Proceeding of the IEEE
International Conference on Robotics and Biomimetics (ROBIO)
Shenzhen, China, December 2013

Feature Selection and Pose Estimation From Known Planar Objects Using Monocular Vision

Shengdong Xu[1] and Ming Liu[2]

[1]ETH Zurich, Switzerland, Email: Samuel.xu1988@gmail.com
[2]The Hong Kong University of Science & Technology, Hong Kong, Email: liu.ming.prc@gmail

***Abstract*—In this paper, we develop a way to accurately and precisely estimate the pose of a calibrated camera with a single picture which includes a known planar object. For the proposed algorithm, we first use SURF detector for feature extraction and matching. Then, we use the information from known reference image to retrieve 3D point coordinates. Based on resulting 2D-2D correspondences and 3D coordinates, multiple-view geometry constraints are adopted to calculate the camera pose. Comparing with previous work, the proposed algorithm introduces an advanced feature selection algorithm, which eliminates pose ambiguity and improve the pose estimation result. The feature selection algorithm is based on the assumption that most 3D feature points should be coplanar. We conduct tests on traffic sign and evaluated the test results. The test results show that pose estimation with standard RANSAC turns out to be ambiguous occasionally. Conversely, the estimation with the proposed feature selection strategy leads to high robustness and accuracy.**

***Index Terms*—Multi-view geometry, Feature selection, Localization with monocular vision.**

I. Introduction

In computer vision, one of the classical problems is to localize accurately and precisely where a photo or video was taken. It has a broad range of applications, including consumer photography, robot localization, and autonomous navigation. It is the basis for visual SLAM problems and reconstructions. This paper focuses on exploring the approaches to accurately and precisely estimate the camera pose from known planar object using a single monocular camera.

The camera pose estimation problem can be formulated as follows: Estimate the attitude and position of the camera with respect to the world frame from feature point correspondences.

Several approaches for pose estimation are known in former works. Most of them work for arbitrary 3D points cloud, some extended to use line features [1], points and lines [2], and some also focus on coplanar points [3], however with limited validation. In most cases, RANSAC [4] algorithm is implemented in the pose estimation process, trying to classify outliers and inliers robustly and thus to improve the pose estimation result. However, due to the random feature selection strategy adopted in RANSAC, the estimated pose is vulnerable to ambiguity. For planar object, it is possible to develop more stringent criteria to select features for pose estimation and eliminate pose ambiguity.

A. Related work

Monocular vision based localization have been widely studied. The algebraic solution to PnP [5] problems provides the way to estimate camera pose based on n pairs of 2D-3D point correspondences. Most approaches use feature point detectors and matching themes to associate 2D points in the image plane with corresponding 3D points in 3D space. Lowe and Skrypnyk [6] proposed a classic system based on the SIFT descriptor for object localization, but computation expense imposes serious problems. More efficient detectors such as SURF [7] provides approximate performance with faster computation speed. These keypoint-based methods are widely applied in robotic applications such as visual homing based navigation [20]–[23] and scene recognition problems [24], [25]. However, one of the shortcoming of most feature descriptors is that they provide false matches. Considering those false matches which will influence the pose estimation result drastically, RANSAC algorithm is usually to be applied to classify outliers and inliers. Otherwise, random feature selection policy adopted in RANSAC algorithm will inevitably introduce ambiguity when used to estimate camera pose. Inspired by these observations, we propose an advanced point feature selection algorithm based on 3D analysis.

Several works on pose ambiguity can be found in literature. Denis Oberkampf [3] analyzed the pose ambiguity caused by large ratio of camera focal length over object depth. Lu [8] searched the way to eliminate pose ambiguity from video images. Haiwei Yang [9] considered pose ambiguity caused by limited number of feature points or special configuration of object points. In this paper, we explore the way to estimate camera pose robustly for localization with monocular vision, where feature points are relatively abundant and randomly distributed. We mainly attend to address the pose ambiguity problem which results from false feature points detected by error prone feature point detector such as SURF and random feature selection policy adopted in RANSAC algorithm.

Normal vector describes 3D geometric properties [26], [27] and can be used to assist point cloud segmentation efficiently. Point cloud clustering based on normal vectors has been applied in lots of previous work. Sagi Filin [10] used normal vector to segment point cloud obtained from laser scanning data and Rabbani T. [11] used local surface normals to segment point cloud. Further segmentation methods such as

978-1-4799-2744-9/13/$31.00 ©2013 IEEE

mutual information based [28] or Markov Process based [29] segmentation was also reported . This paper adopts normal vector voting strategy to segment discrete 3D point cloud. In this method, the neigborhood of each point act as voters to determine whether this point is coplanar with them or not.

