A novel plane based image registration pipeline with CNN scene parsing

1st Dingtian Yan
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
dyan@essex.ac.uk

2nd Huosheng HU
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
hhu@essex.ac.uk

***Abstract*—Plane is one of the most important element for indoor man-made structural rooms, such as broadcasting studios, open lecture room and empty offices. The existing visual mapping algorithms cannot effectively detect and describe the visual features on an empty large or medium size plane. This article aims to introduce a novel frame-to-frame registration pipeline based on one medium-size plane on man-made object instead of multiple planes or small plane patches. By introducing a structural description of reference planar area with its contour data and CNN segmentation information, the proposed approach is able track the pose of camera with high accuracy and robustness in comparison with existing feature-based tracking or dense geometric tracking approaches.**

***Index Terms*—Plane based image registration; CNN scene parsing; 2D contour extraction; Depth completion; contour-to-contour structural matching; Mapping in low-texture scene.**

I. Introduction

When performing the indoor mapping or camera tracking tasks, the large and medium size planes are selectively ignored since they can not provide enough visual features, especially when these planes are only partly visible in current camera perspective. The theory of visual reconstruction through depth information also faces the same problem. Because the imprecision of the experimental instrument and lacking of algorithms that can accurately describe the 3D planar model, most of the existing sense plane mapping algorithms usually use the plane to optimise the mapping result, instead of using the plane to complete the task of camera tracking. Therefore, the problem of visual reconstruction for low-texture scenes has become a practical research direction in recent years [1].

Plane is the most important part of real-world objects, and there are many planar areas in man-made indoor scenes, such as large planes including walls and floors; medium-sized planes like projection screens and surface of tables and various small planes like computer monitor. In one of the pioneering studies, Salas-Moreno deployed the planes to improve mapping performance [1]. It was mentioned that a dense tracking algorithm is more suitable for rich-planar environment, which can improve the tracking performance in the case of accurate extraction and description of planes. Therefore, they use the dense tracking algorithm to estimate the camera pose of each frame at beginning and then optimising each pose through connecting and refining planes.

Recently, Hsiao has introduced a key-frame based Dense planar SLAM with GPU since plane is a much cheaper dense presentation, and correctly extracted plane could be used as a landmark to reduce the drift error [2]. Therefore, a novel approach was proposed to convert the planar area into various small planar patches, and then track the camera by tracking corresponding patch pairs.

Based on existing studies on plane-based dense registration, this paper proposes a novel method with CNN labels to improve the frame-to-frame registration performance by tracking one marked plane inside the scene. This plane model is not needed to be acquired in priority and it is not required to be fully appeared in every frame. To detect and describe this plane, a CNN model and a novel composite contour extractor is utilised, and a depth completion algorithm is applied to get the correct geometric description. Then the camera motion estimation is accomplished by matching the geometric feature of labelled plane between two frames. Due to the use of CNN to label feature area and purposed depth completion process, this algorithm is more suitable for solving off-line visual mapping problems in indoor extreme scenes. The experimental results show that the average running time is 500 to 600 milliseconds without using GPU in 50 inter-frame registration.

There are three main steps in the proposed algorithm. In the first part, a scene parsing CNN model ”DeepLabV3+” is used to segment the image and label the landmark plane [3]. Then, a novel 2D contour extraction strategy is designed to get the complete contour of landmark plane. And a depth completion algorithm is used to compensate the imprecision of depth sensor reconstruct the 3D model of contour area. Finally, a normal dense point-point ICP algorithm is applied to calculate the transformation between two contour models in two different processes. In addition, purposed registering algorithm uses the Ref Model and Ray-cast theory to make every registration accurate and reliable.

The rest of this paper is organised as follows. Firstly, a brief review of recent research on plane-based visual mapping algorithm is reviewed in Section II; then, each part of the proposed algorithm is introduced in Section III and the experiment results are shown in Section IV to show the feasibility of the proposed approach. Finally, a brief conclusion and further plans are described in Section V.

