Real-time 3-D Feature Detection and Correspondence Refinement for Indoor Environment-Mapping using RGB-D cameras

Liang-Chia Chen[1,2]
[1]Department of Mechanical Engineering
National Taiwan University
Taipei, Taiwan
Email: lchen@ntu.edu.tw

Nguyen Van Thai[2]
[2]Graduate Institute of Automation Technology
National Taipei University of Technology
Taipei, Taiwan
Email: t7669033@ntut.edu.tw

Hsien-I Lin[2]
[2]Graduate Institute of Automation Technology
National Taipei University of Technology
Taipei, Taiwan
Email: sofin@ntut.edu.tw

Abstract—**The article presents an efficient method in detecting critical 3-D feature points for efficient and accurate data registration required in real-time indoor environment mapping by using RDB-D cameras. To achieve fast and accurate data correspondence between different 3-D scanned images, in the proposed method, RGB images are first used to detect two-dimensional (2-D) sparse color features for estimating matched pairs between successive scanned depth images. Critically, detected 2-D sparse features are mapped with their corresponding depth information. Consequently, sub-sets of matched pairs in 3-D depth space are established. Moreover, due to potential sensing noises, not all of pairs are valid and useful to 3-D matched pair establishment. Invalid pairs are detected and eliminated using an proposed angle-based filter for 2-D matched pairs, as well as a filter based on Euclidean distance, neighboring area and surface curvature filters for 3-D matched pairs. The experimental results show that the method is efficient and invariant to pose, robust for large-scale indoor environments, and feasible for real-time 3-D indoor environment mapping.**

Keywords—correspondences; RGB-D information; 3-D feature detection; refinement.

I. INTRODUCTION

In previous years, most of three-dimensional (3-D) optical detection systems are based either on laser scanners or stereo cameras. Although these systems have been used widely and successfully for numerous applications in 3-D space because of their high measurement range and accuracy, there are still some undesired limits or difficulties in these solutions. The establishment of corresponding point pairs between two images to estimate the depth information is one of challenges for the stereo systems. Meanwhile, a serious restriction for laser scanners is that measurements are performed line by line. These bottlenecks may lead to negative outcomes for real-time 3-D feature detection. Recently, RGB-D cameras have received much attention for its potential application in 3-D mapping, segmentation, recognition and others applications. RGB-D cameras can provide both of color and depth information, enabling machines perceive the world in 3-D space and translate these perceptions into a synchronized depth image in the same way that human do. In this research, a RGB-D camera was integrated with a mobile robot for utilizing the advantages described above.

Detection of feature points plays an important key in enhancing efficiency and accuracy of manipulation of 3-D point clouds, such as image registration, data segmentation and object recognition. In computer vision, *features* are defined salient elements in 2-D images such as corners or sharp edges, while 3-D feature points are best representation for 3-D objects and scene. Feature detection is one of the most important tasks in many applications, such as 3-D mapping [1]-[7], segmentation [8]-[9] and recognition [10]-[12]. One of prominent research in 3-D image processing is that modeling 3-D objects or scenes. Detected sparse features between consecutive scanned images should be matched together to establish corresponding pairs in between. These matched pairs are critical to applications associated with 3-D models and object reconstruction. Harris method (Harris *et al.* 1988 [13]) was first proposed to enable explicit tracking of image features, in which the image features must be discrete and do not form a continuum like texture or edge pixels (edgels). Matching between edge images on a pixel-by-pixel basis can work for stereo by using the known epi-polar camera geometry. However, for the motion problem, where the camera motion is unknown, the aperture problem may prevent the undertaking explicit edgel matching. This could be further overcome by solving for the motion beforehand. Harris selected a corner-region pixel as a nominated corner feature if its response is an 8-way local maximum. Similarly, edge region pixels are deemed to be edgels if their responses are both negative and local minima in either the x or y direction,