II. Method

A. Algorithm in Outline

Figure 1 shows the proposed method of this paper in an outline.

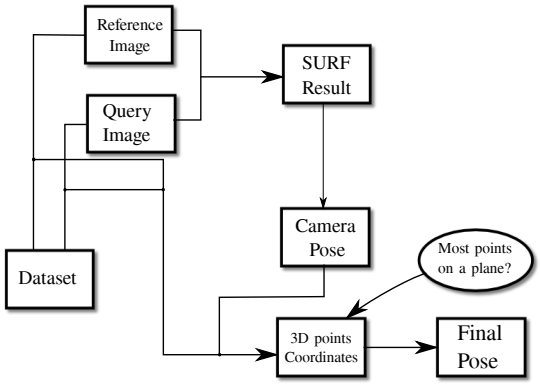

Fig. 1. Research method outline

Given a reference image and a test image which is currently taken by the device, we first extract SURF features and compute possible correspondences using nearest neighbor matching of the descriptors along with a distance ratio threshold. For faster nearest neighbor queries on the descriptors, we use the Fast Library for Approximate Nearest Neighbors FLANN [12], which is publicly available. RANSAC with adaptive termination criteria will be used to eliminate outliers. Camera pose is estimated by solving PnP problem [5], [17]–[19]. In this paper, the reference image was taken from a known pose, hence 3D coordinates of the points on the planar object can be retrieved.

Because the object is planar, once the 3D feature points are reconstructed, we can examine whether they are coplanar or not and thus to classify proper and improper feature points. After initial pose estimation, we then use triangulation method [14]–[16] to reconstruct 3D points based on the data from two images and the initial pose estimation result. If most reconstructed 3D feature points lie on the same plane, the initial pose estimation result should be correct. On the other hand, if lots of 3D feature points are not coplanar, we can conclude that the initial result is prone to be wrong. In this case, we will come back to the pose estimation step and conduct a new iteration to improve the result. These observations show that the key of a reliable feature selection algorithm is to define a robust criteria to guarantee that all points are coplanar.

Before getting into the proposed framework, we would like to give a short tutorial and compare several existing techniques for pose estimation, and indicate the necessity for feature selection.

B. Absolute pose estimation algorithm

For this algorithm, the coordinates of 3D points on the planar object is calculated in a straightforward manner. Once we know the pose of the camera center C and every projection ray direction (f_i) of each corresponding point on the plane, the 3D point (p_i) can be determined as the intersection point between projection ray and XY plane in world frame as shown in Figure 2.

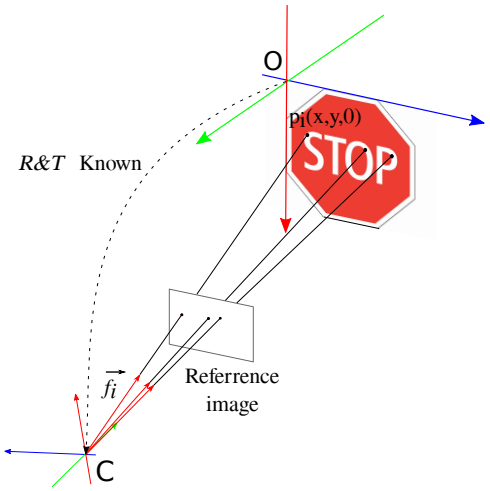

Fig. 2. Retrive 3D point coordinates based on reference image

The rotation and translation from world frame to the reference image camera frame can be therefore reconstructed. We assume the rotation matrix is R and the translation vector is $T = [T_x, T_y, T_z]^T$. For every unit projection ray vector $f_i = [f_{ix}, f_{iy}, f_{iz}]^T$, we define the scalar s as:

$$s = -\frac{T_z}{f_{iz}} \tag{1}$$

Then for every projection ray, its corresponding 3D points on the planar sign is:

$$p_i = s \ * \ f_i + T \tag{2}$$

Basically, for every projection ray, its corresponding 3D point on the planar object can be uniquely determined. Once the 3D coordinates of the points on the planar object is obtained, and we get image points pairs between reference image and query image, the following step is to compute absolute camera pose. Given the intrinsic parameters of a camera and a set of 2D-to-3D point correspondences, the problem of determining the absolute position and orientation of the camera is known as PnP problem, which provides us a method to estimate camera pose based on paired 2D image points and 3D points. The minimal number of points pairs needed to estimate the camera position and orientation is three, considering a fully calibrated camera.