II. Background

Recent plane-based visual tracking and mapping algorithms includes two main approaches: using plane for tracking or using plane for global alignment. In the former case, Taguchi et al. used a multi-stage process to optimise a 3D plane point cloud from raw depth data [4]. They divide the whole 3D data into various planes by extending a random plane point with its neighbours repeatedly, and then compute the pose transformation by matching threes matched plane-plane features or point-point features. Similarly, Kaess identified that at least three plane pairs is required to match two frames. Two angle values and one distance value were used to describe and distinguish each plane, and estimate the camera motion by measuring the distance between matched plane in a tangent space [5]. However, these approaches is not practical in feature-less scenes. Li used the small plane patch to enhance the registration performance, but a 21*21 small patch cannot make the system run in a low-texture environment [6]. Similarly, Hsiao used the key-frame structure for dense mapping [2]. By turning a complete plane into several facets, it solves the problem of inaccurate depth information at the plane boundary. Both photo-metric error and geometric error were used to estimate the pose transform matrix.

In the later case, Salas-Moreno used the connected component labelling technique to detect and mark planes [1]. This is a pioneering work in terms of using planes to optimise normal point-to-point dense registration results. Ma used a soft EM algorithm to extract and label the planes in each image, and register current frame to key-frame through matching points and planes to global model [7]. Then the detected areas are utilised to align the global model. ElGhor et al. detects the feature points and extract plane from point cloud through RACSAC concurrently. By projecting points onto corresponding plane and complementing the same plane, the influence of sensor error on the construction is greatly reduced [8]. Guo used the edge and structure of planes to optimise the low-texture visual odometry problems [9], and LSD and LBD were used to extract and describe the edge of planes respectively. However, this algorithm requires a very simple scene structure and is impractical. In large man-made indoor scenes, structural information is also considered as an important feature. Kim introduced their theory used for empty structural scene, such as the corridors [10]. The scene structure is consistent with the Cartesian coordinate. When a new frame was inserted into the loop, users could define its orientation by analysing the wall plane, ground plane and ceiling plane, and then comparing its orientation with Cartesian coordinate to estimate current pose.

III. Algorithm Overview

The purposed registration algorithm is based on point-point ICP algorithm. However, classic ICP-based dense visual tracking algorithm can not meet the requirements of our contour-contour matching theory, and it mainly includes two reasons: on the one hand, captured depth data, RGB image and CNN scene segmentation information are not accurate enough, and the contour extracted through these raw data will directly affect the registration performance; on the other hand, ICP algorithm is easy to fall into the local minimum problem due to its computation limitation when the difference between two images is large, and the matching score obtained in this case is unreliable. To solve these limitations, a novel frame-frame registration process is proposed based on global ref model and ray-casting theory. Figure 1 presents the processing pipeline of the proposed contour-to-contour registration algorithm. The target and process of each part is described in the following subsections: section A includes a novel contour extraction and modelling process; section B describes the raw contour-contour ICP registration and its limitations; section C and section D present the ref model theory and a self-calibration process respectively.

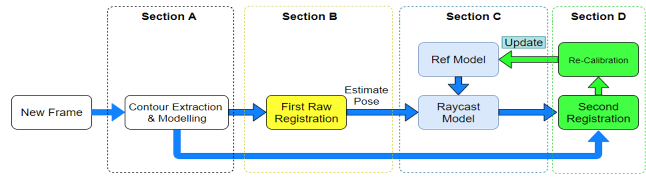


Fig. 1. The flowchart of the proposed Contour-to-contour registration pipeline

A. Contour extraction and Data Fusion

This section introduces a novel contour extraction and modelling process, which aims at recovering an accurate 3D contour model of landmark plane with noisy visual data and imprecise CNN image segmentation. This process consists of two parts. The first part is to accurately extract the edge points of the landmark plane in the non-HD image through the rough CNN scene segmentation and contour extractors. The other part is to correct the raw depth data of the contour region with normal vector of this plane, and then construct an accurate 3D contour model.

