3D Reconstruction of Dynamic Vehicles using Sparse 3D-Laser-Scanner and 2D Image Fusion

Dennis Christie, Cansen Jiang, Danda Paudel, Cédric Demonceaux

3D Reconstruction of Dynamic Vehicles using Sparse 3D-Laser-Scanner and 2D Image Fusion

Dennis Christie*, Cansen Jiang†, Danda Paudel‡, and and Cédric Demonceaux§
*Gunadarma University, Indonesia. †‡§LE2I - Université de Bourgogne Franche-Comté, France.
Email: *dennis.aprilla@student.gunadarma.ac.id, Dennis_Christie@etu.u-bourgogne.fr,
†Cansen.Jiang@u-bourgogne.fr, ‡danda-pani.paudel@u-bourgogne.fr, ‡cedric.demonceaux@u-bourgogne.fr

***Abstract*—Map building becomes one of the most interesting research topic in computer vision field nowadays. To acquire accurate large 3D scene reconstructions, 3D laser scanners are recently developed and widely used. They produce accurate but sparse 3D point clouds of the environments. However, 3D reconstruction of rigidly moving objects along side with the large-scale 3D scene reconstruction is still lack of interest in many researches. To achieve a detailed object-level 3D reconstruction, a single scan of point cloud is insufficient due to their sparsity. For example, traditional Iterative Closest Point (ICP) registration technique or its variances are not accurate and robust enough to registered the point clouds, as they are easily trapped into the local minima. In this paper, we propose an 3-Point RANSAC with ICP refinement algorithm to build 3D reconstruction of rigidly moving objects, such as vehicles, using 2D-3D camera setup. Results show that the proposed algorithm can robustly and accurately registered the sparse 3D point cloud.**

***Index Terms*—3D Reconstruction, 2D camera, 3D camera, Registration, Point Cloud, ICP**

I. INTRODUCTION

A. Background

Map building becomes one of the most interesting research in computer vision field nowadays, due to its wide future applications such as localization, navigation, and autonomous driving vehicles. Simultaneously Localization and Mapping (SLAM) technique is widely used by researchers to build the 3D map of an environment. Multiple cameras and odometry sensors are involved to serve the system, which allow us to acquire 3D data of the scene to build the 3D map of the environment. A benchmark, provide by Geiger et al. [1], utilized 3D laser scanner as well as 2D cameras on the vehicle, which produces very accurate but sparse 3D measurement of the scene. SLAM approach can be used to build a 3D reconstruction for the static objects such as buildings or streets.

However, in the real world, there are usually not only static objects, but also moving objects in outdoor environments. It is very common that the outdoor environment is a dynamic scene, for instance, people are walking around, vehicles are running, cyclists are crossing the roads, which make the outdoor environment barely a completely static scene. When there are a large number of dynamic objects around, 3D map building by SLAM will not be accurate. 3D reconstruction in a dynamic scene is still an unsolved and challenging task.

A complete framework of static map and dynamic object reconstruction in outdoor environments was proposed by Jiang et al. [2], [3]. They succeeded to detect and segment the dynamic objects from the scene to reconstruct the static scene parts as well as the dynamic objects independently. The proposed static map building algorithm achieved very satisfactory performances on realistic outdoor environment.

This paper is a complementary of [2] with a smaller scope. We mainly focus on improving the reconstruction quality of the dynamic objects from a moving camera setup. As the output of the framework, multiple dynamic objects, such as vehicles, are obtained from different view ports due to the relative motions between the camera and the objects. In such kind of scenarios, reconstructing the moving objects from a moving camera setup is very challenging. To tackle this problem, the objective of this work is to build a 3D reconstruction of rigidly moving vehicle, using a 2D-3D calibrated camera setup.

B. Problem Definition

To reconstruct the rigidly moving objects from a moving camera setup, we propose to use 3D laser-scanner, such as Velodyne [4], due to its stability and accuracy in acquiring the 3D data. Yet, the laser scanner provides only sparse 3D point cloud, especially for the object that is far away from the camera. Moreover, the laser scanner does not give us texture information of the point cloud. We define a high quality 3D reconstruction as: object is reconstructed accurately, densely and textured. The dilemma comes when texture is obtained by using 2D cameras, while the 3D points are acquired from laser scanner.