Inevitably, the pose estimation result highly depends on the result of feature matching. We need at least five positive matches to recover unique pose. If the query image is blurred,

the result is less likely to be right. For those planar object which is highly symmetric and contains limited amount of texture, feature matching is prone to making even more mistakes.

C. 3D point coordinates reconstruction algorithm

Because the object is planar, if we reconstruct 3D points of the features, most of them should be coplanar. Here, we use triangulation method to reconstruct 3D points coordinates based on the information from both query image and reference image. The algorithm is shown as Figure 3.

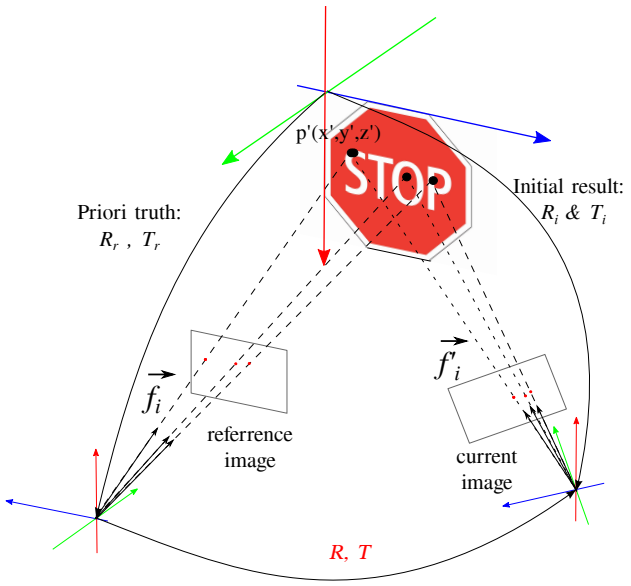

Fig. 3. Reconstruct 3D points using triangulation method

Using classical absolute pose estimation matrix, we can get an initial result (rotation matrix R_i and translation vector T_i from the world frame to query image camera frame). Given that the reference pose is known as a priori (rotation matrix R_r and translation vector T_r from world frame to reference image camera frame are known), the rotation matrix R and translation vector T from the reference image camera frame to query image camera frame can be easily computed:

$$R = R_r^{-1} \ * \ R_i \tag{3}$$

$$T = T_i - Tr \tag{4}$$

For every 3D feature point, its corresponding projection rays from two images can be obtained based on projective geometry. Once every point's corresponding projection ray ($\vec{f}_i$,$\vec{f'}_i$) is known, and the rotation matrix R and translation vector T from world frame to reference image camera frame are feasible to compute, by which we can reconstruct the 3D coordinate of every feature point using triangulation method.

D. Normal vector of feature points

In this section, we assume that the number of positive feature points is more than 9 [1]. For every 3D feature point, we choose its k ($k = 8$) nearest neighbors to vote whether it is coplanar or not. A plot in 3D space is shown as figure 4:

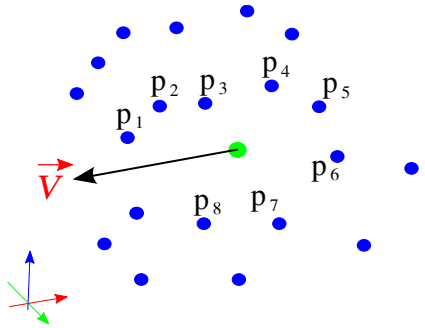

Fig. 4. Computation of normal vector for a feature point

For a feature point p, it has 8 nearest neighbor feature points $p_1, p_2, p_3 ... p_8$. Assuming that after 3D reconstruction, the coordinate of the point p is $[x, y, z]^T$ and its nearest neighbors' coordinate are $[x_i, y_i, z_i]^T$ for neighbour point p_i. Let 3×8 variation matrix H be:

$$[(p_1 - p) \ (p_2 - p) \ (p_3 - p) \ ... \ (p_8 - p)] \tag{5}$$

Then the 3×3 Covariance Matrix C is derived as:

$$C = H \ * \ H^T \tag{6}$$

The normal vector of point p is the unit eigenvector of the matrix C with the maximum corresponding eigenvalue. Every feature point obtained from the triangulation method will have a corresponding normal vector though the above steps. This can be shown as Figure 5.