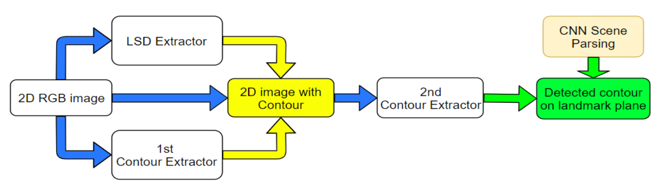
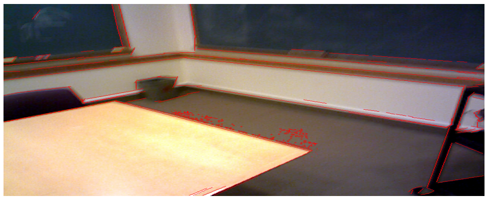
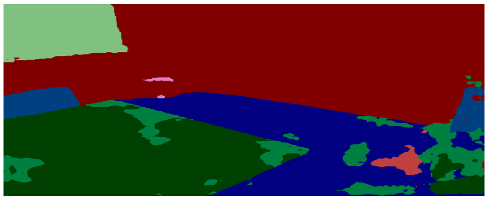


Fig. 2. The flowchart of the proposed Contour-Extraction Pipeline

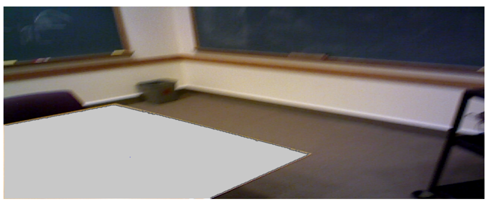
1) Composed Feature extraction process: The proposed contour extraction process is the combination of LSD extractor and basic contour extractor based on image intensity value, and its structure is illustrated in Figure 2. LSD is a frequently used line detection algorithm in recent studies and often used in conjunction with the LBD line description algorithm. However, it is difficult for LSD to find accurate edge feature points in non-HD images due to unbalanced illumination and shooting-jitter problems. Similarly, the normal contour extractor based on gradient change has the same limitation. Therefore, the purposed contour extraction pipeline combines the two extractors to compensate for each other's detection error. This process start with extracting the contour features on 2D RGB image with LSD and contour extractor at the same time, and project all the contour points back to original image to force an edge to large areas. Then, contour extractor



(a) Imprecise Contour Detection



(b) Rough Image Segmentation



(c) Optimised Contour Detection on landmark plane

Fig. 3. Contour Extraction on Labelled plane. (a) the red points describes the raw contour extraction result with LSD contour extractor and Intensity gradient based contour extractor. The false detection occurs because of unbalanced illumination (errors inside plane) and image quality (errors outside plane); (b) describes the rough scene segmentation from CNN model, the dark green indicates the right label, and the light green pixels are the false detection; (c) presents the plane and its contour extraction result with our approach, the table surface is labelled with light grey, and its contour points is drew with dark grey.

is used again to extract the contour in the enhanced image. By filtering out all the open contours and short contours, only the contours which describes the landmark plane and other large areas are saved. After 2D contour extraction, the CNN scene parsing model "DeepLabV3+" is used to define which contour describes the landmark plane. Due to the imprecision of CNN scene segmentation, if more than 70 percent of the whole pixels inside one contour area belongs to the landmark label, it means this area is describing the landmark plane. The Figure 3(a), 3(b) and 3(c) describe the raw contour detection, rough image segmentation and optimised contour extraction result with purposed algorithm respectively.