according to whether the magnitude of the first gradient in the *x* or *y* direction is larger. This results in thin edges. In the 1990s, several corner and edge detectors were introduced. In 1994, Wang *et al.* [14] proposed a new corner detection algorithm based on the observation of surface curvature. The algorithm utilized a linear interpolation scheme for intermediate pixel addressing in the differentiation step, which results in improved accuracy of corner localization and reduced computational complexity. Noise was reduced by a combination of Gaussian convolution, non-maximum suppression and false corner response suppression. Three years later, Smith *et al.* [15] described a new approach for low level image processing, edge and corner detection and structure preserving noise reduction. In this work, they used non-linear filtering to define which parts of the image are closely related to each individual pixel in which each pixel is associated with a local image region having similar brightness. The new feature detectors were based on the minimization of this local image region, and the noise reduction method takes this region as the smoothing neighborhood. Another well-known corner detector was also published one year later, 1998, by Trajkovic *et al.*, called FAST corner detector [16]. The algorithm was based on the property of corners that the gradient of image intensity should be a local maximum in all directions. Consequently, the corner response function was computed as a minimum change of intensity over all possible directions. To compute the intensity change in an arbitrary direction, an inter-pixel approximation was used. A multi-grid approach was employed to reduce the computational complexity and to improve the quality of the detected corners.

In general, all of the above reviewed feature detection methods are called the feature-based algorithms. They have been widely used since corners and edges can correspond to image locations with high information content and can be matched between consecutive scanned images more reliably. However, most of feature-based algorithms may not work so well for images subjected to variations in scale, illumination variances, unknown rotation or affine transform. To solve these restrictions, in 2004, Lowe *et al.* [17] presented a method for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene, named SIFT (Scale Invariant Feature Transform). The extracted features were invariant to image scale and rotation, and were shown to provide robust matching across a substantial range of affine distortion, change in 3-D viewpoint, addition of noise, or random variances in illumination. The feature was highly distinctive, in the scene that a single feature can be correctly matched with high probability in a large image database. The major stages of computation used to generate the set of image features lie on performing scale-space extrema detection and searching for keypoint localization, orientation assignment and keypoint descriptors. Each of these stages were performed in a descending order and on every stage a filtering process was made so that only the robust key points were allowed to jump to the next stage. The cost of detecting the features was significantly reduced with SIFT. However, the 128-dimensions of the descriptor vector may turn the whole feature detection into a relatively time consuming process. For this limitation, SURF (Speed-up Robust Feature) method which was published in 2006 by Bay *et al.* [18] can ensure high speed in three of the feature detection steps: detection, description and matching. SURF is a novel scale- and rotation-invariant detector and descriptor of interest points. Like SIFT, SURF is invariant to image scaling, translation and rotation. However, SURF gains its performance because the computation and the comparison between features are faster and it is more robust against different image transformation.

In addition, in the 2000s, several 3-D feature detection algorithms for point cloud datasets have been introduced and become well-known in the 3-D vision field. The early publication in this area was PFH (point feature histogram) and FPFH (fast point feature histograms) algorithms. Both of them was developed by Rusu *et al.* in 2008 [19] and 2009 [20], respectively. The main idea of the PFH descriptor is to analyze the surrounding information of every point present in a point cloud data structure in order to obtain a geometrical representation using a "multidimensional histogram of values" based on the estimated surface normal of the point "*k*-neighborhood". Fundamentally, the steps to use this descriptor are summarized into three major steps. The first step consists of taking the closest neighbors of each point into the point cloud data model. In the second step, it is necessary to evaluate the Euclidean distance and compute the "three angular values" algorithm for each couple of neighbors. In the last step, the results of comparison are put into the output feature histogram. However, the computational complexity of the algorithm is high; only with thick clouds of data around one point; again, real-time applications cannot be established using PFH descriptor. Rusu *et al.* had improved PFH's restriction by creating FPFH descriptor. This descriptor allows reduce the computation time to make the algorithm suitable for real-time applications. Its first step consists of applying the PFH to obtain the estimations between a point and its *k*-neighborhood and so on for each point. After this step the weighting scheme is introduced by recalculating the estimations considering not only the points which are present into the *k*-neighborhood, but a wide range which allows having extra FPFH connections. However, the methods discussed above are either computationally intensive or tediously complicated. They may not be totally suitable for real-time 3-D environment mapping.

