A 3D-2D Registration Method for Stereo Scan Overlay on Structure from Motion Model

Deepak Rajamohan
School of Engineering and Information Technology
University of New South Wales
Canberra, ACT, Australia
Email: Deepak.Rajamohan@student.adfa.edu.au

Matthew Garratt
School of Engineering and Information Technology
University of New South Wales
Canberra, ACT, Australia
Email: M.Garratt@adfa.edu.au

Mark Pickering
School of Engineering and Information Technology
University of New South Wales
Canberra, ACT, Australia
Email: M.Pickering@adfa.edu.au

***Abstract*—The ability to detect and analyze changes or understand the scene while navigating close to buildings is very important for autonomous aerial and ground vehicle based surveillance applications. For this, the latest textured 3D scan of the platform's view frustum has to be placed accurately in the context of a big map like a Structure from Motion (SfM) map of the region. However, due to the drift in the camera trajectory, the scans are usually not aligned with the SfM model. This paper proposes a novel registration algorithm that aligns the 3D scan using known 2D images of the SfM model. The proposed 3D-2D registration method uses a heuristic approach which first performs a robust 2D-2D registration between the projection of the 3D scan and the SfM images and then calculates the 3D alignment parameters by combining registration results of multiple camera views of the SfM model. The results presented compare the robustness of the proposed registration techniques with traditional approaches.**

Fig. 1: A low resolution graffiti wall SfM map overlaid by dense stereo scan

I. Introduction

Unmanned vehicle platforms are built with a variety of sensors depending on the application requirements. Sensors like a stereo camera or an RGB-D camera on these platforms can create small spatially accurate 3D scans of the view frustum. Overlaying such 3D scans on a pre-built Structure-from-Motion (SfM) map can help obtain a spatial context for the scan and the derived information, under complex navigation environments such as when close to buildings. One of the example applications would be to overlay a dense stereo scan over interested regions of a low resolution SfM map. For example, the figure 1 shows the SfM map of a graffiti wall, partly overlaid by a dense 3D scan of the same. Scan overlay could be used to perform external inspections of large real-estate properties and log maintenance requests with a spatial context. Since stereo cameras are best suited for outdoor applications, we decided to overlay the textured 3D scan patches from a stereo camera over a SfM model. However, when overlaying such information, the spatial correctness depends on the estimated trajectory of the platform using Simultaneous Localization and Mapping techniques (SLAM). Even if the coordinate systems of SfM and SLAM are aligned initially, a gradual drift misaligns the scan. Also, navigating close to buildings is challenging for GPS aided navigation due to reduced satellite visibility and multi-path reflection. Hence, stereo scan alignment on SfM map needs to use additional information apart from relying on the estimated trajectory.

This paper proposes a novel 3D scan alignment procedure which overlays the scan on the SfM map, provided that the scan pose initialized using the trajectory information is close to the ground truth. The method first registers the projection of the textured 3D scan across multiple camera views of the SfM map by combining edge and point feature information. The registration information across multiple views are then combined to estimate a rigid 3D affine transformation using a least squares approach. Figure 2 summarizes the approach of the proposed scan alignment procedure.

II. Related Work

Most of the reported algorithms related to this work purely solve the localization and mapping problem. Dense mapping techniques, used in the Kinect Fusion [1], [2] family of work, utilise RGB-D sensors to incrementally stitch the depth maps using the Iterative Closest Point (ICP) algorithm [3] to perform odometry calculations. Similarly, the so called direct methods [4], [5], [6] use a dense approach which uses the rich pixel information of a single camera to solve the SLAM problem. In all these methods, the latest scan is registered on to a previous scan within a short time interval between the two scan operations. But in case of scan overlay on a SfM map, there could be a significant change in appearance that poses challenges in the alignment. Our proposed method follows a different approach for scan alignment using 3D-2D registration.