2) Depth Completion: After locating the 2D edge of a landmark plane, a depth completing process is triggered to correct the raw depth value acquired with off-shelf commodity cameras, and obtain an accurate 3D contour model of plane in each frame. In this section, a novel plane depth compensation algorithm based on normal vector equation is tested to get a correct 3D model of plane‘s contour. This process includes three steps:

a Split the extracted plane model with two points groups: $Cloud_{Inside}$ and $Cloud_{Edge}$;
b Choose two random point P_1 and P_2 in $Cloud_{Inside}$ and one point P_3 in $Cloud_{Edge}$;
c Build the residual equation with P_3 and P_1, and P_3 and P_2 to estimate the depth of each P_3 based on Normal Vector equation.

In Step a, the landmark plane is projected to space with raw depth value and then distinguished with the inside point cloud $Cloud_{Inside}$ and edge point cloud $Cloud_{Edge}$. The $Cloud_{Edge}$ is composed with the detected edge points and its nearest 15 neighbours toward the centroid of this plane, and the $Cloud_{Inside}$ is made up of the remaining points. Then, the normal vector of $Cloud_{Inside}$ is extracted as (a, b, c, d) in Step b. This normal vector is reliable since the inside points have less false depth values. Finally, By choosing two random point P_1 and P_2 in $Cloud_{Inside}$ and one point P_3 in $Cloud_{Edge}$, and then establishing the residual equations based on normal vector extracted in Step b, the depth of each point P_3 is acquired by solving these residual equations.

$$a\vec{x} + b\vec{y} + c\vec{z} + d = 0; \tag{1}$$

where a, b, c and d are derived by normal vector extraction, vectors **x**, **y**, **z** could be derived by ($P_3.x$ - $P_1.x$), ($P_3.y$ - $P_1.y$), ($P_3.z$ - $P_1.z$) or ($P_3.x$ - $P_2.x$), ($P_3.y$ - $P_2.y$),($P_3.z$ - $P_2.z$). Note that **x**, **y** and **z** of P_1 and P_2 are derived directly, $P_3.x$ and $P_3.y$ could be derived by Equations (2), (3) and (4).

$$P_3.z = depth/camera_{scale}; \tag{2}$$

$$P_3.x = (n - camera_{cx}) * p_3.z/camera_{fx}; \tag{3}$$

$$P_3.y = (m - camera_{cy}) * p_3.z/camera_{fy}; \tag{4}$$

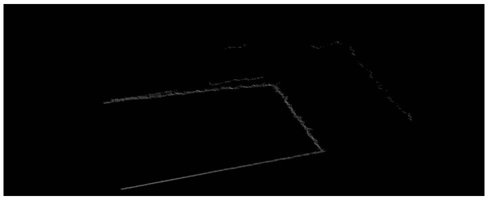
Where m and n are the 2D pixel position of P_3, $camera_{scale}$ is the camera depth scale; $camera_{cx}$ and $camera_{cy}$ are the camera principle points; $camera_{fx}$ and $camera_{fy}$ are the camera focal length. These camera intrinsic parameters could be derived from calibration directly.

By inserting these values in normal vector formula, we can obtain two residual equations, which could be solved through Ceres optimisation lib.

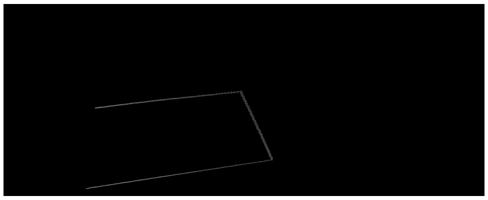
$$\begin{aligned} residual[0] = a * (P_3.x - P_1.x) + b * (P_3.y - P_1.y) \\ + c * (P_3.z - P_1.z) + d \end{aligned} \tag{5}$$

$$\begin{aligned} residual[1] = a * (P_3.x - P_2.x) + b * (P_3.y - P_2.y) \\ + c * (P_3.z - P_2.z) + d \end{aligned} \tag{6}$$

Figure 4 has shown the modelling result of our approach and proved the reliability of our proposed algorithm. When the 3D model is accurate, it could be treated as a two-dimensional image feature represented by three-dimensional information. Moreover, tracking this geometric feature is similar to 2D visual feature-matching theory with less number of visual features.

