PEGUS: An Image-Based Robust Pose Estimation Method

S. S. Mehta†, P. Barooah, W. E. Dixon
Department of Mechanical and Aerospace Engineering,
University of Florida,
Gainesville, FL-32611
Email: {siddhart,pbarooah,wdixon}@ufl.edu

E. L. Pasiliao, J. W. Curtis
Air Force Research Laboratory,
Munitions Directorate,
Eglin AFB, FL-32542
Email: {eduardo.pasiliao,jess.curtis}@eglin.af.mil

***Abstract*—In this paper, a robust pose (i.e., position and orientation) estimation algorithm using two-views captured by a calibrated monocular camera is presented. A collection of pose hypotheses is obtained when more than the minimum number of feature points required to uniquely identify a pose are available in both the images. The pose hypotheses - unit quaternion and unit translation vectors - lie on the $\mathbb{S}^3$ and $\mathbb{S}^2$ manifolds in the Euclidean 4-space and 3-space, respectively. Probability density function (pdf) of the rotation and translation pose hypotheses is evaluated by griding the unit spheres where a robust, coarse pose estimate is identified at the mode of the pdf. Further, a "refining" pdf of the geodesic distance from the coarse pose estimate is constructed for the hypotheses within a grid containing the coarse estimate. A refined pose estimate is obtained by averaging the low-noise hypotheses in the neighborhood of the mode of refining pdf. Pose estimation results of the proposed method are compared with RANSAC and nonlinear mean-shift (NMS) algorithms using the Oxford Corridor sequence and the robustness to feature outliers, image noise rejection, and scalability to number of features is analyzed using the synthetic data experiments. Processing time comparison with the RANSAC and NMS algorithms indicate that the deterministic time requirement of the proposed and NMS algorithms is amenable to a variety of visual servo control applications.**

***Keywords*-robust pose estimation, two-view geometry, outlier rejection**

I. Introduction

MOTIVATED by practical applications such as autonomous guidance, navigation and control (GNC), and intelligence, surveillance, target acquisition and reconnaissance (ISTAR) various image-based techniques have been developed including visual servo control, visual odometry, and structure from motion. The underlying framework of these methods is based on an estimate of the relative pose (i.e., position and orientation) between the two-views obtained by an imaging device. For a monocular camera, the rotation and direction of translation are estimated, whereas in case of a stereo-vision system, the rotation and Euclidean translation vectors are estimated.

The physical and geometric features of an object or scene such as points, lines, and circles are typically used to determine the relative pose between the two viewpoints of a moving camera or the pose of a moving object with respect to a stationary camera (see [1]–[5] and references therein). In this paper, we focus on the problem of estimating the relative pose of a camera using point features as the camera moves between the two viewpoints. Often the feature point correspondence between the two images is provided by a feature tracking algorithm, such as the KLT tracker [6], SIFT [7] and SURF [8]. We assume that the feature correspondence or "loosely speaking" the matching problem is solved and a set of point correspondences is available between the two-views. The assumption merely requires a set of point correspondences between the views but does not make any assumption regarding the outliers or false matches present in the correspondences. Given a minimal set of point correspondence, the relative pose can be estimated using a number of algorithms, e.g., the eight point algorithm [9], seven point algorithm [10], five point algorithm [11]. However, point correspondences returned by a feature tracker invariably contain gross mismatches or large errors in the feature point locations, which are commonly referred to as *outliers*. The central issue in pose estimation is devising a class of robust estimators that can reject such outliers. Various optimization schemes such as the iterative weighted least-squares [1], M-estimators [12], [13], and least-median of squares (LMedS) [12]–[15] gained significant interest but suffer from the drawbacks of the optimization methods. Kumar et al. [12] found experimentally that the M-estimators are susceptible to initial estimates and demonstrate a low breakdown point. While LMedS-based methods can achieve a higher breakdown point of 0.5 [14]–[17] additional measures, e.g., weighted least-squares, may be required due to poor efficiency in the presence of Gaussian noise [13], [17]. The most popular approach to robust pose estimation problem has been the hypothesize-and-test methods, such as RANSAC [18] and its variants: MLESAC [19], PROSAC [20], GOODSAC [21], pre-emptive RANSAC [22], randomized RANSAC [23], [24], SCRAMSAC [25], etc. Wherein the hypotheses are generated by random, *a prior* assessment driven [21] or online adaptive [20] selection of the minimal set of feature point correspondences required to generate a pose hypoth-

†Corresponding author, ph. +1(850)833-9350 x227
This research is supported in part by the US Air Force, Eglin AFB, grant FA8651-08-D-0108/025.