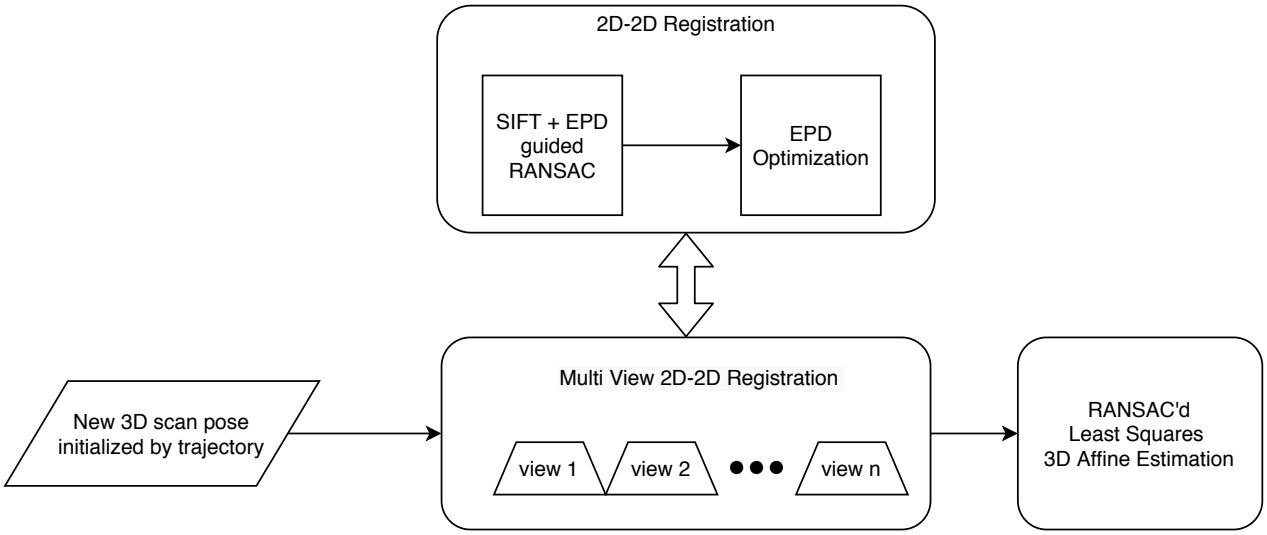


Fig. 2: Process block diagram

Traditionally, 3D-2D registration has taken a heuristic approach that uses successive 2D-2D rigid affine registration and approximates the 3D transformation each time [7], [8]. Hence the robustness of the process improves with the robustness of the image registration algorithm used. Image registration techniques are heavily application dependent. Registering the projection of textured 3D scans created using a stereo camera and the SfM images can be challenging due to the change in lighting conditions and the geometrical distortions in the scan. Although the application considered does not involve purely multi-modal registration, the challenging conditions prompted us to adopt a few of the techniques commonly used to solve such problems. In one such approach [9], Pickering used the conditional distribution of pixel values in the moving image with respect to each pixel value of the reference image to register X-Ray and CT-Scan medical images. Edge information of an image is robust to minor geometrical distortions and has been successfully used for optimal image registration in a method called Edge Position Difference (EPD) [7]. This heuristic method was adopted for a 3D-2D registration problem that aligns stereo camera scans in [8]. In addition, the authors used pixel information along the edges to improve registration. Both these techniques use a brute-force search for closing in on the solution and hence are time consuming for wide-baseline registration.

Feature based approaches use interest point detectors and descriptors such as SIFT [10], SURF [11], and ORB [12] to first find the point pairs between the reference and moving images and then fit a transformation model using the Random Sample Consensus (RANSAC) [13] algorithm which filters outliers. They are suited for wide baseline registration while the edge base approach provides fine registration. We propose a novel image registration technique that combines the benefits of the edge based approach, EPD, with SIFT feature-based registration. The contribution here is the design of a guided RANSAC approach that uses both the information and optimizes the estimated model using EPD algorithm. Another contribution of the paper is the novelty of our overall framework (figure 2) which uses the novel 2D-2D image registration method to achieve a 3D-2D scan alignment.

The remainder of the paper is organized into two major sections, with section III describing the proposed methods for 3D scan alignment and section IV explaining the experimental setup and the results from scan alignment.