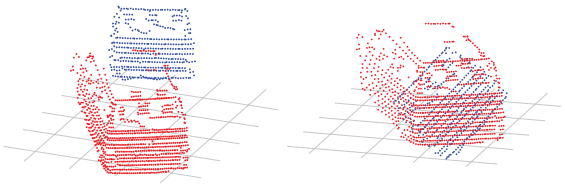


Fig. 1: Illustration of Poor 3D Geometric Structure Problem: left image, the red and blue point clouds are acquired from two different scans. The right image shows the failure registration result by using ICP algorithm due to the poor geometric structure of the point clouds.

A calibrated 2D-3D camera setup can overcome this dilemma, which means that there exist correspondences between 2D image points with 3D laser points. For instance, giving a mobile robot equipped with calibrated 2D and 3D cameras, one can reconstruct photometric high quality 3D

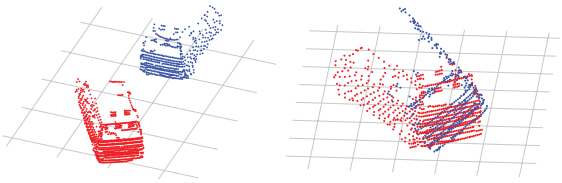


Fig. 2: Illustration of Small Partially Overlapping Problem: left image, the red and blue point clouds are acquired from two different scans. The right image shows the failed registration result using ICP algorithm.

scenes for static objects by registering a sequence of observations [2], [3]. Differently, registering the moving objects are more challenging due to their sparsity, small overlapping area, and non-uniform point density distribution of the moving objects. Two typical failure cases are shown in Fig. 1 and Fig. 2. The registrations using traditional ICP algorithm are failed due to the reasons of poor 3D geometry structure and small partially overlapping, respectively. This is due to traditional ICP algorithm only consider the mean square error (MSE) minimization between points of two point clouds. Minimum MSE, doesn't guarantee it converge into correct registration.

To overcome problem that is stated, we propose a series of registration algorithm, a 3D reconstruction of dynamic objects using 2D-3D camera setup attached to a moving vehicle, as the contribution of this paper.

II. LITERATURE REVIEW

Finding feature correspondences is a fundamental problem in computer vision applications, such as stereo vision [5], struct-from-motion [6], motion analysis [7], and especially point cloud registration [17]. In literature, there are many methods to match the features, such as statistical approach with similarity measurement [8], feature descriptor based matching [9]–[11].

One of the most robust descriptor based feature matching algorithm is Scale Invariant Feature Transform (SIFT) by Lowe [11]. SIFT is able to detect keypoints and build the scale invariant and orientation invariant descriptors. Although descriptor based feature matching is robust and efficient in rich-texture regions, features from low-texture area are easily wrongly matched. Thus, feature descriptor matching is generally reliable in matching highly distinctive features.

For feature matching between successive frames, KLT [12] feature tracking algorithm is applied to establish the dense correspondences [2]. The main advantage of OF feature tracking and matching comes from its ability to produce very dense feature matches unlike the sparse feature matches from feature descriptors. In addition, a cross validation can be applied to improve the tracking accuracy and robustness. Since image sequence (video) is involved in this work, optical flow is the optimal choice for our case.

One of the most challenging tasks in SLAM is the ego-motion, which contains the rotation and translation of the camera, estimation of the camera. Least Square Fitting is a non iterative registration algorithm firstly proposed by Arun et al. [13]. The advantage of using Least Square Fitting algorithm is its nature of efficiency. However, since using Least Square Fitting algorithm consider all of the matching pairs, it is very sensitive to outliers. Another technique to estimate rigid motion between two point clouds is the minimum 3-Point RANSAC [2] algorithm. This registration algorithm takes iteration to maximize number of inliers under the RANSAC framework [14]. The advantage of using 3-Point RANSAC is its nature to be robust to outliers.

Since the introduction of ICP by Chen and Medioni [15] and Besl and McKay [16], the algorithm was further developed into various type of ICP by many researchers to make it more robust. Rusinkiewicz and Levoy [17] nicely classified these variance of ICP into six main stages.