978-0-7695-4683-4/12 $26.00 © 2012 IEEE
DOI 10.1109/CRV.2012.18

esis. Each hypothesis is scored based on the number of feature points in both the views that are well-explained by it and a hypothesis with the best score is declared as the desired estimate, and the corresponding feature points are classified as "inliers". Often the pose estimation is followed by pose refinement where the least squares optimal solution is computed using the obtained set of inliers. Most of the extensions to basic RANSAC scheme focus on reducing the computation time, since generating a large number of hypotheses (which is required to obtain a good estimate with high probability) and scoring them is computationally expensive. Due to the hypothesize-and-test framework, the solution as well as computation time of RANSAC are inherently non-deterministic, i.e., for the given set of point correspondences the pose estimate and computation time may vary between the different runs. Further, RANSAC and other hypothesize-and-test methods choose only one of the many hypotheses that are or can be generated; all other hypotheses are ignored even those that may be quite close to the true pose. Each hypothesis can be considered as a noisy "measurement" of the relative pose that is to be estimated. In principle, one should be able to average the measurements in the vicinity of true pose in an appropriate sense to compute a more accurate estimate than any of the individual measurements (i.e., hypotheses). Clustering-based methods [26], [27] follow the above principle by generating a large number of pose hypotheses and identifying a pose estimate in the clustered space, e.g., the $\mathbb{SE}(3)$ manifold.

In this paper, we propose a robust pose estimation algorithm based on the clustering principle using multiple pose hypotheses generated from the feature point correspondences between the two views. There are two challenges that impede the development. First, many of the pose hypotheses will be corrupted by outliers and will show poor accuracy. Including these corrupted measurements in the averaging process may lead to little improvement, if any. The second difficulty is that since a pose is not a member of vector space, it is not clear how to average multiple noisy pose measurements.

To address these challenges, we estimate the rotation and (unit) translation in a decoupled manner. The rotation hypotheses are expressed as unit quaternions, which lie on a 3-sphere (i.e., a unit sphere in Euclidean 4-space). The probability density function (pdf) of unit quaternions on 3-sphere is obtained, wherein the dominant cluster corresponding to the mode of pdf gives rise to a coarse pose estimate. A "refining" pdf of the geodesic distance from the coarse pose estimate is constructed for the hypotheses within a grid containing the coarse estimate. The "low-noise" rotation hypotheses identified within a small geodesic distance of the mode of refining pdf are averaged according to [28] to produce a refined rotation estimate. Estimating the unit translation proceeds in an identical manner, except now the data lies on a unit 2-sphere in Euclidean 3-space. When a Euclidean translation vector (direction as well as magnitude) is available, say from a stereo camera, the mode estimation and averaging is simpler since the data lies in a vector space. Because of the role played by gridding of the unit sphere in 3 or 4 dimensions, the proposed algorithm is called the Pose Estimation by Gridding of Unit Spheres (PEGUS).

In contrast to hypothesize-and-test methods wherein the objective is to determine the largest inlier set, the proposed algorithm averages the information from a number of hypotheses that are likely to be close to the true pose. As a result, it comes up with a more accurate estimate than RANSAC-type methods. Our algorithm has certain similarities with the non-linear mean shift algorithm proposed in [27]; the similarities and differences between the two are discussed in Section II. Another advantage of the PEGUS algorithm is that it does not involve any iterative search, so that the time required for its execution is highly predictable making PEGUS suitable for various closed-loop visual servo control applications.