Therefore, in our work, RGB-D images are employed to develop for feature detection. A calibration procedure is first performed for the RGB-D camera so that every pixel in the colour image is incorporated with its corresponding depth value. Calibrated RGB images are then employed to detect 2-D sparse features using SURF algorithm and the detected features are used to estimate the matched pairs between successively scanned depth images. Detected 2-D sparse features are then associated with their corresponding depth information. Consequently, subsets of the matched pairs in 3-D space are established. The efficiency of the matching process can be greatly enhanced, so that real-time 3-D environment mapping is achieved.

II. METHODOLOGY

A. Data Acquisition

In this work, to acquire colored point clouds, a Kinect camera is used. Kinect camera is a RGB-D camera which can provide both of color and depth information per pixel. It acquires 640x480 registered color images and depth data at 30 frames per second. Figure **1** shows the developed mobile robot system which consists of a Kinect camera used to acquire 3-D point cloud datasets.

Figure 1. The robot system developed to acquire 3-D and color images using a Kinect camera.

By incorporating each color pixel with its corresponding depth value, colored clouds are generated. Note that pixels having corresponding depth value being equal to zero are skipped in the resulting cloud. This process is useful to eliminate invalid measured points in the resulting cloud. However, there are still many noises in the generated clouds which may distort the result of the feature detection. Thus, these noises must be rejected to guarantee the clouds contain only valid points. The generated clouds contain some hollows due to eliminated points. Figure **2** shows an example of how to generate a colored point cloud employing RGB and depth information captured from a Kinect camera.

Figure 2. The color image (left-upper) and the depth image (left-lower) acquired by Kinect. (right) The colored cloud generated by incorporating every pixel in the color image with their corresponding depth value in the depth image.

B. Generation of sparse 3-D features and refinement of invalid correspondences

RGB images captured from the Kinect camera are preprocessed by the SURF algorithm to extract sparse 2-D color feature points. The SURF detector can establish the matched pairs of the detected features between two consecutively scanned images. However, not all of pairs are valid due to potential sensing noises. In our method, an angle-based filter is developed to reject these invalid pairs. Although this filter can remove many invalid pairs. However, there may be still other invalid pairs undesirably left in the result. They will be rejected by the proposed filtering strategy . The procedure of generating sparse 3-D features and rejecting invalid correspondences is shown in Figure **3**.

According pairs retained by the angle-based filter and depth images corresponding with RGB images, an incorporating procedure is introduced to generate matched pairs in 3-D space, called correspondences. Finally, a correspondence refinement algorithm is applied to reject all invalid corresponding pairs. The developed strategy consists of three major stages, in which the first stage is distance-based filtering, following by the area-based filter as the second filtering, and the curvature-based filter as the third filter. The details of the developed strategy are described in the following sections.