III. Methodology

A. Image Registration for 3D Scan Alignment

The projection of a textured 3D scan will match the camera image of the scene if both the 3D scan and the camera view are registered on to a global map. In our system, the global map is a SfM model. However, the scan is not perfectly aligned with the map due to the drift in the camera trajectory. So with known camera poses and corresponding images of the SfM model, the 3D scan alignment is achieved by solving a 3D-2D registration problem. This type of problem arises commonly in medical imaging applications, such as CT scan to X-Ray registration [14], [15], where the 3D-2D registration is achieved by successive 2D-2D image registration. EPD is a 2D-2D registration method used as part of the proposed 3D scan alignment procedure. The following is a brief description of its operation. When edges R_e of the reference image and those of the moving image M_e are overlapped without any registration error, the sum of distance between each edge pixel p_i in a moving image, to its closest edge pixel q_j in the reference image would be at its global minimum. This sum is called the Edge Position Difference (EPD) as per equation 1 and is a continuous function in the space of parameters in the transform T.

$$EPD = \sum_{\forall p_i \in M_e} \min_{\forall q_j \in R_e} dist(p_i, q_j)) \tag{1}$$

To formulate EPD as a cost function, a Chamfer distance transform [16], [17] is applied to the reference edge image leading to a distance map D. It provides an instantaneous EPD value when the edge of the moving image is overlayed on to it

(a)

(b)

Fig. 3: An example reference image and its distance map

as per equation 2. One of the images used for experimentation and its corresponding distance map is shown in figures 3a and 3b respectively. The distance map D can be used to calculate the derivatives of the transformation parameters needed for registration. These include x - transformation along x axis of the image plane, y - transformation along y axis of the image plane , θ - transformation around the axis perpendicular to the image plane and a scaling parameter s. The derivative of D with respect to each of the transformation parameters ($\frac{\partial D}{\partial x}$, $\frac{\partial D}{\partial y}$, $\frac{\partial D}{\partial \theta}$, $\frac{\partial D}{\partial s}$) is calculated either directly or using the chain rule as per the equation 3. The parameters are denoted by a common variable $\boldsymbol{m}$. A gradient descent optimization method was used to search for the parameters leading to the registration.

$$EPD = \sum_{\forall p_i \in L_{M_e}} D(p_i) \tag{2}$$

$$\frac{\partial D}{\partial m} = \frac{\partial D}{\partial x} * \frac{\partial x}{\partial m} + \frac{\partial D}{\partial y} * \frac{\partial y}{\partial m} \tag{3}$$

B. Avoiding local minima

The EPD method enables a smooth convergence to the solution as long as a coarse registration step initializes it so that the EPD optimization starts close to the global minimum. In most cases of wide baseline registration, the gradient descent optimization will converge to a local minima, as there can be repeated edge patterns. The applications discussed in [7] and [8] use a brute force approach for the coarse registration where the search space was uniformly scanned, leading to a waste of resources. To reduce the search space, a novel heuristic approach is proposed. In our proposed algorithm, a SIFT based wide baseline registration precedes the EPD optimization step. SIFT features are extracted from both the images and matched in the Euclidean space. Depending on a selection criteria such as Lowe's ratio test [10], only the best point pairs are selected for the next step. The RANSAC is an iterative algorithm that removes outliers that do not obey the Gaussian noise assumption. The idea behind this algorithm is that inliers lie clustered together while outliers are scattered from each other. At every iteration, it selects a random set of the minimum number of samples needed to fit a transformation model. For the case of an affine transformation, just two point pairs are required. The model is then used to find the consensus set, or the inliers. The model that has the most number of inliers is the best fit. The inliers corresponding to the best fit are then used to re-estimate the model. RANSAC uses the number of inliers to probabilistically calculate the minimum iterations required to estimate a model such that no outliers are chosen in the sample set.