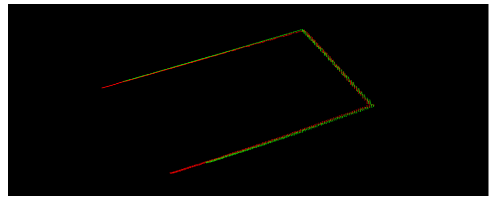


(a) Raw depth model

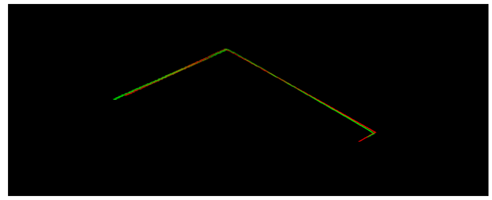


(b) Optimised model

Fig. 4. Comparison of modelling results between the original depth image and the ones from the depth completion algorithm. (a) is the 3D model with normal depth data. Some edge points are no longer in the plane area because of imprecise depth value. (b) is the 3D model optimised with the proposed depth completion algorithm, which has kept all the important contour information and could be considered as a 3D description of 2D contour area of plane.



(a) limitation 1



(b) limitation 2

Fig. 5. These two images describe the limitation of Raw Contour-contour Registration Results, green points and red points are from adjacent different frames. (a) the matching score of this scene = 0.00010304. As one model is totally overlapping the other one, it still gets a small matching score even when the registration is not accurate. (b) When the two models are partially overlapped, it is easy to fall into the local minimum problem, and its matching score is not reliable.

B. contour-to-contour Dense Registration

Dense registration could be classified by geometric approach and photo-metric approach. The geometric approach is using various ICP algorithm to match the geometric feature and the photo-metric approach is relying on intensity value. Our algorithm belongs to the geometric approach. With the inspiration of 2D line feature based visual tracking algorithm, this paper registers two planes by matching their 3D spatial edge features, which could be considered as a 3D matching approach on 2D image features. By accurately matching the edges with each other, two plane models could be accurately aligned, as well as the whole 3D scene. Point-point ICP algorithm also derives a matching score, which describes the matching accuracy. And a smaller score indicates the better registration performance.

This approach provides good matching results when two 3D models are similar to each other. However, when two models are very different (e.g. long distance between two frames), this approach is easy to fall into local minimum and its derived matching score is no longer reliable, as shown in Figure 5. This limitation is caused by point-point ICP algorithm itself. Therefore, a novel reference model theory and re-calibration method based contour-contour matching algorithm is deployed in the next subsections for optimisation.

C. Ref Model in inter-frame registration

Using Ray-casting theory and Global Model to replace normal frame-frame tracking has proved its efficiency in recent dense visual mapping algorithms, since it can effectively reduce the drift error as this method matches each newly added frame with the global model instead of a continuous inter-frame registration. Based on this theory, this paper introduces a novel global reference model and ray-casting theory specific to purposed contour-contour matching algorithm to improve its robustness and accuracy. Compare with existing theory, our algorithm has made three main changes:

1) Reconstructing and updating a special ref model besides the global reconstructed model. This ref model is presented with point cloud instead of voxel-based reconstructed model. This makes the system saves the process of point-voxel converting and easy to change or delete the points.
2) The ref model updates its viewed part in current perspective when the pose of current inserted frame is confirmed, and only part of new model will be added to the ref model. This keeps the simplicity and accuracy of ref model.
3) When tracking failure occurs, global ref model will stop updating and the current ref model will be saved as a piece-wise contour model.