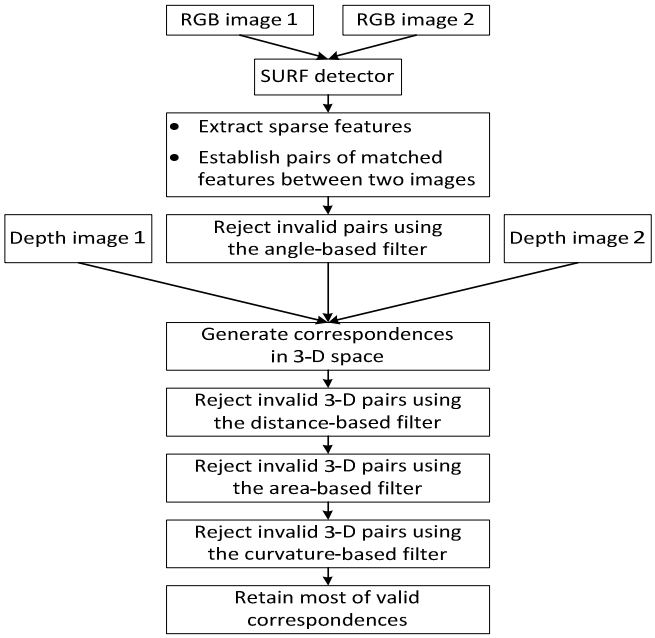


Figure 3. Flowchart demonstrating the algorithm of real-time 3-D feature detection and correspondences refinement.

1) Angle-based filter

SURF detector is a fast algorithm for detecting sparse 2-D features for color images. However, due to potential sensing noises, there exist some data pairs having poor matching accuracy. Eliminating the false matches is essential for establishment of accurate matching of feature points between two neighboring images. In our research, a checking procedure is developed based on the mean angle among pairs to decide the validity of the matching. The mean angle is defined as follows:

$$\overline{\mu}_{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\left|\left\langle pair_i, x_{axis}\right\rangle\right| \qquad (1)$$

where $\left|\left\langle pair_i, x_{axis}\right\rangle\right|$ represents the i^{th} angle between the i^{th} pair and the horizontal axis; and N is number of pairs.

With a mean $\overline{\mu}_{\alpha}$ and a standard deviation, a point pair having an angle exceeding a preset threshold is eliminated. The filtered pairs are then associated with their corresponding depth values to establish correspondences in 3-D space. The process is iterative to eliminate all the unmatched pairs.

2) Distance-based filter

According to an important property is that the Euclidian distance between two matched key-points for correct pairs of correspondence should be invariant, an algorithm for filtering bad pairs can be defined as:

$$\overline{\mu}_l = \frac{1}{N}\sum_{i=1}^{N} dis\left(p_{t_i} - p_{s_i}\right) \tag{2}$$

where N is number of correspondences; and $dis\left(p_{t_i} - p_{s_i}\right)$ is Euclidean distance between point p_{t_i} and point p_{s_i}.

With a mean $\overline{\mu}_l$ and a standard deviation, any pair having a distance beyond a preset threshold is rejected.

3) Area-based filter

Similarly, the Area-based filter is designed to compare each triangle in the target with the corresponding triangle in the source. For accurate matching, the area of these corresponding triangles should be invariant. This means the areas of two matching triangles should be equal or very close.

4) Curvature-based filter

The curvature-based filter sets a radius, r, to search the neighboring region around a query point. Based on its Euclidean distance, the closest point is determined for every query point. Based on the searched neighbors, the curvature of the surface formed by the neighbors is determined for both source and target cloud. A threshold is set to discriminate bad matching for a data pair with an over-deviated surface curvature. This procedure work well for eliminating bad correspondences, but it could time consuming for searching neighbors.

C. Indoor environment mapping using refined corresponding pairs

To do mapping, in our work, the mobile robot moves in the indoor environment to collect its surroundings. The Kinect camera mounted on the robot provides both of color and calibrated depth information at each detection. Each pair of two consecutive measures is used to establish sub-sets of matched pairs in 3-D depth space. As mentioned above, these sub-sets certainly consist of invalid pairs due to the potential sensing noises. Therefore, the proposed strategy of the filtering is next used to reject all of invalid corresponding pairs; hence, the processed data only contains valid matched pairs. According to two sub-sets, P_t and P_s, of 3-D key points extracted from the refined corresponding pairs for each measure in the pair of two consecutive measures, the rotation matrix R and translation vector T between P_t and P_s should be precisely determined. To obtain the final transformation, the mean squares error, *MSE*, is defined as follows:

$$MSE_k = \frac{1}{N}\sum_{i=1}^{N}\left\|P_{t,i} - R\left(P_{s,i}\right) - T\right\| \tag{3}$$

where N is number of correspondences; and, k denotes the k^{th} iteration.