In cases where the number of outliers outnumber inliers, RANSAC might result in an incorrect model fit, especially when the outliers are clustered together. Projection of the textured 3D scans occupy only a small section of the reference image and hence results in more outliers. As a result, the RANSAC search needs to be augmented with more information that can help filter the outliers. Tal et al. [18] discusses a modified RANSAC approach for matching images with computer generated 3D models of the subject. The approach uses multiple quality measure information to calculate a total hypothesis score that is used to calculate the best model. Inspired by this method, the EPD score is used, along with the normalized inlier count, to arrive at the best model as per equation 4. W_1 and W_2 are the empirical weights that control the influence of both the information in the hypothesis score H_{score}.

$$H_{score} = W_1 \frac{\sum Inliers}{\sum (Inliers + Outliers)} + W_2(1 - EPD) \tag{4}$$

The EPD score calculated at every iteration serves as a guided approach to reject clustered outliers. This causes the iteration count update step of the RANSAC algorithm to occur less often than when using only the inlier count, but also ensures more outliers are rejected. Though this modification takes additional time, the proposed algorithm helps the optimization to converge to the global registration solution. The selected inliers are finally used to calculate a more accurate solution and serves as a good initialization for the EPD optimization.

C. Multi-view registration

The scan alignment is repeated across multiple adjacent views of the SfM camera poses and a single homography matrix is estimated for each view. Figure 4 is a scene from one of the experiments which shows a scan initialized by the SLAM trajectory and hence is misaligned when viewed from the different SfM camera views. Figure 5 shows the relation between the different coordinate systems involved in the 3D alignment application. It shows how a 3D transformation RT is observed as different 2D transformations H_n from N camera views. The local co-ordinate system where the 3D transformations are estimated has its XY axis plane XY_{pl} parallel to the camera's image plane. The origin of the co-ordinate system is the intersection between the camera principal axis and the plane XY_{pl}. So the in-plane transformation parameters include the translation along the X and Y axis, and also the rotation around the Z axis. Also, the scaling in the image plane is proportional to the translation along the Z axis of the 3D scan's local coordinate system and hence this is also picked

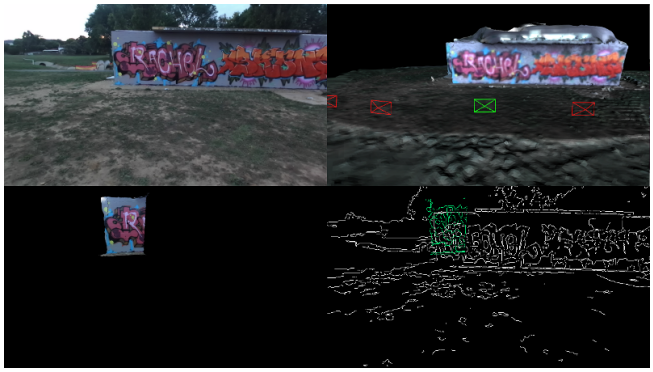


Fig. 4: A scene from a multi-view registration experiment. SfM camera image (top left), cameras selected in the SfM model (top right), misaligned ZED stereo scan (bottom left), overlap of edges from camera image and the rendered scan (bottom right)

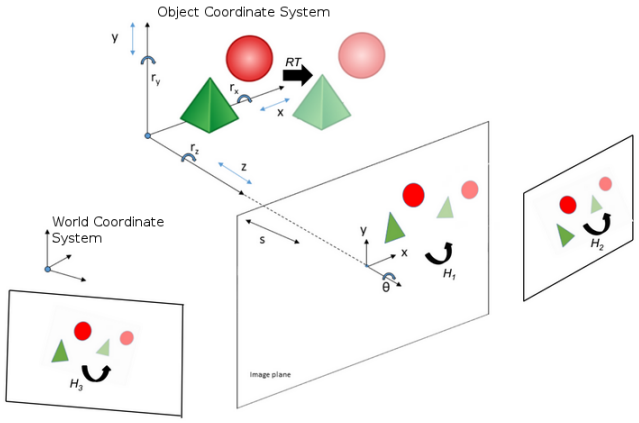


Fig. 5: Multi-view 3D-2D Registration Setup

up by the 2D homography matrix. Parallel image registration is done between the selected SfM images and the respective projection image. The result is a set of N 2D homography matrices H_n, capturing the 3D in-plane transformation that occurs with respect to the corresponding camera view.

Since the global location of the 3D scan before alignment is known, the image coordinates $(dc1, dc2)$ are calculated using the projective geometry of 3D points X on the scan. The new display coordinates are calculated using the homography matrix estimated from the image registration process. It can also be calculated and by projecting X multiplied by the unknown matrix RT (equation 12) that includes 9 elements r_{ij} of a 3x3 rotation matrix and 3 elements t_i describing the translation parameters. This can be formulated into a system of linear equations that is solvable provided we consider image registration from multiple views. The textured scan projection across multiple SfM camera views needed to use 3D graphics rendering software library. Also the stereo scan from the ZED camera and dense SfM data were available in standard 3D graphics file formats. Hence the least squares solution was obtained using the graphics projection equations.