III. MATHEMATICAL MODEL

A. 3-Point RANSAC Point Cloud Registration

3-Point RANSAC algorithm can robustly and precisely estimate the rigid motion parameter $\mathbf{R}$ and $\mathbf{t}$ between two 3D point clouds in a linear manner. Let $\mathrm{X} = [x, y, z]^{\mathrm{T}}$ and $\mathrm{Y} = [x', y', z']^{\mathrm{T}}$ be two corresponding 3D points under rigid transformation, we have:

$$\mathrm{X} = \mathbf{R}\mathrm{Y} + \mathbf{t}, \tag{1}$$

where $\mathbf{R}$ is the 3×3 rotation matrix and $\mathbf{t}$ is the 3×1 translation matrix. Let g be the Gibbs representation of rotation matrix $\mathbf{R}$, we have $\mathrm{G} = [g]_{\times}$ is a 3×3 skew-symmetric matrix where $g = \mathbf{e} \tan \frac{\theta}{2}$ with $\mathbf{e} = [e_x, e_y, e_z]^{\mathrm{T}}$ is the Euler rotation axis and rotation angle θ. Applying the Cayley Trasformation, R can be represented as:

$$\mathrm{R} = (\mathbf{I}_3 + \mathrm{G})^{-1}(\mathbf{I}_3 - \mathrm{G}), \tag{2}$$

where $\mathbf{I}_3$ is a 3×3 identity matrix. Replacing Eq. 1 using Eq. 2 and multiplying $(\mathbf{I}_3 + \mathrm{G})$ on both sides, we have:

$$\mathrm{X} - \mathrm{Y} = -\mathrm{G}(\mathrm{X} + \mathrm{Y}) + \tilde{\mathbf{t}}, \tag{3}$$

where $\tilde{\mathbf{t}} = (\mathbf{I}_3 + \mathrm{G})\mathbf{t} = [\tilde{t}_x, \tilde{t}_y, \tilde{t}_z]^{\mathrm{T}}$. Eq. 3 is a linear system such that the parameters can be estimated using a Linear Least Square approximation.

To robustly find the good estimation, a Random Sample Consensus (RANSAC) [14] algorithm is adopted to maximize the number of inliers. Let $\mathcal{N}$ be the number of matching pairs that have the RMSE less than threshold τ, so called inliers, we aim to maximize the number of inliers by randomly select 3 sample matching pairs as candidates to estimate the transformation matrix:

$$\mathcal{N} = \underset{\mathbf{T}}{\mathrm{argmax}} \Xi_{i=1}^{n} \|\mathrm{X}_i - \mathbf{T}\mathrm{Y}_i\|_2 < \tau, \tag{4}$$

where $\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix}$ is the desired transformation matrix in homogeneous coordinate. n is the total number of matching pairs. Symbol $\|\cdot\|_2$ denotes the $l_2 - norm$. Notation $\Xi_{i=1}^{n}$

counts the number of pairs satisfying the criteria that the point-pair distance is less than threshold τ.

B. ICP Point Cloud Registration

Different from 3-Point RANSAC algorithm, the ICP algorithm establishes the matching pairs by considering the nearest point from the new model to the reference. Let $\mathbf{X} = \{\mathrm{X}_1, ..., \mathrm{X}_j\}$ be the reference point cloud, and $\mathbf{Y} = \{\mathrm{Y}_1, ..., \mathrm{Y}_k\}$ be the new model, a point to point ICP matching pairs can be established by:

$$\mathrm{X}_j \leftrightarrow \mathrm{Y}_k := \min\{\|\mathrm{X}_j - \mathbf{T}\mathrm{Y}_k\|_2\}, \tag{5}$$

with $j \in [1, ..., m], k \in [1, ..., K]$, where symbol $\leftrightarrow$ denotes the point to point matching property, and symbol $:=$ can be interpreted as *defined by*. m and K are the maximum index of the reference and new model point cloud, respectively. Eq. 5 builds the matching pairs according to the closest point distance from reference point cloud to the new model. Since the ICP has complete point-to-point matching between two point clouds, its performance is highly depending on the level of noise or number of inliers.

$$\mathcal{E}_{icp} = \underset{\hat{\mathbf{T}}}{\operatorname{argmin}} \frac{1}{m} \sum_{j=1}^{m} \|\mathrm{X}_j - \hat{\mathbf{T}}\mathrm{Y}_j\|_2, \tag{6}$$

where $\mathcal{E}_{icp}$ is the normalize root mean square error distance (RMSE) required to be minimized. $\hat{\mathbf{T}}$ is the refined transformation matrix after ICP.

IV. METHODOLOGY

A. Point Correspondences

To search for feature correspondences in 3D space, ideally, 3D feature descriptors [18] should be used. However, in practice, matching the 3D points using 3D descriptors is difficult to be maintained due to lack of robust 3D feature descriptors [2], [3]. In contrast, 2D image feature descriptors is much more robust and reliable in the case of 2D-2D matching. Therefore, to overcome this problem, we take the advantage of the 2D-3D camera setup to infer the 3D-3D feature correspondences based on the detected 2D-2D correspondences.