If MSE_k is less than a given threshold, it indicates the transformation between P_t and P_s is reached. The procedure will be repeated for the next pair of consecutive measures. Final transformations are used to reconstruct 3-D environment map of the environment. **Figure 4** shows the flowchart of the algorithm of real-time 3-D mapping for indoor environments using the refined corresponding pairs.

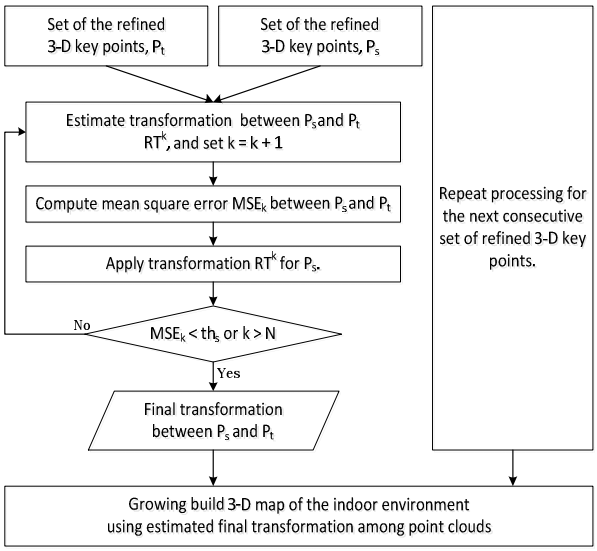


Figure 4. Flowchart demonstrating the algorithm of real-time 3-D mapping for indoor environments using the refined corresponding pairs.

III. EXPERIMENTAL RESULTS

In this work, the proposed algorithm is only applied for indoor environments due to the limitations of the RGB-D camera employed for a measurement range to be within 6.5 meters. However, for other data acquistion systems, such as systems employing laser rangefinders, the developed algorithm can be also applied effectively.

To demonstrate the performance of the algorithm, many various real-world experiments were performed for large-scale indoors space. Using the 3-D data acquisition system to acquire consecutive datasets for the following experiments: (1) pairs of far-field scene, (2) pairs of near-field scene, and (3) pairs of middle field scene. Figure **5** shows the result of feature detection and correspondence refinement for a far-field scene captured at a radiation testing room. The distance from the Kinect camera to the scene is around 2.5 meters. The number of points for the first point cloud dataset was 242,294 points; meanwhile, the one for the second dataset was 251,051 points. The threshold

set for the angle-based filtering was 5 degrees as 10 centimeters was set for the distance-based filtering, and 0.25 squares of centimeters for the area-based filtering. For the matched pairs of SURF features between two RGB images, the number of matched pairs without data filtering was first 324 pairs and reduced to 317 pairs when using the angle-based filter. The number of 3-D correspondences established by incorporating the SURF matched pairs using the corresponding depth was first 278 pairs in 3-D space. After applying the correspondence refinement algorithm based on three filters, the number of retained (valid) correspondences was reduced to 108 pairs

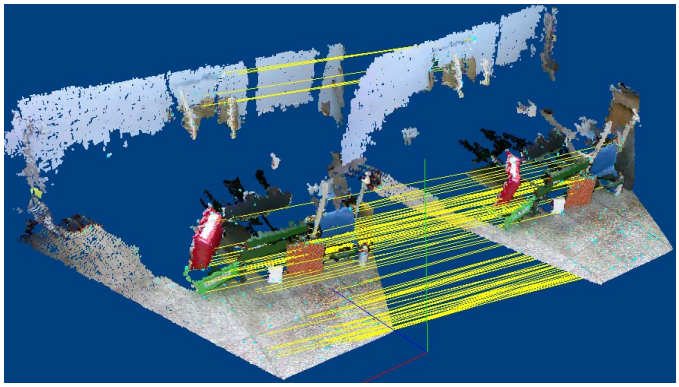


Figure 5. The result of feature detection and correspondence refinement for a far-field scene at a radiation source room in the Institute of Nuclear Energy Research (INER), Taiwan.