The conversion of homogeneous 3D world point X to 2D display coordinates $(dc1, dc2)$ involves projecting the world point to 4D viewport coordinates V as in equation 5. After applying perspective divide in equations 7 and 8, the display coordinates are calculated as per the equations 9 and 10. The 3D scan needs to undergo an unknown 3D transformation RT that will satisfy all the H_n estimated 2D transformations using the image registration method discussed in sections III-A and III-B. This leads to a new set of viewport coordinates V_{new} and thereby new display coordinates dc_{1new}, dc_{2new} as per equations 11 and 13.

$$V = [vc_1 \quad vc_2 \quad vc_3 \quad vc_4]^T = MVP \cdot X \tag{5}$$

where MVP is the known 4X4 model-view-projection matrix

where X is the homogeneous 3D point given by

$$X = \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{6}$$

$$vc_1 = \frac{vc_1}{vc_4} \tag{7}$$

$$vc_2 = \frac{vc_2}{vc_4} \tag{8}$$

$$dc_1 = (vc_1 + 1) \cdot \frac{wn_x(b_3 - b_1)}{2} + wn_x \cdot b_1 \tag{9}$$

$$dc_2 = wn_y - (vc_2 + 1) \cdot \frac{wn_y(b_4 - b_2)}{2} + wn_y \cdot b_2 \tag{10}$$

$$V_{new} = MVP \cdot RT \cdot X \tag{11}$$

where

$$RT = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{12}$$

$$\begin{bmatrix} dc_{1new} \\ dc_{2new} \\ 1 \end{bmatrix} = H \cdot \begin{bmatrix} dc_1 \\ dc_2 \\ 1 \end{bmatrix} \tag{13}$$

where wn_x and wn_y are the graphics render window dimensions and b_1, b_2, b_3, b_4 are bounds of the graphics viewport. All these values are known. The right hand side of equation 13 is known, and the left hand side can be expanded in terms of the 12 unknowns of the matrix RT. The unknowns include 9 elements r_{ij} of a 3x3 rotation matrix and 3 elements t_i describing the translation parameters. This results in two equations for a pair of points related by homography. So, for a single view, if we take n known point pairs, we obtain $2n$ equations forming a system of linear equations of the form $Ax = b$. The matrix A was found to be rank deficient if we consider just one view, but can be formulated into a solvable system of linear equations when considering multiple views. All the N views selected for scan alignment may not provide a good estimate of the matrix H_n. Hence the camera views are filtered into inliers and outliers using the RANSAC algorithm with the 3D transformation RT as the model being estimated.

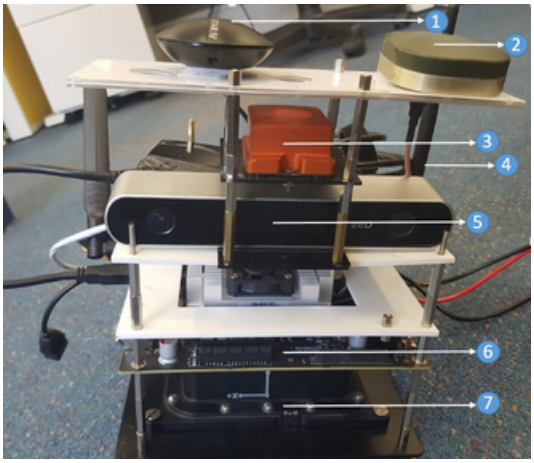


Fig. 6: Sensor platform

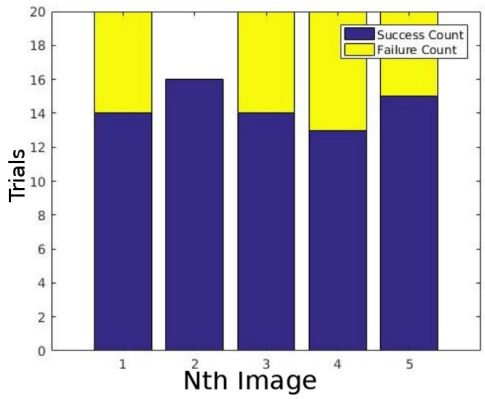
A final estimate of RT is calculated only using the inlier views. Figure 4 shows a mis-aligned 3D scan from the perspective of one of the selected camera views. The figure also shows the edges in the SfM camera image and the projection from the 3D scan being overlapped to begin the 2D-2D registration procedure for that particular view.