II. Related work

There are certain similarities between our approach and the non-linear mean shift algorithm by Subbarao *et. al.* [27], in which a set of generated hypotheses are used to construct a kernel-based estimate of the pdf of the pose hypothesis in $\mathbb{SE}(3)$. A non-linear version of the mean-shift algorithm is then used to iteratively search for the mode of pdf starting from an arbitrary initial condition. The identified mode is declared the pose estimate. Since all the hypotheses used to construct the pdf contribute to the mode, and the mode may not coincide with any of the hypotheses, the resulting estimate can be thought of as an average of the hypotheses, though the averaging is of an implicit nature. In short, the approach in the proposed PEGUS algorithm as well as that in [27] treat pose estimation as a clustering problem. Both construct estimates of the probability density (or mass) function from a set of generated hypotheses and returns an averaged hypothesis as the pose estimate rather than a single hypothesis from those generated.

Despite the similarities between the two approaches, there are significant differences between the proposed PEGUS algorithm and the non-linear mean shift algorithm of [27]. First, the PEGUS algorithm is more robust to multi-modal densities of the generated hypotheses than the mean shift method. In the presence of multi-modal pose distribution, the iterative search involved in the mean shift algorithm may converge to a local maxima depending on the initial condition. In contrast, since we construct a histogram-based estimate of the pmf (probability mass function) of the hypotheses locating the global mode is trivial even with multi-modal densities. The pmf of the rotation hypotheses is constructed by gridding a unit sphere in 4 dimensions on which the unit quaternion representations of the corresponding rotations lie. The same approach is applicable to unit translations where gridding is done on a unit sphere in

3 dimensions. If both the magnitude and direction of translation can be estimated then the histogram is constructed by dividing a region of $\mathbb{R}^3$ into a number of equal volume cells.

The second major difference is that the non-linear mean shift algorithm returns the mode as the estimate, whereas the proposed method uses the mode only to identify a set of hypotheses that are likely to be close to the true pose, i.e., to obtain the rough pose estimate. These "low-noise" hypotheses are then explicitly averaged in an appropriate manner to construct the final refined pose estimate. In addition, the proposed method does not involve iterative computation, whereas the mean-shift algorithm requires an iterative search for the mode. On the other hand, the non-linear mean-shift algorithm is applicable to a wide variety of estimation problems in which data lies on Riemannian manifolds, whereas the proposed method is only applicable to problems in which the data lies on spherical surfaces or real coordinate spaces.

III. Problem statement and approach

The objective is to develop a robust pose estimation algorithm using two images captured by a monocular camera (or four images if a pair of cameras are used) and without the knowledge of the scene. Let R denote the *true rotation* between two views and t be the *true translation*. The translation can be a *unit* translation if the scale information is not available.

If there are M pairs of feature points between two views captured by the camera and the minimal number of feature point pairs required to generate a pose hypothesis are P, then the total number of pose hypotheses that can be computed is $N_{\max} := \binom{M}{P}$. We first generate n such hypotheses, where n is typically much smaller than $N_{\max}$. Each pair in the generated rotation and translation hypotheses is a "noisy measurement" of the true rotation R and true (unit) translation t, respectively. Some of these measurements, i.e., hypotheses, suffer from a large inaccuracy. Our approach is to select a subset of "low-noise" hypotheses from the set of all possible hypotheses so that they are close to the true rotation and translation. The low-noise hypotheses are then appropriately averaged to compute a pose estimate.

To facilitate extraction of the low-noise hypotheses, each rotation hypothesis is expressed in terms of its unit-quaternion representation. Since the unit quaternions q and $-q$ represent the same rotation, we ensure that the unit-quaternion representation of a rotation hypothesis has the first component positive, i.e., if $q = q_r + iq_1 + jq_2 + kq_3$ then $q_r > 0$. A unit quaternion representation of a rotation matrix can now be thought of as a unit-norm vector in $\mathbb{R}^4$ whose first component is positive. That is, it lies on the "top" half of the 3-sphere $\mathbb{S}^3$. The d-sphere $\mathbb{S}^d$ is defined as

$$\mathbb{S}^d := \{x = [x_1, \ldots, x_{d+1}]^T \in \mathbb{R}^{d+1} | \ \|x\| = 1\} \quad (1)$$

where $\|\cdot\|$ denotes the Euclidean norm. Similarly, we define

$$\mathbb{S}^{d+} = \{x \in \mathbb{R}^{d+1} | \ \|x\| = 1, x_1 \geq 0\}. \quad (2)$$

Therefore, each rotation hypothesis is an element of $\mathbb{S}^{3+}$. Similarly, each hypothesis of the unit translation is an element of $\mathbb{S}^2$. If scale information is available, translation hypotheses are elements of $\mathbb{R}^3$ instead of $\mathbb{S}^2$.