Figure **6** shows the result of a middle-field scene captured at a mechanical workshop. The distance from the Kinect camera to the scene is around 1.5 meters. The numbers of the first and second point cloud dataset were 220,628 and 226,858 points, respectively. The thresholds were set the same as the previous experiment. For the matched pairs of SURF features between two RGB images, the number of matched pairs was 522 pairs and reduced to 517 pairs after data filtering. The number of 3-D correspondences established by incorporating SURF matched pairs was 435 pairs. By applying the correspondence refinement algorithm based on three developed filters, the number of retained correspondences was reduced to 90 pairs.

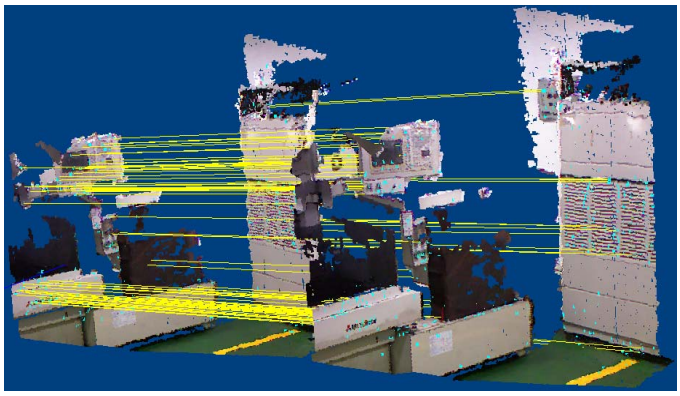


Figure 6. The result of a middle-field scene detected at the mechanical workshop in the National Taipei University of Technology (NTUT), Taipei, Taiwan.

Figure **7** shows the result of another near-filed scene. The distance from the Kinect camera to the scene is around 0.8 meters. The numbers of the first and second point cloud dataset were 248,751 and 249,247 points, respectively. The thresholds were set the same as the previous experiment. For the matched pairs of SURF features between two RGB images, the number of matched pairs was 1036 pairs and reduced to 1024 pairs after data filtering. The number of 3-D correspondences established by incorporating SURF matched pairs was 891 pairs. By applying the correspondence refinement algorithm based on three developed filters, number of retained correspondences was reduced to 456 pairs.

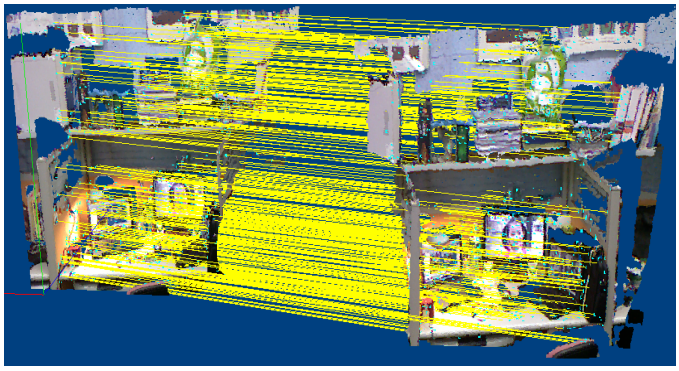
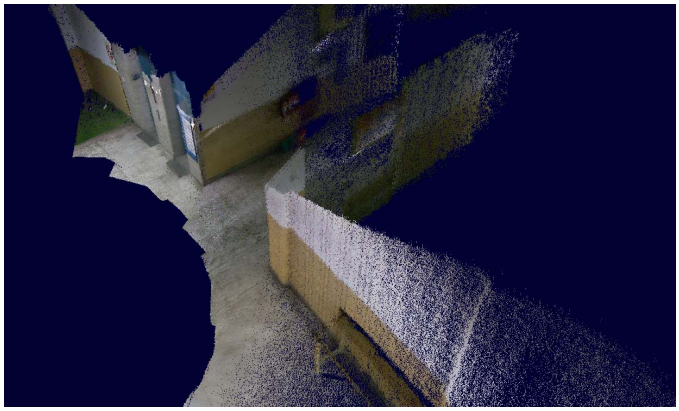