IV. Experimental Results

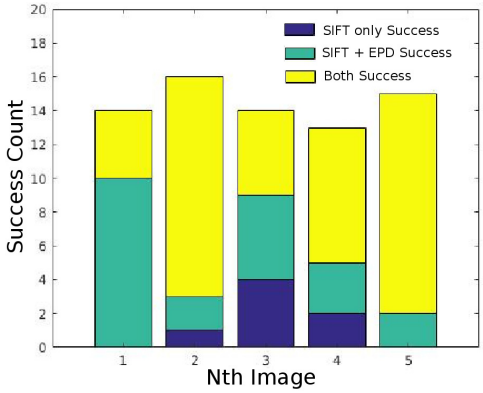
Figure 6 shows the prototype device that was built for conducting the experiments. The parts marked in the figure are: 1) Low cost GPS, 2) Novatel GPS antenna, 3) Xsens Inertial Measurement Unit (IMU), 4) Radio link for Differential GPS corrections, 5) ZED stereo camera, 6) NVIDIA Jetson board, and 7) High Fidelity Novatel SPAN RTK-GPS. The dimension of the sensor platform is 17cm in length and 17 cm in breadth and around 20 cm in height and weighs approximately 2Kg. The experiment was done offline by first collecting the trajectory and scan data and then using the data to demonstrate scan alignment.

The proposed novel 2D-2D registration procedure uses two different types of information. As discussed in section III-B, the EPD based similarity measure is used to guide feature based registration, resulting in better initialization for the succeeding optimization step. Figure 8 compares the registration success with and without using the EPD measure inside the RANSAC algorithm, constituting the two test cases being compared. In order to establish the performance improvement when using the proposed technique, experimental trials were conducted on five scan datasets.

Random sections of a large SfM map were chosen. A ZED stereo camera was used to collect 3D scans of the chosen scenes. After performing a one-time scale and field-of-view calibration, the stereo scan was aligned manually using multiple SfM camera views of the scene. This aligned pose was considered to be the ground truth. Then the scan was perturbed by a bounded, but arbitrary, 6DOF transformation with fixed scale. The projection images on the chosen SfM camera views were then used for the multi-view 3D-2D registration procedure. A set of N random 3D points on the surface of the stereo scan were chosen and the projection of those



(a)

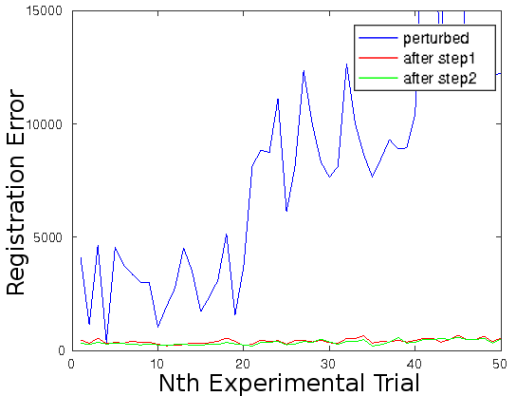


(b)