In details, the point correspondences are not established directly between 3D feature points, but we infer them indirectly by building the 2D feature matching pairs from their corresponding image frames. Once corresponding points between two images are obtained, we associate the 2D features with their corresponding 3D points by the 2D-3D correspondences from the calibrated 2D-3D setup. Figure 3 illustrates how the 3D-3D correspondences are established through the inferences of 2D-2D correspondences.

B. Feature Matching

As discussed about, the SIFT or other feature detectors can detect only sparse feature points, and those feature points can be easily lost tracked during a long sequence tracking. Moreover, since the moving objects in the scene are relatively small, there are very few features can be detected from the moving objects in most cases. On the contrary, Optical Flow feature matching can be used instead, which produces very dense feature matching between two consecutive frames. Remind that image sequence (video) is involved, OF feature is the optimal choice for our system. The main advantage of optical flow feature matching is that every pixel can be consider as a feature point, such that we have very dense features. Accordingly, features from moving objects are dense enough to be tracked and registered.

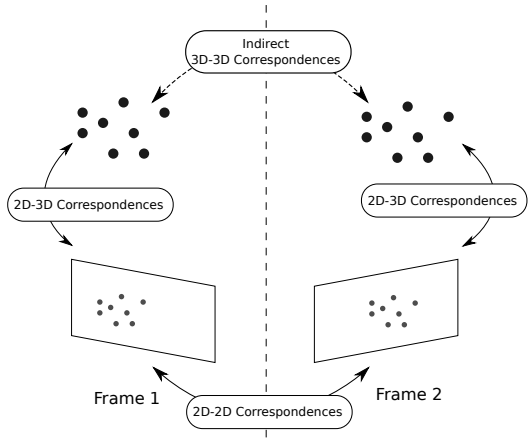
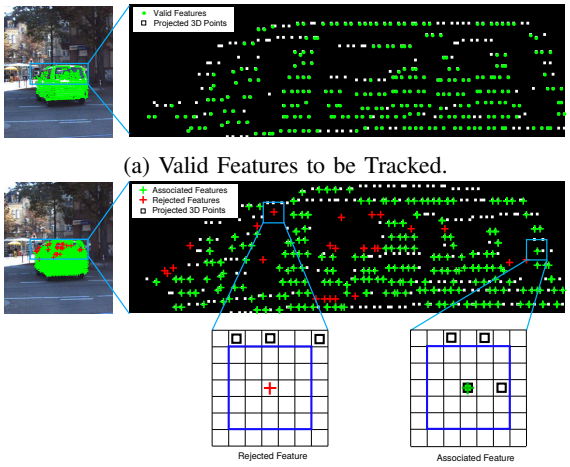


Fig. 3: Inferring 3D-3D correspondences from 2D-2D feature correspondences.

C. Feature Tracking

In practice, the point density of the 2D image is much denser than the 3D point cloud. A problem raises that not all the image points has 3D correspondences. Fig. 4 explicitly shows the problem that sparse 3D points corresponds to dense 2D image points.



(a) Valid Features to be Tracked.

(b) Features and 3D Projections Association in the Next Frame.

Fig. 4: Closest Feature Association Problem.

In this figure, the white dots in both sub-figures represents the projections from 3D points. The green dots in sub-figure (a) are the valid features after forward-backward OF

validation, while the crosses (labelled in green or red) in subfigure (b) represents the predicted location of the features using optical flow values in the next frame. In Fig. 4, the red crosses (lonely features) are rejected due to their having no 3D correspondences. Note, a correct association is defined by have a 3D projection within a 5×5 searching window.

D. 3D Registration Algorithm

1) Registration Pipeline: To achieve long term registration, we introduce a keyframe-based local sequence registration pipeline. Instead of taking a global reference frame, we divide long sequence with keyframes into local sequences., see Fig. 5.

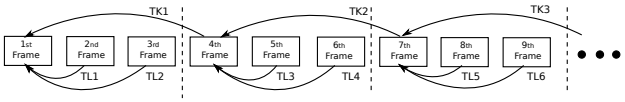


Fig. 5: Key-Frame based Registration Pipeline: TK stands for the transformation between two key frames, while TL represents the local transformation within the subsequence.

