graph structure. In the graph each node represents a sample frame meanwhile each edge represents geometric constraints between two frames. The information this graph maintains includes Spatial Index, Feature Index and Virtual Camera.

### 4.1 Graph Construction

First we initialize an empty graph. In 3D reconstruction only relative transformation between two consecutive frames is computed. Therefore the graph structure is linear. And only if matched data are similar in colors the new RGB-D can be fused into the 3D model. After that we should update the graph. In this process considering huge amount of RGB-D data are acquired and the motion from one frame to the next is very small, so we just take some sample to update the graph. The sample frame is determined by geometric relationship.

### 4.1.1 Spatial Index

After ICP algorithm used to obtain relative transformation ti between two frames the transformation should be applied to a global coordinate so that the data can be fused into a 3D model. Correspondingly this transformation is denoted as $T_i^{'}$ . Using $T_i^{'}$ camera's corresponding location (X, Y, Z) is calculated. As mentioned before, before adding data into the graph we calculate the distance between current camera position and the closest sample frame's position (X', Y', Z'). If the distance does not exceed a pre-defined threshold, then there is no need to update the map. However as the camera continues to move this distance will eventually exceed the threshold. Then a new sample frame is added. And Spatial Index will record the 3D location.

#### 4.1.2 Feature Index

ICP algorithm is critical to the selection of initial point pairs. Otherwise the algorithm can easily fall into local minimum. When camera tracking fails, the assumption that high degree of similarity between two frames exists will no longer tenable. So it is obligatory to provide initial point pairs for ICP algorithm. Note that it is difficult to extract effective feature information from depth image. In practical applications visual features are more reliable and visual features can measure the degree of similarity between two frames. So in SLAM visual features are applied to detect loop closure. For the establishment of Feature Index, RGB image is used to extract SIFT features. In practice it has been proven that SIFT feature is invariant, even for images with scale change and rotation.

#### 4.1.3 Virtual Camera

Raw depth image is not accurate and contains lots of noise. The reason that voxel-based data fusion can generate 3D model with high geometric fidelity is that at same area RGB-D data are collected from multiple angles. So the generated 3D model is a weighted average of those data. To improve the accuracy of registration, virtual depth image is generated combining the 3D model with current camera position $T_i$. Compared to raw depth it has a higher geometric precision.

As we have a dense surface reconstruction and camera's global position Ti, per pixel ray-cast can be performed (Parker, 98). Each pixel's corresponding ray is marched starting from the minimum depth for the pixel and stopping when a zero crossing is found indicating the surface interface. Then the distances corresponding to its pixel position is recorded which is used to generate a virtual depth image.

### 4.2 Sample Frame Extraction

There are two main components in the process of camera relocalization. First, a process of data collection, a set of sample frames is extracted from the map. Second, a sample frame which is the best match of current frame is determined from the data set. Note that when the graph is constructed geometric constraints between consecutive frames are recorded. And there is only small motion of the camera from frame to frame. So we can utilize the geometric relationship to reduce the number of data to be used soon afterwards. As we know that visual information is often to be used to evaluate the degree of similarity between frames. So in this paper we extract SIFT feature from RGB image and match them between current frame and sample frames. Besides, SIFT pairs can provide initial point pairs for ICP algorithm. SIFT is widely used feature detector and descriptor (Lowe, 2004). Though the descriptors are very distinctive, they must be matched heuristically and there can be false matches. To determine a subset of feature pairs corresponding to a consistent rigid transformation RANSAC algorithm is used. Additionally, the RANSAC associations act as an initialization for ICP, which is a local optimizer. For the 3D point clouds we employ a point-to-plane ICP algorithm to compute rigid 6DOF transformation. To evaluate the accuracy of those matches color similarity measurement is conducted. Eventually the best match is selected to recover current camera position. The overall process of camera relocalization is as follows in listing 1,