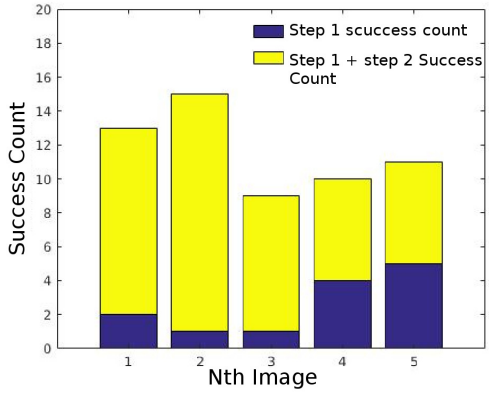
Fig. 7: 2D-2D registration success comparison

points on the image plane of any chosen SfM camera pose was estimated. The projected 2D coordinates corresponding to the ground truth pose, perturbed pose and corrected pose were all recorded. A registration was said to be successful if the mean-square-error between the ground truth and the corrected coordinates are below an empirical threshold value. This experiment was repeated 20 times for each dataset with both the test cases. Figure 7a shows the success and failure counts out of the total number of trials. Here, the success count indicates the number of trials where at least one of the test cases resulted in successful registration. Failure count is the number of trials when both the methods resulted in failure. Due to the randomness of the images and the perturbation trials, some trials were more difficult than others. Hence the failure trials where both the cases failed are ignored. Figure 7b shows the success split of the registration results. It shows the number of times when both the methods lead to success and the number of times when only one of them was successful. As can be seen from these results, using EPD inside RANSAC makes the trials successful more often than without using it.

Considering the overall 2D-2D registration procedure, the optimization step helps fine-tune the solution after initial coarse alignment. The results in figure 8 emphasize the benefits of the optimization step. For a single image, the registration error after the random perturbation, after the initial coarse



(a)



(b)

Fig. 8: Benefit of the optimization step

alignment and after the optimization step is shown in figure 8a. The EPD optimization reduces the registration error further in most cases. Graph 8b highlights this fact for all the datasets used in the experimental trials. It compares the number of times the feature match method followed by EPD optimization helped to obtain a lower registration error versus the number of times the feature match resulted in a lower value but the optimization step deviated from the solution. As can be seen in the bar graph, the additional optimization step leads to better image registration during most experimental trials. The datasets used for the registration experiment are shown in figure 9. A comparison of scan overlay on a SfM map while using the traditional ICP method and our heuristic approach is shown in figure 10.

V. Conclusion

In this paper, a new 3D-2D registration procedure is proposed for a 3D scan alignment application. Unlike earlier approaches that used brute force searches to initialize the optimization step, the proposed algorithm uses a feature based approach with a modified RANSAC outlier rejection scheme to enable better 2D-2D registration. A multi-view rigid 3D scan alignment step was also developed to ensure 3D scans were registered correctly by using both depth as well as image information. As a logical extension to this work in future, we

Fig. 9: Reference and moving images from the dataset

plan to use non-rigid alignment of the stereo-scan using SfM images and one that achieves a seamless overlay for generating more photo-realistic models.

References

[1] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon, "KinectFusion: Real-time Dense Surface Mapping and Tracking," in *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, Washington, DC, USA, 2011, ISMAR '11, pp. 127–136, IEEE Computer Society.

[2] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon, "KinectFusion: Real-time 3D Reconstruction and Interaction Using a Moving Depth Camera," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA, 2011, UIST '11, pp. 559–568, ACM.

[3] P. J. Besl and N. D. McKay, "A Method for Registration of 3-D Shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[4] Jakob Engel, Thomas Schöps, and Daniel Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