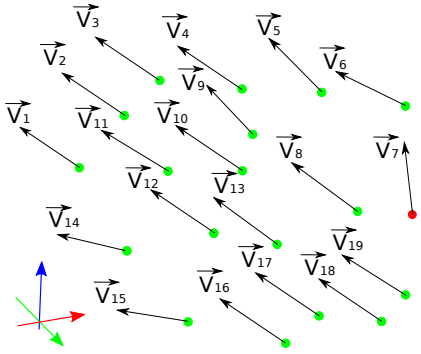

Fig. 5. Normal vector of every feature point

Therefore, coplanar points should have normal vectors that point to the same direction (or opposite). As the object is planar, if the initial result (rotation matrix R_i and translation

[1]It is related to the number of neighbours to be taken as reference. The number must be bigger than 3, in order to provide enough constraints on planar geometry.

vector T_i from the world frame to query image camera frame) is correct, most points should lie on the same plane and thus have normal vectors pointing to the same direction (or opposite), improper features should have normal vectors that point to an arbitrarily different direction. Thus, we can classify features according to their normal vectors. A typical example is sketched as the red point in Fig 5. The detailed algorithm is explain in the next subsection.

E. Feature selection process

The process is explained as follows. After calculation of normal vectors of each feature, we extract their median normal vector, and compute the dot product between each normal vector and the median normal vector which represents the cosine from each normal vector to the median vector. Because most feature points should be coplanar, the deviation should be small. And those feature points with relatively small normal vector median normal vector dot products is improper and thus should be erased in pose estimation process. The feature selection process can be shown in Figure 6.

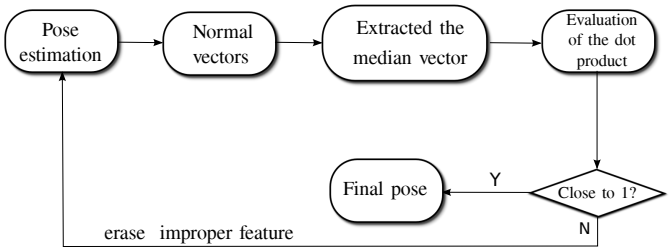

Fig. 6. Flow chart of feature selection process

After improper feature point erasion, most feature point that take part in the pose estimation procedure should be positive, thus the result would be accurate and robust.

III. EXPERIMENT

We carried out tests with the Vicon motion tracking system, which provides sub-millimeter precision on 3D ground truth. We took 15 images of a stop traffic sign from 15 different poses and chose one of them as reference image (the pose where we took this reference image is assumed to be known and thus can be directly used). What we want to do is to estimate the other 14 camera poses using the information from reference image and query image. Some of the test images have high pixel noise, and some was taken from distance or large angle. The estimated camera poses can be expressed by a 3×4 matrix which denotes the rotation (the first tree columns) and translation (the last column) from the object frame. We estimate 14 camera poses separately with standard RANSAC algorithm and with feature selection strategy developed in this paper. The comparative results are evaluated afterwards.

The test method is illustrated in Figure 7. First, we use standard RANSAC algorithm to calculate the camera pose. When the query image is taken quite far away from the reference image, there might be only a few matched points after SURF, and the result might be vulnerable to ambiguity.

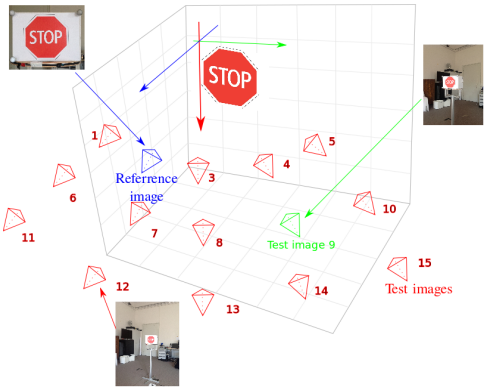

Fig. 7. Experimental Design

Test results of 100 times experiment using the test image 9 in Figure 7 is shown as Figure 8.