**Listing 1** Camera Re-localization

```
Nodes = Find_Sub_Data_Set_From_Graph(distance)
F* = Extract_RGB_SIFT_Features(P_s)
If {F} = Find_Similar_Samples(Feature_Index; F*)
   For each sample in {F}
      t = Perform_RANSAC_Alignment(F_j; F*)
      Repeat
         T′= Compute_Closest_Points(t; P_s; P_t)
      until ( Error Converge(t) ≤ θ ) or (max Iteration reached)
      return T′
S_i = Computer_Color_Similarity_Measurement(P_s; P_t; T′)
T = Get_Max_Similarity(S_i)
Relocate_Camera_Position_(T)
```

Consecutive depth frame, with an associated live camera pose estimate, is fused incrementally into one single 3D reconstruction. After a period of time a huge amount of RGB-D data are accumulated and maintained by the graph structure. Note that camera moves at a certain rate. So even if camera tracking has failed, current position of the camera must locate around the last sample frame within a certain range. Consequently if the last sample frame in the graph is picked as data centre and a radius is pre-defined we can employ Spatial Index to extract a data set {F}. For simplicity, sample frames are noted as $\mathbf{F_j}$ and current frame $\mathbf{F}^*$. $\mathbf{F_j}$ which is the best match of $\mathbf{F}^*$ will lie in the data set. However, as $\mathbf{F_j}$ varied from each other we have to measure the degree of similarity between $\mathbf{F_j}$ and $\mathbf{F}^*$. Firstly we extract sparse visual features from $\mathbf{F}^*$ and associate them with their corresponding depth values to generate feature points in 3D which will be used later. Then those features are matched heuristically with features kept by Feature Index of each node in the data set. This part will be elaborated in section 1.1.2. Then the number of successfully matched feature pairs is an indication of the degree of data similarity. And we will choose 2~3 sample frames from the data set which rank the top in this procedure.

### 4.3 Selection Strategy

The ICP algorithm iterates between associating each point in one time frame to the closest point in the other frame and computing the rigid transformation that minimizes distance between the point pairs. However the important first step of ICP is to find correspondences between frame pairs, otherwise the ICP algorithm will easily converge to a local minimum. So we use feature selection to provide initial corresponding point pairs for ICP algorithm. Through the above procedures 2~3 sample frames are selected. To determine which frame should be used to recover camera pose color similarity is introduced. Combining the color similarity criterion, registering is much more robust in difficult cases and the result becomes more reliable.

#### 4.3.1 SIFT+RANSAC

In the process of SIFT match the best candidate match for each keypoint of $\mathbf{F}^*$ is found by identifying its nearest neighbor in the database of keypoints of $\mathbf{F_j}$. Consider N pairs of initial feature pairs between frame $\mathbf{F}^*$ and $\mathbf{F_j}$, represented by vectors **(X; Y)** in their respective coordinate system. RANSAC samples the solution space of **(R; T)** (rotation and translation) and counts the number of inliers, **f**,

$$f(R,T)=\sum_{i}^{N} I(X^i,Y^i,R,T) \qquad (1)$$

Where I will be inlier if $\mathbf{X^i}$ and $\mathbf{Y^i}$ fit well with a pre-defined threshold under the constraint of **(R; T)**. A inlier will be counted as 1 otherwise 0. RANSAC chooses the transform consistent with the largest number of inlier matches.

#### 4.3.2 ICP

In 2D because of the scale indeterminacy the frame pairs are not finely aligned. That means the registration accuracy is not precise enough. ICP is a popular and well-studied algorithm for