To detect the key frames automatically, a MSE threshold is defined such that new keyframes are chosen whenever the MSE of a local sequence is greater than the threshold.

2) Initialization and ICP Refinement: Sparsity and poor geometric structure (see Fig. 1) make ICP is not effective to register point clouds in our case, it will easily trapped in local minima. Even though 3-Point RANSAC can register the point clouds nicely, registration through transformation propagation leads to an error accumulation problem. It is already reduced by using pipeline in Fig. 5. However, to achieve a high quality registration, a refinement step is necessary. Therefore, combining 3-Point RANSAC as initialization and ICP for refinement can perform a very effective registration (see Fig. 8).

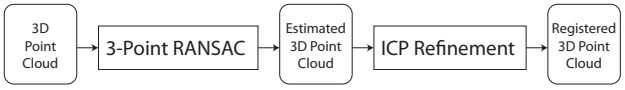


Fig. 6: Key-Frame based Registration Pipeline

3) Recognizing Part of Vehicle: ICP refinement doesn't guarantee correct registration in all case. It fails in registering two point clouds with small partial overlapping (see Fig. 2). In this case, the ICP tries to minimize the registration error by rotating and translating the model, which produces incorrect registration.

To avoid that situation, the registration should be performed only for set of points that share the same side of the object (Figure 7a). Normal vector is able to give that information. Thus, by using normal vector, we are able to obtain those set of point.

To differentiate the faces of rigid object, a plane fitting algorithm performed. A plane is defined as $z = ax + by + c$. By using 3 non-colinear points, we can define a plane. To robustnessly fit the plane, the RANSAC framework is applied. Figure 7a and 7b shows the fitted back-side planes and left-side planes, respectively. Getting the fitted planes, only the points close to the fitted planes will be considered to perform registration.

V. RESULT AND EVALUATIONS

To evaluate the performances of the proposed algorithm, experiments are conducted on realistic outdoor environment by using KITTI dataset [1]. The descriptions of the dataset are summarized in Table I. The software is developed using a computer with Intel Quad Core i7-2640M, 2.80GHz, 7.8GB Memory.

TABLE I: Dataset Profile

Name of Dataset		Van Dataset
Number of Frame		44
Number of Side		3
Transition Frame	Left-Back	1-30
	Right-Back	31-44
Number of Keyframe Produced		11

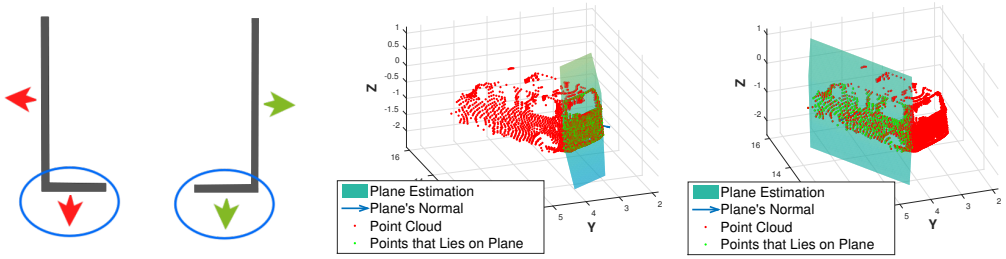
To evaluate the performances with and without registration refinement, Mean Square Error (MSE) is chosen to quantify the performance. 3D point cloud from first frame is considered as a global reference model. And the registration error is defined as MSE of corresponding 3D points from registered model to the global reference model.

Figure 8 shows that both ICP refinement with and without interpolation (red and blue plot) have registration failure in around frame 30-35. These frames is when small partial overlapping occurred (see Fig. 9). Also, a denser point cloud interpolation is applied to increase the stability of the ICP registration.Satisfactory result appear for refinement with the proposed algorithm with plane fitting (green plot). After dividing different parts of vehicle by considering the normal consensus of point cloud subset, registration performed accurately, even for small partially overlapping cases.

Figure 10 shows qualitative results of overall registration from three different viewports. 3D point clouds from multiple frames are registered very satisfactorily. Comparing to single scan point cloud by 3D laser scanner (see Fig. 1 and Fig. 2) the the registered 3D point cloud of vehicle have much denser points and richer geometry structure details.