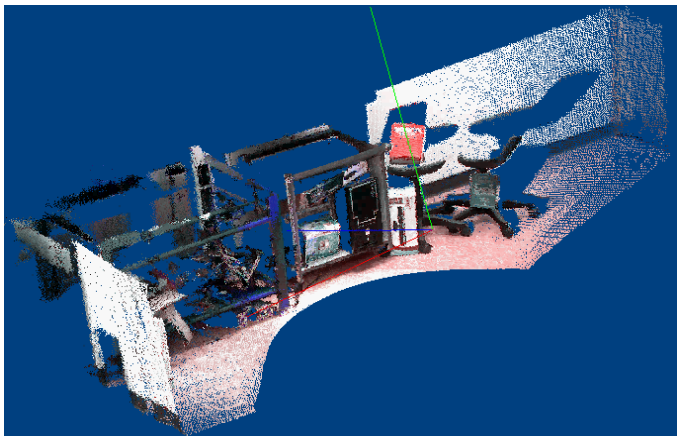
Figure 7. The matching result of a near-field scene of Thai's desktop at the AOI lab in the National Taipei University of Technology (NTUT), Taipei, Taiwan.

The number of retained valid correspondences after data filtering for scenes is different depending on the number of detected SURF features. However, for near-field scenes, the number of retained valid correspondences is usually larger than for other scenes due to the resolution of the point clouds in 3-D space.

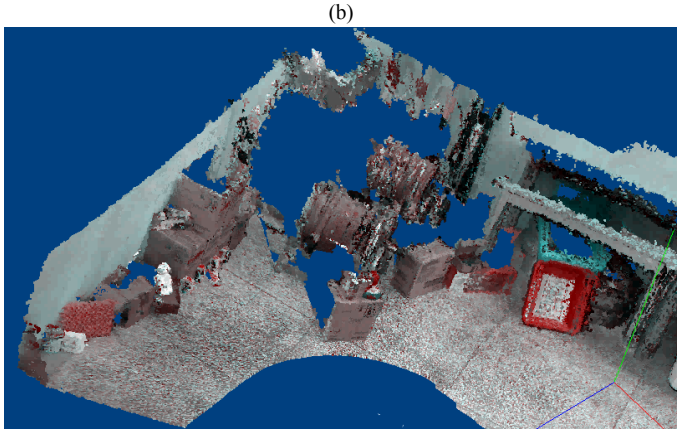
Three various environments have been configured for testing the algorithm of building maps using valid correspondences. Applying the proposed method to establish valid correspondences between pairs of consecutive scenes. These valid correspondences are then used to estimate transformations for mapping. Experimentally, each pair of consecutive scenes takes an average of around 1 second for alignment. **Figure 8** shows the reconstructed 3-D space maps of three different locations.



(a)



(b)



(c)

Figure 8. 3-D environment maps reconstructed by the proposed 3-D mapping algorithm for large-scale indoor environments. (a) The 3-D map of the corridor in front of the AOI lab in the National Taipei University of Technology (NTUT), Taipei, Taiwan. (b) The 3-D map of a small corner at the OI Lab in NTU; (c) The 3-D map of a small corner at the radiation source testing room in the Institute of Nuclear Energy Research (INER), Taiwan.

IV. CONCLUSIONS

The proposed algorithm for feature detection and correspondence refinement has been developed and tested under various indoor environments. The feasibility the algorithm was primarily verified. The number of valid correspondences is refined to keep high accuracy of matching between scanned images for applications in 3-D space such as mapping, segmentation and recognition. The developed filters are effective in removing invalid matched point pairs from the reconstructed images. The advantages achieved by the method lie in its enhanced registration efficiency and robustness. Some experiments were performed for data captured from the RGB-D camera and the results indicate the algorithm works satisfactorily for indoor environments. In the future work, the proposed algorithm could be further developed for outdoor environments by employing other 3-D sensors such as laser range finders for even wider applications.