3D shape alignment. And ICP has been shown to be effective when two point clouds are nearly aligned. Since we have two frames aligned in the process of SIFT+RANSAC the prerequisites for ICP algorithm has been satisfied. To generate more accurate alignments than point-to-point ICP an ICP variant based on point-to-plane error metric has been shown to improve convergence rates and is the preferred algorithm when surface normal measurements are available (Rusinkiewicz, 2002) (Segal, 2009). In the previous section we have mentioned that virtual depth image is more accurate and less noisy compared to raw depth image. So in this part the point-to-plane ICP will be applied between virtual depth image and current RGB-D frame. In the first iteration of ICP algorithm **(R; T)** is initialize by SIFT+RANSAC match. When the point-to-plane error metric is used, the object of minimization is the sum of the squared distance between each source point and the tangent plane at its corresponding destination point. More specifically, if **$s_i$ = ($s_{ix}$, $s_{iy}$, $s_{iz}$, 1)$^T$** is a source point, **d = ($d_{ix}$, $d_{iy}$, $d_{iz}$, 1)$^T$** is the corresponding destination point, and **n = ($n_{ix}$, $n_{iy}$, $n_{iz}$, 1)$^T$** is the unit normal vector at then the goal of each ICP iteration is to find **($R_{opt}$; $T_{opt}$)** such that (Low, 2004)

$$(R_{opt};T_{opt}) = \arg\min_m \sum_i (((R;T)\bullet s_i - d_i)\bullet n_i)^2 \qquad (2)$$

After the registration of 3D point clouds the final transformation **($R^*$; $T^*$)** is computed.

### 4.3.3 Color Similarity Measurement

The framework above only utilizes little part of pixels corresponding to SIFT feature in the depth image. It is assumed that if **($R^*$; $T^*$)** applied on frame pairs common areas should overlap perfectly. However the rigid transformation may be unreliable under difficult circumstances. So it is not always the case in practical situations. To compute color similarity, we choose a set of points from RGB image including all SIFT feature points and some other visual features such as Harris Descriptor. SIFT features often locate at the edge of object while point clouds are not sensitive in these areas. So we put larger weight on those pixels corresponding to SIFT features in color similarity measurement.

Every feature point has information including location, gradient magnitude and orientation. For each image sample, L(x, y), the gradient magnitude, m(x, y), is precomputed using pixel differences to produce weight W(x, y):

$$m(x,y) = \sqrt{(I(x+1,y)-I(x-1,y))^2 + (I(x,y+1)-I(x,y-1))^2} \qquad (5)$$

$$W(x,y) = 1/m(x,y) \qquad (6)$$

To measure color similarity coefficient method is used. First we set **$F^*$** as master image and **$S^j$** as slave image and get pixels corresponding to SIFT features from both RGB data. The difference is that pixel window in **$F^*$** is 4*4 and $F^j$ larger 16*16. The coefficient of the stereo-pair pixel of matching window and the target window can be calculated by formula below. This final coefficient r is the max value of the window.

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{[\sum_{i=1}^{n}(X_i-\bar{X})^2][\sum_{i=1}^{n}(Y_i-\bar{Y})^2]}} \qquad (7)$$

Combining with the pre-defined weight we sum all of those coefficients up. Each sample frame obtained in section 1.1.2 there will be a coefficient. So for the reason to compare color similarity S we have to normalize the coefficient value:

$$S = \frac{\sum_{i=1}^{m} W(x_i,y_i)\bullet r(x_i,y_i)}{\sum_{i=1}^{m} W(x_i,y_i)} \qquad (8)$$

## 5. RESULTS & DISCUSSION

We have conducted a number of experiments to investigate the performance of our system. These and other aspects, such as the system's ability to keep track during very rapid motion and the performance of automatic relocalization, are tested. In our experiment an indoor space is reconstructed. Figure 1 shows an example frame observed with this RGB-D camera.

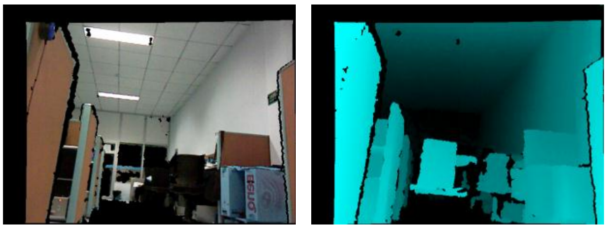

Figure 2: (left) RGB image and (right) depth information captured by an RGB-D camera. Black pixels in the right image have no depth value, mostly due to max distance, or surface material.