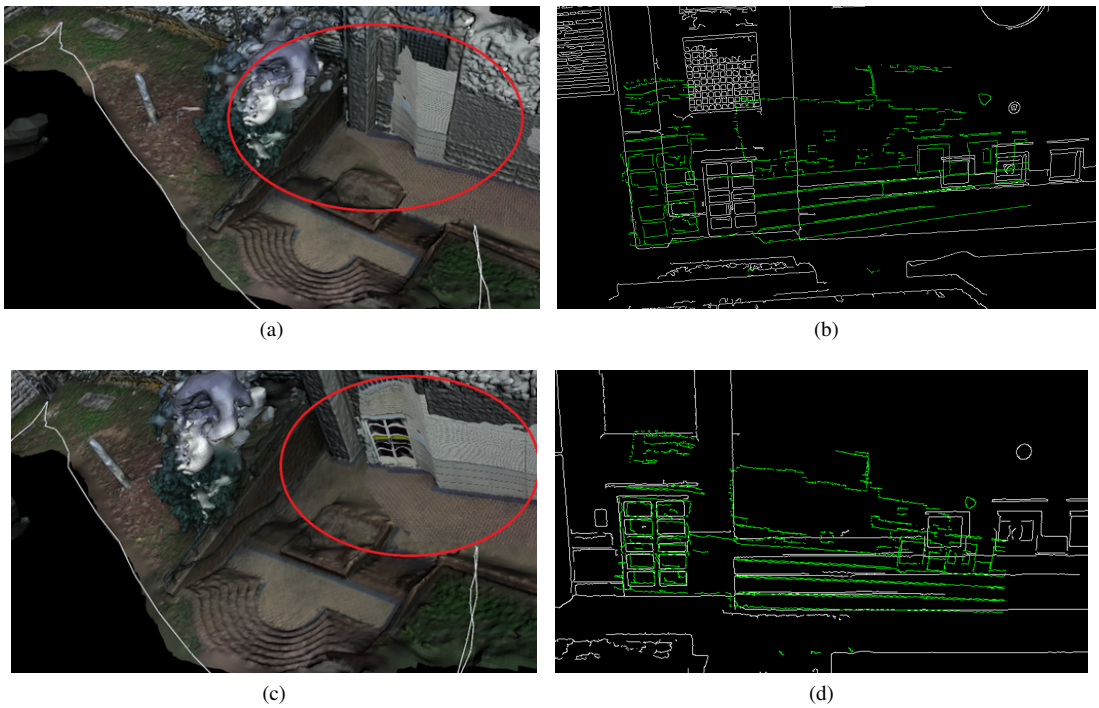
[5] J. Engel, J. Stückler, and D. Cremers, "Large-scale Direct SLAM with Stereo cameras," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, Sept 2015, pp. 1935–1942.

[6] Vladyslav Usenko, Jakob Engel, Jörg Stückler, and Daniel Cremers, "Direct Visual-Inertial Odometry with Stereo Cameras," .

[7] Shabnam Saadat, Mark R Pickering, Diana Perriman, Jennie M Scarvell, and Paul N Smith, "Fast and Robust Multi-Modal Image Registration for 3D Knee Kinematics," in *Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on*. IEEE, 2017, pp. 1–5.

[8] Deepak Rajamohan, Mark Pickering, and Matthew Garratt, "Using Edge Position Difference and Pixel Correlation for Aligning Stereo-Camera Generated 3D Scans," 12 2018, pp. 1–5.

[9] Mark R Pickering, "A New Similarity Measure for Multi-Modal Image Registration," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2273–2276.



(a) (b)

(c) (d)

Fig. 10: Scan overlay comparison. (a) ICP only alignment, (c) proposed method (b,d) corresponding edge overlap on SfM images

[10] David G Lowe, "Object Recognition from Local Scale-Invariant Features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Ieee, 1999, vol. 2, pp. 1150–1157.

[11] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up Robust Features (SURF)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[12] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2011, pp. 2564–2571.

[13] Martin A Fischler and Robert C Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[14] Md Nazmul Haque, Mark R Pickering, Moyuresh Biswas, Michael R Frater, Jennie M Scarvell, and Paul N Smith, "A Slice based Technique for Low-Complexity 3D/2D Registration of CT to Single Plane X-ray Fluoroscopy," in *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*. IEEE, 2012, pp. 1–6.

[15] Md Nazmul Haque, Mark R Pickering, Abdullah Al Muhit, Michael R Frater, Jennie M Scarvell, and Paul N Smith, "A Fast and Robust Technique for 3D-2D Registration of CT to Single Plane X-ray Fluoroscopy," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 2, no. 2, pp. 76–89, 2014.

[16] Gunilla Borgefors, "Distance Transformations in Digital Images," *Computer vision, graphics, and image processing*, vol. 34, no. 3, pp. 344–371, 1986.

[17] Gunilla Borgefors, "Hierarchical Chamfer Matching: A Parametric Edge Matching Algorithm," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 10, no. 6, pp. 849–865, 1988.

[18] Tal Hassner, Liav Assif, and Lior Wolf, "When standard RANSAC is not enough: Cross-media visual matching with hypothesis relevancy," *Machine Vision and Applications*, vol. 25, no. 4, pp. 971–983, 2014.