REFERENCES

[1] L.C. Chen and N.V. Thai, "Real-time 3-D mapping for indoor environments using RGB-D cameras," *Advanced Materials Research*, vol. 579, pp. 435-444, 2012.

[2] T. Whelan, H. Johannsson, M. Kaess, J.J. Leonard, and J. McDonald, "Robust tracking for real-time dense RGB-D mapping with Kintinuous," *Technical report, Computer Science and Artificial Intelligence Laboratory*, Cambridge, 2012.

[3] A.S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an RGB-D camera," *The 15th International Symposium on Robotics Research*, Flagstaff, Arizona, USA. August 28-September 1, 2011.

[4] H. Du, P. Henry, X. Ren, M. Cheng, D.B. Goldman, S.M. Seitz, and D. Fox, "Interactive 3D modeling of indoor environments with a consumer depth camera," *Proceedings of the 13th International Conference on Ubiquitous Computing*, Beijing, China. September, 2011.

[5] C. Audras and A.I. Comport, "Real-time dense appearance-based SLAM for RGB-D sensors," *Proceedings of Australasian Conference on Robotics and Automation*, Monash University, Melbourne Australia, December, 2011.

[6] K. Pathak, N. Vaskevicius, J. Poppinga, M. Pfingsthorn, S. Schwertfeger, and A. Birk, "Fast 3D mapping by matching planes extracted from range sensor point-clouds," *The 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, St. Louis, USA. October, 2009.

[7] L.P. Ellekilde, S. Huang, J.V. Miró, and G. Dissanayake, "Dense 3D map construction for indoor search and rescue," *Journal of Field Robotics*, vol. 24, no. 1/2, pp. 71-89, 2007.

[8] T. Masuda, "A robust method for registration and segmentation of multiple range images," *Computer Vision and Image Understanding* vol. 61, no. 3, pp. 295-307, 1995.

[9] R. Newcombe, S. Izadi, O. Hilliges, D. Kim, D. Molyneaux, J.D.J. Shotton, P. Kohli, A. Fitzgibbon, S.E. Hodges, and D.A. Butler, "Moving object segmentation using depth images," *Patent Application Publication*, United States, 2012.

[10] B.B. Amor, M. Ardabilian, and L. Chen, "New experiments on ICP-based 3D face regconition and authentication," *The 18th International Conference on Pattern Recognition*, Hong Kong, China, August, 2006.

[11] A.S. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2D-3D hybrid approach to automatic face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, 2007.

[12] H. Chen and B. Bhanu, "Human ear recognition in 3D," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, 2007.

[13] C. Harris, "A combined corner and edge detector," *Four Alvey Vision Conference,* Manchester, UK, pp. 147-151, 1988.

[14] H. Wang, "A practical solution to corner detection," *Proceedings of the IEEE International Conference on Image Processing,* Austin, Texas, November 1994.

[15] S.M. Smith and J.M. Brady, "SUSAN – A new approach to low level image processing," *International Journal of Computer Vision,* vol. 23, no. 1, pp. 45-78, 1997.

[16] M. Trajkovíc and M. Hedley, "Fast corner detection," *Image and Vision Computing,* vol. 16, pp. 75-87, 1998.

[17] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[18] H. Bay, A. Ess, T. Tuytelaars, and L.V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding,* vol. 110, no. 3, pp. 346-359, 2008.

[19] R.B. Rusu, N. Blodow, Z.C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," *Proceedings of the 21st IEEE/RSJ International Conference on Intelligent Robots and Systems,* Nice, France, September, 2008.

[20] R.B. Rusu, N. Blodow, Z.C. Marton, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," *Proceedings of the IEEE International Conference on Robotics and Automation,* Kobe, Japan, May, 2009.