The ref model $Model_{Ref}$ starts with the initial contour model in a image sequence. Assuming the current frame is $Frame_i$ and the ref model is $Model_{Ref}$, the up coming Frame is $Frame_{i+1}$, its contour depth image is $Depth_{i+1}$, and its spatial point cloud in world coordinate is $Cloud_{i+1}$. Updating the ref model includes four steps: (1) When the pose P_{i+1} of new inserted frame $Frame_{i+1}$ is defined, the $Model_{Ref}$ is split into the viewed model $Model_{viewed}$ and unviewed model $Model_{unviewed}$ in the first stage; (2)only the points in $Cloud_{i+1}$ that is surrounded by the points in $Model_{viewed}$ will be added to the $Model_{viewed}$. (3) the viewed model $Model_{viewed}$ is warped to $Frame_{i+1}$ and then projected back to world coordinate for updating. (4) the updated viewed model $Model_{viewed}$ is combined with unviewed model $Model_{unviewed}$ to form an updated landmark ref model. '

When the new model is out ranging the ref model, the system will detect the out-ranged points by detecting their surrounding points. For example, if a point P_{i+1} in $Cloud_{i+1}$ is only surrounded by points of the same model and this point is not connecting to the $Model_{viewed}$, then this point belongs to the newly added points. In order to maintain the accuracy, the proposed system will only add new points to the ref model when the number of newly added points is greater than 30, which is described in Figure 6.

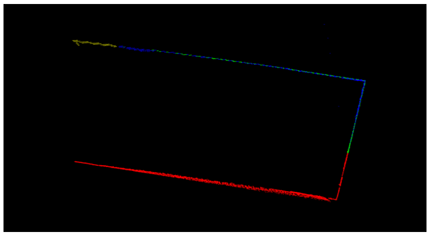


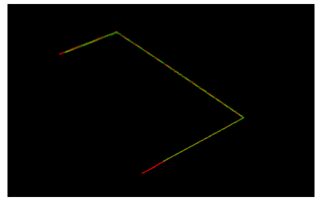
Fig. 6. The updated ref model. The yellow point in the up-left corner is the newly added points; the red describes the unviewed part of ref model $Model_{unviewed}$ in the current camera perspective; the blue points are the viewed points in the ref model and the green points are the updated points to the viewed model $Model_{viewed}$.

D. Self-Calibration

Self-calibration aims at solving the limitations of raw contour-contour ICP algorithm and deriving an accurate matching score, and it is the second ICP process to optimise the raw pose estimation result from first contour-contour registration. Its main computing process is using estimated camera pose $\hat{Pose}_{i+1}$ from the first contour-contour ICP registration to ray-cast ref model and get its viewed part $\hat{Model}_{viewed}$, and then optimising the predicted pose $\hat{Pose}_{i+1}$ by defining the spatial relationship between predicted $Model_{viewed}$ and real model $Cloud_{i+1}$. This method can effectively optimise the local minimum problem and significantly improve the accuracy of pose tracking since the difference between the estimated camera pose $\hat{Pose}_{i+1}$ and the real pose $Pose_{i+1}$ are smaller than $Pose_i$ of $Frame_i$ and $Pose_{i+1}$ of $Frame_{i+1}$. Moreover, the matching score derived by this approach is reliable and accurate enough to describe the inter-frame matching accuracy due to $Model_{viewed}$ and $Cloud_{i+1}$ are similar, And it can be used as the main criteria in visual tracking algorithms to determine if tracking lost happens.