Since each rotation hypothesis is a noisy measurement of the true rotation, the rotation hypotheses can be thought of as realizations of a random variable whose distribution is defined over the half-sphere $\mathbb{S}^{3+}$. By dividing the surface of the sphere $\mathbb{S}^3$ and counting the number of rotation hypotheses (rather, their unit-quaternion representations), we can estimate the pmf of rotation random variable. The mode of the pmf is a point in the bin that has the largest number of unit-quaternions. A subset of these quaternions that are within a predetermined geodesic distance of the mode is selected, and then averaged in an appropriate manner to obtain the desired rotation estimate. Estimation of translations proceed in a similar manner. The algorithm is described in detail in the following section.

IV. Proposed algorithm

A. Rotation estimation

Step 1: Hypotheses generation engine: Not all of the possible pose hypotheses are computed. Instead, we use a sampling with replacement strategy to generate a number of hypotheses that have small "correlation" among one another. The number of such hypotheses to be generated, n, is a design parameter that has to be specified *a priori*. The sampling strategy consists of selecting the first feature point pair at random from the M pairs, then selecting the second pair from the remaining $M-1$ pairs, and so on until the P-th pair is selected. These P pairs of point correspondence are used to generate a hypothesis. This sampling procedure is repeated n times to generate n hypotheses, which are denoted by q_i, t_i, where q_i is a unit-quaternion and t_i is a translation vector (unit-norm or otherwise), for $i = 1, \ldots, n$. The set of these n rotation hypotheses is denoted by S_q, and the set of translation hypotheses is denoted by S_t.

The reason for not computing all the possible hypotheses is that total number of possible pose hypotheses, $N_{\max}$ is typically extremely large, since $N_{\max} = \binom{M}{P}$, where M is the number of point correspondence and P is the minimal number needed to generate a hypothesis. For example, even a small value of M, e.g., 21, with $P = 8$ yields $N_{\max} = 203490$. Processing such a large number of hypotheses is computationally infeasible, especially since the pose hypothesis generation is computationally intensive. In addition, processing all N_{max} hypotheses is not necessary since most of these hypotheses will be "correlated", as they are generated from overlapping feature point sets.

It turns out that the method of generating the n hypotheses described above leads to a "uniform sampling" over the set

of all the possible hypotheses. This is described in more detail in Appendix A. As a result, even with a small value of n (≈ 50) the method yields good pose estimates.

Step 2: Estimating the mode: Each q_i is imagined to be the realization of a random variable $\mathbf{q}$ with an unknown distribution defined over $\mathbb{S}^{3+}$. The 3-sphere $\mathbb{S}^3$ is divided into a number of regions of equal area, or bins, that are denoted by B_j, $j = 1, \ldots, K_q$, where K_q is the number of regions. The algorithm described in [29] is used for this purpose. The pmf of the random variable $\mathbf{q}$ over the bins B_j, which is denoted by $p^{(q)}$, is an array of K_q numbers: $p_j^{(q)} = \mathrm{P}(\mathbf{q} \in B_j)$, where P denotes probability. A histogram estimate $\hat{p}^{(q)}$ of the pmf $p^{(q)}$ is computed by counting the number of points q_i inside each bin: $\hat{p}_j^{(q)} = \frac{1}{n}\sum_{i=1}^{n} I_{B_j}(q_i)$, where $I_A(x)$ is the indicator function of the set A. That is, $I_A(x) = 1$ if $x \in A$ and 0 otherwise. A useful property of the histogram-based estimate is that $\hat{p}_j^{(q)}$ is an unbiased estimate of $p_j^{(q)}$ even if the samples used to construct the estimates are correlated. Let B_{j^*} be the bin in which the pmf attains its maximum value, i.e., $j^* = \arg\max_j(\hat{p}_j^{(q)})$. An estimate of the mode of the pmf $p^{(q)}$ is obtained by taking the arithmetic mean of the unit-quaternions q_i's lying in B_{j^*} and then normalizing the mean, giving the coarse pose estimate denoted by $q^* \in \mathbb{S}^{3+}$.