VI. CONCLUSION AND FUTURE WORK

In extreme cases, Iterative Closest Point (ICP) or its variance is not able to perform registration correctly. It easily trapped in the local minima. Highly sparse 3D point cloud, small partial overlap, existence of noise and inliers, create a new challenge to perform a registration, especially a registration dedicated to long sequence. Feature point reduction and error accumulation is the main problem in long sequence 3D point cloud registration. To overcome all of the problem defined, we took advantages from a calibrated setup of 2D-3D camera and proposed a algorithm for long sequence registration. The algorithm conducted in realistic outdoor environment by using KITTI dataset [1]. From experiments, the proposed algorithm produced very satisfying results both quantitatively and qualitatively.



(a) Similar side. (b) Back side plane. (c) Left side plane.

Fig. 7: Different Part the Rigid Vehicle.

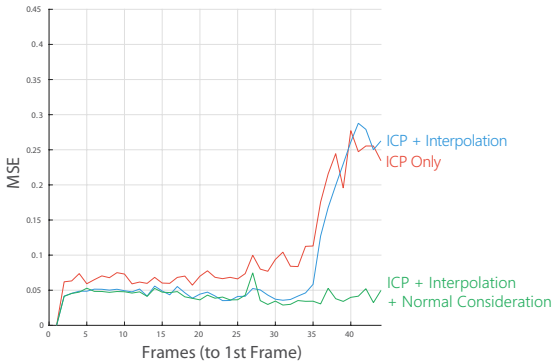
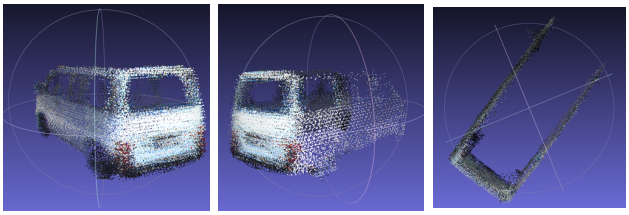


Fig. 8: Comparison of Different Registration Algorithms.

(a) Image from Frame 30. (b) 3D Model.

Fig. 9: Poor 3D Geometric Structure Frame.



(a) Back-left View. (b) Back-right View. (c) Top View.

Fig. 10: Registration Results using Our Algorithm.

However, the algorithm is still not sufficiently generic based on the assumption of plane fitting is adequate. To robustify the registration, robust estimation techniques, such as M-Estimator, can be applied to the algorithm. Furthermore, 2D images can also be employed as additional information to build up the Absolute Pose Estimation problem to estimate camera poses to register the point cloud.

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012.

[2] Cansen Jiang, Danda Pani Paudel, Yohan Fougerolle, David Fofi, and Cedric Demonceaux. Static-map and dynamic object reconstruction in outdoor scenes using 3d motion segmentation. pages 324–331, 2016.

[3] Danda Pani Paudel, Cédric Demonceaux, Adlane Habed, Pascal Vasseur, and In So Kweon. 2d-3d camera fusion for visual odometry in outdoor environments. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 157–162. IEEE, 2014.

[4] Frank Moosmann and Christoph Stiller. Velodyne slam. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 393–398. IEEE, 2011.

[5] David Marr, Tomaso Poggio, Ellen C Hildreth, and W Eric L Grimson. *A computational theory of human stereo vision*. Springer, 1991.

[6] Robert C Bolles, H Harlyn Baker, and David H Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[7] M Michela Del Viva and M Concetta Morrone. Motion analysis by feature tracking. *Vision research*, 38(22):3633–3653, 1998.

[8] Frank Candocia and Malek Adjouadi. A similarity measure for stereo feature matching. *Image Processing, IEEE Transactions on*, 6(10):1460–1464, 1997.

[9] Philippe Montesinos, Valérie Gouet, Rachid Deriche, and Danielle Pelé. Matching color uncalibrated images using differential invariants. *Image and Vision Computing*, 18(9):659–671, 2000.

[10] Florica Mindru, Tinne Tuytelaars, Luc Van Gool, and Theo Moons. Moment invariants for recognition under changing viewpoint and illumination. *Computer Vision and Image Understanding*, 94(1):3–27, 2004.

[11] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[12] Carlo Tomasi and Takeo Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[13] K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3-d point sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (5):698–700, 1987.

[14] 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*, 24(6):381–395, 1981.

[15] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

[16] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

[17] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001.

[18] Luís A Alexandre. 3d descriptors for object and category recognition: a comparative evaluation. In *Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vilamoura, Portugal*, volume 1, page 7. Citeseer, 2012.