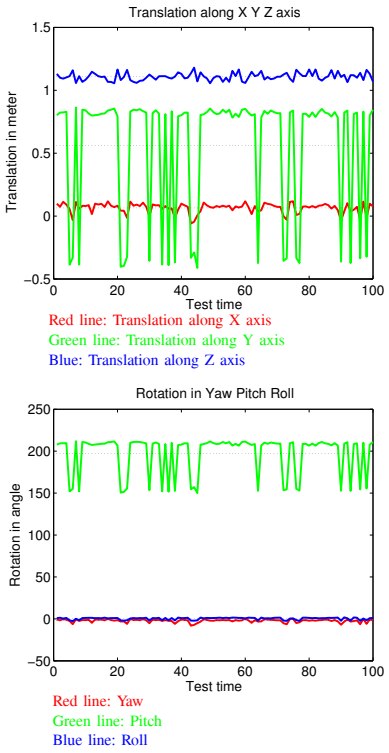

Fig. 8. Test result before feature selection

The ground truth of the test is:

$$T = [T_x \ T_y \ T_z]^T = [0.140 \ 0.7593 \ 1.133]^T$$

$$R = [roll \ yaw \ pitch]^T = [-1.433 \ 3.213 \ 203.495]^T$$

In this case, about 20 out of 100 test times, the estimated result jumps to a totally wrong value, which is far from satisfactory.

When the proposed algorithm is implemented, theoretically, most feature points will be coplanar after feature selection and thus the dot product between each normal vector and their median vector should be close to 1 or -1.

We classify proper and improper feature points according to the absolute value of the dot product. Therefore, after feature selection, the dot products (absolute value) distribution should lie close to one with small variation. The comparative results of the dot products (absolute value) before and after feature selection are shown in Figure 9. We see that after erasion of improper feature points, dot products (absolute value) distribution lies closer to 1 with smaller variation.

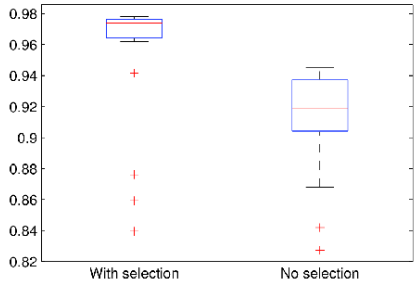

Fig. 9. Dot products distribution

The comparison between the pose estimation result with feature selection and that without is shown in Figure 10. Comparing with the result of standard RANSAC in Fig. 8, we see that the false pose estimations have dropped from 20 to 6. It means the pose estimation precision rises from 80% to 94%.

Moreover, from the test result, we see that the proposed feature selection strategy can not only eliminate pose ambiguity, but also improve pose estimation result of query images with high pixel noise. The comparison between the test result of query image in the presence of high pixel noise with feature selection and that without are shown in Figure 11.

The ground truth of the test is:

$$T = [T_x \;\; T_y \;\; T_z]^T = [0.160 \;\; -0.185 \;\; 2.113]^T$$

As we can see, in the presence of high pixel noise, feature selection strategy developed in this paper can make the result more robust and accurate.

IV. Disscusion

From above experiment result and further analysis, we can see that if there are serious pixel noise, the initial pose estimation result will not be reliable. Especially, if the number of feature points is limited and only a part of them are proper feature points, pose ambiguity imposes a serious problem. After looking carefully into the pose estimation process, the reason can be easily found out. SURF result provides large amount of improper matches which will influence the pose estimation process drastically. In order to classify inliers and outliers, RANSAC algorithm needs to be applied for automatic processing of point cloud with the aim of 3D building modeling. But model fitting approaches will possibly provide spurious segmentation result when dealing with different point

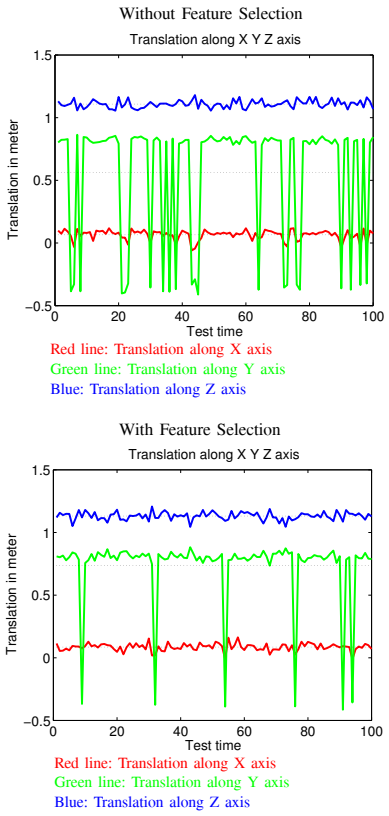

Fig. 10. Test result comparison

cloud sources. As a result, pose ambiguity will occure if we fail to find an efficient way to segment point cloud in a proper manner. Advanced feature points selection strategy can assist accurate pose estimation and improve final result. The feature selection strategy developed in this paper proves to be reliable and robust because normal vector of each feature point provides a safe indicator of the properness of each feature point. As shown in the experiment and its result, after elaborate feature points selection process developed in this paper, fewer improper feature points will take part in the pose estimation process, the final result is more likely to be robust, accurate and precise.