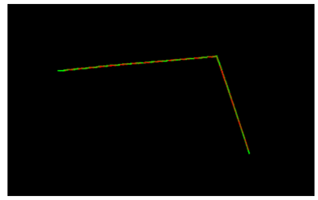
IV. Experimental Results

The experiments were conducted to verify the validity of the proposed algorithm in this section. To verify the practicability of algorithm with different camera motions, we first extracted the key-frame sequence through orb-slam, and use the proposed algorithm for key-frame to key-frame registration. In this experiment, the image sequence and camera parameters are obtained from SUN3D Database (Harvard- c11/hv-c11-2) [11], and the surface of the table inside this scene is considered as the landmark plane. Three tests were implemented to prove the reliability and robustness of the proposed algorithm: the first one is demonstrating the feasibility of purpose theory



(a) matching result with three edges

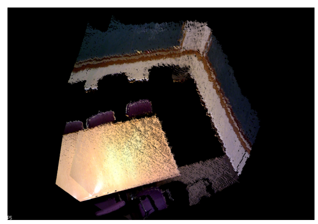
(b) matching result with three edges



(c) matching result with two edges

(d) matching result with two edges

Fig. 7. Registration results, the green point and red point are from different images: (a) and (b) Registration with 3 edges, the matching score = 0.00007362; (c) and (d) Registration with 2 edges, the matching score = 0.000082763



(a) Scene-scene ICP

(b) Contour-contour ICP

Fig. 8. Comparison between normal scene-scene ICP (a) and our contour-contour ICP (b)between two adjacent key-frames.The matching score derived from (a) equals to 0.067, and then matching score obtained from (b) equals to 0.011

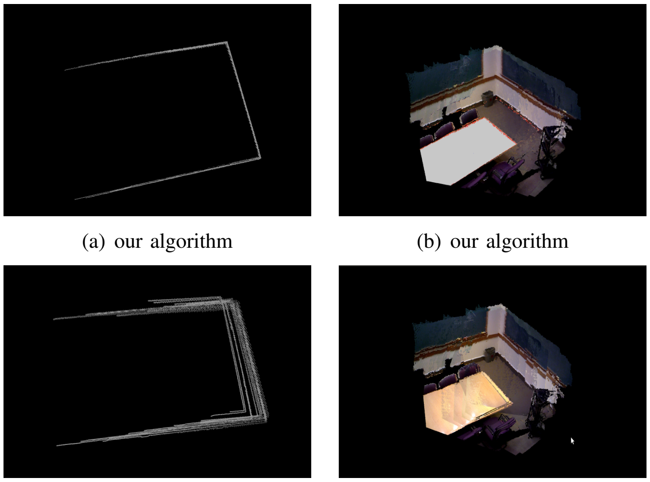
when part of landmark plane is visible, and the other two are describing the comparison between scene-scene ICP and existing ORB-SLAM respectively.

In the first experiment, the inter-frame registration performance is tested when there are two edges or three edges are visible in current perspective, and the experimental result is shown in Figure 7.

Figure 8 describes the second test, which aims at identifying purposed algorithm is more efficient than normal scene-scene ICP registration. Since scene-scene ICP is very time consuming, only 1 key-frame pair is displayed as a comparison.

The third experiment is proving proposed algorithm is able to implement in visual mapping programs through mapping a sequence of continuous key-frames and comparing the mapping result with ORB-SLAM. Both the reconstructed contour model and scene model are present to make the comparison clearer. This experimental result is described in Figure 9, both the reconstructed contour model and scene model are present to make the comparison clearer.

In Table 1, the 3D contour models were transferred with the estimated pose computed by three algorithms to obtain a balanced and numerical comparison. The distance between clouds was computed by the average distance between points.



(a) our algorithm (b) our algorithm

(c) ORB-SLAM (d) ORB-SLAM

Fig. 9. The comparison between our algorithm and ORB-SLAM with mapping a 15 key-frame sequence without global loop closure optimisation. (a) and (b) are describing the mapping result with our algorithm. (a) is the recovered contour model and (b) is the recovered scene model. The white are in (b) is the ref plane model and the red points are its contour. (c) and (d) are describing the mapping result with ORB-SLAM without loop closure optimisation. To get the contour model, the pose of each key-frame is estimated firstly, and then register each contour model to space with estimated poses.

The derived matching score is the average score of 25 different matching frame pairs.