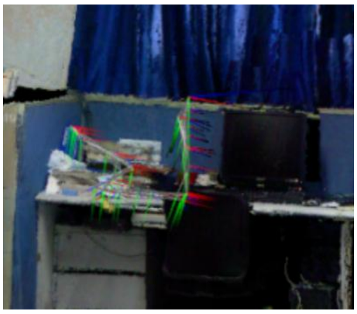

Figure 3: Demonstration of the reconstructed 3D model. The colored points in the middle linked as a polygonal line are representatives of sample frame. They are also the representatives of camera positions.

During mapping, the camera was carried by a person, meanwhile to test the performance of automatic relocalization the camera was moved shiftily. As shown in Figure 3 that there is no explicit "holes" or "ghost image" existing in the reconstructed model. Some holes on the edge of object are caused by the missing of data information where camera cannot reach. In our experiment camera tracking failure happened. However the system only takes some milliseconds to re-initialize camera position. So the efficiency of our method to achieve camera relocalization has been proven.

## 6. CONCLUSION

Building accurate, dense models of indoor environments has many applications in robotics, gaming. In this paper We investigate how potentially inexpensive depth cameras-Kinect-can be utilized to reconstruct 3D model using voxel-based method. To maintain the stability of our system graph-based method along with SIFT and Colour Similarity Measurement has been proposed. And we get a prospective result of camera relocalization in 3D reconstruction process.

## REFERENCES

D. Lowe. 2004. Discriminative Image Features from Scale-invariant Keypoints. International Journal of Computer Vision, 60(2).

A. Segal, D. Haehnel, and S. Thrun. 2009. Generalized-ICP. In Proc. of Robotics: Science and Systems (RSS).

S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. Real-time 3D model acquisition. In ACM Transactions on Graphics (SIGGRAPH), 2002.

Blais, G., Levine, M. 1995. Registering Multiview Range Data to Create 3D Computer Objects, Trans. PAMI, Vol. 17, No. 8.

Besl, P., McKay, N. 1992. A Method for Registration of 3-D Shapes, Trans. PAMI, Vol. 14, No. 2.

Y. Chen, G. Medioni. 1992. Object modeling by registration of multiple range images. Image and Vision Computing (IVC), 10(3):145–155.

B. Curless, M. Levoy. 1996. A volumetric method for building com-plex models from range images. In ACM Transactions on Graphics(SIGGRAPH).

Shahram Izadi et al. 2011. KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera, SIGGRAPH,

Richard A. Newcombe et al. 2011. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In ISMAR.

P. Henry et al. 2010. RGB-D mapping: Using depth cam-eras for dense 3D modeling of indoor environments. In Proc. of the Int. Symposium on Experimental Robotics (ISER).

R. A. Newcombe and A. J. Davison. Live dense reconstruction with a single moving camera. In Proc. of the IEEE CVPR,2010.

R. A. Newcombe, S. Lovegrove, and A. J. Davison. 2011. DTAM: Dense tracking and mapping in real-time. In Proc. of the Int. Conf. on Computer Vision (ICCV).

S. Rusinkiewicz, M. Levoy. 2001. Efficient variants of the ICP algorithm. 3D Digital Imaging and Modeling, Int. Conf. on , 0:145.

S. Parker, P. Shirley, Y. Livnat, C. Hansen, and P. Sloan. 1998. Interactive ray tracing for isosurface rendering. In Proceedings of Visualization.

K. Low. 2004. Linear least-squares optimization for point-to-plane ICP surface registration. Technical report, TR04-004, University of North Carolina.

## ACKNOMLEDGEMENT

The authors wish to thank China Postdoctoral Science Foundation (180789) and the Basic Research Universities Special fund (274988), and the National 973 Program (171065).