V. Conclusion

In this paper, we proposed an advanced feature selection algorithm and showed its application in pose estimation problems. In general, pose estimation results by only using standard RANSAC are fair. It means that if the query image is taken near to where the reference image was taken, the number of feature point is high enough to get satisfactory pose estimation result. However, if the resolution of the query image is low, the pixel noise might be too serious to conduct pose estimation. In this case, the pose estimation result is not robust and only the translation along focal axis is not sensitive to noise. Moreover, if the number of feature points is limited and contains improper ones, the problem of pose ambiguity may occur because RANSAC algorithm initially randomly

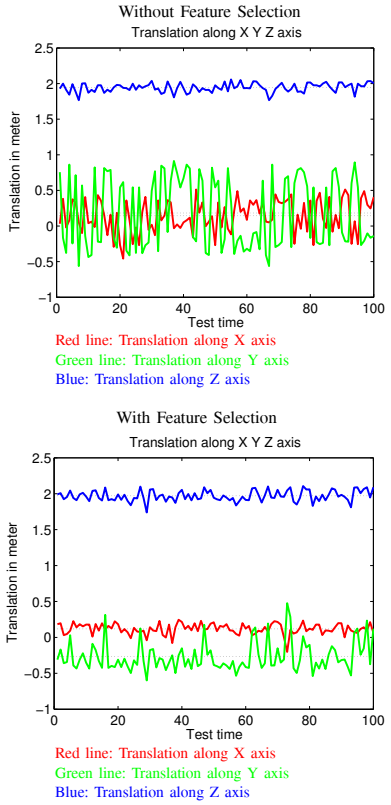

Fig. 11. Test result comparison of image in the presence of pixel noise

chooses feature points to fit model. Pose estimation with feature selection strategy developed in this paper addressed above problems and improved the estimation result in a great deal. The result is highly robust even with presence of high pixel noise or limited number of proper feature points.

Real world applications may include but not limit to: localize smart phone users in large indoor environment though taking a photo of a known planar object.

REFERENCES

[1] MING LIU, BEKIR TUFAN ALPER AND ROLAND SIEGWART: *An adaptive descriptor for uncalibrated omnidirectional images - towards scene reconstruction by trifocal tensor*. IEEE International Conference on Robotics and Automation (ICRA), 2013.

[2] A.ANSAR AND K.DANIILIDIS: *Linear pose estimation from points or lines*. European Conference on Computer Vision, A.H. er al. Ed.,vol.4 Copenhagen, Denmark: Srpinger,May 2002,pp.282-296

[3] D. OBERKAMPF: *Iterative pose estimation using coplanar points*. IEEE on Computer Vision and Pattern Recognition, pp.626-627 1993.

[4] MARTIN A. FISCHLER, ROBERT C. BOLLES: *Random Sample Consensus: A Paradigm for Model Fitting with Apphcatlons to Image Analysis and Automated Cartography*. Communications of the ACM, vol.24, pp.381-395, 1981.

[5] KNEIP, L.,SCARAMUZZA, D.,SIEGWART, R.: *A Novel Parametrization of the Perspective-Three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2969-2976, 2011.

[6] SKRYPNYK, I., LOWE, D. G.: *Scene modelling, recognition and tracking with invariant image features*. In International symposium on mixed and augmented reality (pp. 110119). Arlington,VA.

[7] H.BAY,T.TUYTELAARS, AND L.V. GOOL: *Surf: Speeded up robust features*. Proc. European Conf.on Computer Vision, pp. 404-407, 2006.

[8] LU, C.-P., HAGER, G. D., MJOLSNESS,: *Fast and globally convergent pose estimation from video images*. IEEE Transactions on Pattern Analysis and Machine Intelligence,22(6), 610622. 2000.