TABLE I
THE COMPARISON BETWEEN NORMAL SCENE-SCENE ICP, ORB-SLAM AND PURPOSED CONTOUR-CONTOUR REGISTRATION RESULT

Method	**Matching Score**
scene-scene ICP	0.07635
ORB-SLAM	0.03602
purposed contour-contour ICP	0.00932

V. CONCLUSIONS AND FUTURE WORK

This paper has introduced an off-line plane-based inter-frame dense registration algorithm. Alternative to recent photometric error based visual tracking and mapping algorithms, the purposed algorithm estimates the pose of each input frame by tracking the geometric feature of one medium-size plane inside the scene. As a conclusion, our system has made three main contributions as follows:

- Firstly, the purposed inter-frame registration algorithm only depends on one medium-size plane, and this plane is not required to be completely visible in every frame. This makes our algorithm implementable in low-texture indoor scenarios.
- Secondly, a line detection and depth completion algorithm is introduced to get the optimal 3D contour data of planar area with raw RGB, Depth image and rough CNN scene parsing result. The obtained 3D contour model describes the 2D contour features in 3D space, and by tracking this 3D geometric feature, system could more accurately and robustly obtain the spatial relationship between two frames.
- Thirdly, specific to purposed contour-contour ICP registration, we suggest to use a global ref model and ray-casting theory to optimise the limitation of ICP algorithm itself. This approach makes our registration algorithm practical in long-distance inter-frame matching, such as key-frame to key-frame registration. Moreover, the derived matching score with our approach accurately describe the matching accuracy, which makes it implementable in visual tracking and mapping programs.

However, this algorithm has an inevitable limitation. This algorithm is tested only in off-line processes, since CNN scene parsing and depth completion process is time consuming. For registration process, the experiments show that it spent an average of 500ms for registration without using GPU. In the further studies, this algorithm will be implemented on GPU and real-time visual mapping programs. Moreover, this registration process will be firstly tested in more scenarios, and labelled area will be changed to different typed of planar area. Then, this algorithm will be used in the complete indoor mapping algorithms, and a complete visual mapping system based on single plane for low-texture environment will then be researched, including local loop-closure optimisation, global loop-closure detection and loss tracking recovery.

REFERENCES

1 Salas-Moreno, R. F., Glocken, B., Kelly, P. H., and Davison, A. J., “Dense planar slam,” in *2014 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 2014, pp. 157–164.

2 Hsiao, M., Westman, E., Zhang, G., and Kaess, M., “Keyframe-based dense planar slam,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5110–5117.

3 Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H., “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *ECCV*, 2018.

4 Taguchi, Y., Jian, Y.-D., Ramalingam, S., and Feng, C., “Point-plane slam for hand-held 3d sensors,” in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 5182–5189.

5 Kaess, M., “Simultaneous localization and mapping with infinite planes,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 4605–4611.

6 Li, R., Liu, Q., Gui, J., Gu, D., and Hu, H., “A novel rgb-d slam algorithm based on points and plane-patches,” in *2016 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 2016, pp. 1348–1353.

7 Ma, L., Kerl, C., Stückler, J., and Cremers, D., “Cpa-slam: Consistent plane-model alignment for direct rgb-d slam,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1285–1291.

8 Elghor, H. E., Roussel, D., Ababsa, F., and Bouyakhf, E. H., “Planes detection for robust localization and mapping in rgb-d slam systems,” in *2015 International Conference on 3D Vision*. IEEE, 2015, pp. 452–459.

9 Guo, R., Zhou, D., Peng, K., and Liu, Y., “Plane based visual odometry for structural and low-texture environments using rgb-d sensors,” in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2019, pp. 1–4.

10 Kim, P., Coltin, B., and Jin Kim, H., “Linear rgb-d slam for planar environments,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 333–348.

11 Xiao, J., Owens, A., and Torralba, A., “Sun3d: A database of big spaces reconstructed using sfm and object labels,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1625–1632.