[9] HAIWEI YANG: *A Robust Pose Estimation Method for Nearly Coplanar Points*. International Workshop on Nonlinear Circuits, Communications and Signal Processing NCSP'12. 2012.

[10] FILIN, S. AND PFEIFER,: *Segmentation of airborne laser scanning data using a slope adaptive neighbourhood*. ISPRS Journal of Photogrammetry and Remote Sensing, 60(2), pp. 71-80. 2006.

[11] RABBANI, T., VAN DEN HEUVEL, F. A. AND VOSSELMAN, G.: *Segmentation of point clouds using smoothness constraint*. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 36(5), pp. 248253.

[12] MUJA, M.: *Flann, fast library for approximate nearest neighbors*. http://mloss.org/software/view/143/

[13] R.HARTLEY A.ZISSERMAN: *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[14] LONGUET-HIGGINS, H.: *A computer algorithm for reconstructing a scene from two projections*. Nature 1981.

[15] H.-Y. SHUM, R. SZELISKI, S. BAKER, M. HAN, P. ANANDAN: *A Method for Interactive 3D Reconstruction of Piecewise Planar Objects from Single Images*. BMVC, pp. 265-274, September 1999.

[16] AMNON SHASHUA: *Projectivedepth: A geometric invariant for 3D reconstruction from two perspective/or thographic views and forvisual recognition*. In Proc. International Conference on Computer Vision, pages 583-590, 1993.

[17] R. HORAUD, B. CONIO, AND O. LEBOULLEUX: *An analytic solution for the perspective 4-point problem*. Computer Vision, Graphics, and Image Processing, 47:32-44, 1989.

[18] D. DE MENTHON AND L. DAVIS: *Exact and approximate solutions of the perspective-three-point problem*. IEEE on Pattern Analysis and Machine Intelligence,14(11):1100-1105, 1992.

[19] X. GAO, X. HOU, J. TANG, AND H. CHENG: *Complete solution classification for the perspective-three-point problem*. IEEE on Pattern Analysis and Machine Intelligence,25(8):930-943, 2003.

[20] MING LIU, CEDRIC PRADALIER, ROLAND SIEGWART: *Visual Homing from Scale with an Uncalibrated Omnidirectional Camera*. IEEE Transactions on Robotics, 2013. DOI: 10.1109/TRO.2013.2272251

[21] MING LIU, CEDRIC PRADALIER, FRANCOIS POMERLEAU, ROLAND SIEGWART: *Scale-only Visual Homing from an Omnidirectional Camera*. IEEE International Conference on Robotics and Automation (ICRA), 2012.

[22] MING LIU, CEDRIC PRADALIER, FRANCOIS POMERLEAU, ROLAND SIEGWART: *The Role of Homing in Visual Topological Navigation*. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012.

[23] MING LIU, CEDRIC PRADALIER, ROLAND SIEGWART AND QIJUN CHEN: *A Bearing-only 2D/3D-homing method under a visual servoing framework*. IEEE International Conference on Robotics and Automation (ICRA), 2010

[24] MING LIU, ROLAND SIEGWART: *Topological mapping and scene recognition with lightweight color descriptors for omnidirectional camera*. IEEE Transactions on Robotics, 2013. DOI: 10.1109/TRO.2013.2272250

[25] MING LIU, DAVIDE SCARAMUZZA, CEDRIC PRADALIER, ROLAND SIEGWART AND QIJUN CHEN: *Scene Recognition with Omnidirectional Vision for Topological Map using Lightweight Adaptive Descriptors*. International Conference on Intelligent Robots and Systems (IROS), 2009

[26] MING LIU, FRANCOIS POMERLEAU, FRANCIS COLAS AND ROLAND SIEGWART: *Normal Estimation for Pointcloud using GPU based Sparse Tensor Voting*. IEEE International Conference on Robotics and Biomimetics (ROBIO), 2012

[27] MING LIU, ROLAND SIEGWART: *Information Theory based Validation for Point-cloud Segmentation aided by Tensor Voting*. IEEE International Conference on Information and Automation (ICIA), 2013

[28] MING LIU, FRANCIS COLAS AND ROLAND SIEGWART: *Regional topological segmentation based on Mutual Information Graphs*. IEEE International Conference on Robotics and Automation (ICRA), 2011

[29] MING LIU, FRANCIS COLAS, FRANCOIS POMERLEAU, ROLAND SIEGWART: *A Markov Semi-Supervised Clustering Approach and Its Application in Topological Map Extraction*. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012