Further, the geodesic or Riemannian distance $d_q(q^*, q_i)$ between the coarse pose estimate q^* and q_i's lying in B_{j^*} is computed. The pmf of $d_q(q^*, q_i)$ $\forall q_i \in B_j^*$ is obtained using $K_{q\varepsilon}$ equidistant bins of size $\varepsilon_q = \lceil \max d_q(q^*, q_i) - \min d_q(q^*, q_i) \rceil / K_{q\varepsilon}$. The dominant cluster in B_{j^*} is identified corresponding to the bin $B_{\varepsilon k^*}$, $\varepsilon k \in \{1, 2, \ldots, K_{q\varepsilon}\}$, where the pmf of $d_q(q^*, q_i)$ attains maximum value. The choice of the design parameter $K_{q\varepsilon}$ depends on the noise present in the measurements, such that $K_{q\varepsilon}$ should be chosen sufficiently large to reject the noisy measurements. The objective of the geodesic pmf is to find a refined pose estimate by identifying a cluster within B_j^*.

Step 3: Extracting low-noise measurements: Once the refining pmf of $d_q(q^*, q_i)$ is obtained, a subset $Q_q \subset S_q$ of rotation hypotheses $q_i \in S_q$ is selected such that

$$(B_{\varepsilon j^*} - 1)\varepsilon_q + \eta < d_q(q^*, q_i) < B_{\varepsilon j^*}\varepsilon_q + \eta, \qquad (3)$$

where $\eta \in \mathbb{R}$ denotes the minimum geodesic distance $\min d_q(q^*, q_i)$. The distance function $d_q(\cdot, \cdot)$ in (3) is the Riemannian distance between two rotations $q_1, q_2 \in \mathbb{S}^{3+}$ is given by

$$d(R_1, R_2) = \frac{1}{\sqrt{2}} \| \log(R_1^T R_2) \|_F, \qquad (4)$$

where $R_1, R_2 \in SO(3)$ are the rotation matrix representation of q_1, q_2, and the subscript F refers to the Frobenius norm.

Step 4: Averaging low-noise data: Let N_1 be the number of elements in the low-noise measurements of rotation Q_q obtained as described above and let R_i denote the rotation matrix corresponding to $q_i \in Q_q$. The *optimal* average of the rotation matrices $R_1 \ldots R_{N_1}$ in the Euclidean sense is the matrix $\hat{R}$ that satisfies [28]

$$\hat{R} = \underset{R \in SO(3)}{\operatorname{argmin}} \sum_{i=1}^{N_1} ||R_i - R||^2{}_F. \qquad (5)$$

It was shown by Moakher [28] that $\hat{R}$ defined by (5) can be computed by the orthogonal projection of the arithmetic average $\bar{R} = \sum_{i=1}^{N_1} \frac{R_i}{N_1}$ onto the special orthogonal group $\mathbb{SO}(3)$ by

$$\hat{R} = \bar{R}U \mathrm{diag}\left(\frac{1}{\sqrt{\Lambda_1}}, \frac{1}{\sqrt{\Lambda_2}}, \frac{s}{\sqrt{\Lambda_3}}\right) U^T, \qquad (6)$$

where the orthogonal matrix U is such that

$$\bar{R}^T \bar{R} = U^T D U \ \textit{and}\ D = \mathrm{diag}(\Lambda_1, \Lambda_2, \Lambda_3), \qquad (7)$$

and $s = 1$ if $\det \bar{R} > 0$ and $s = -1$ otherwise.

The matrix $\hat{R}$ computed using (6) is the desired estimate of the true rotation R.

B. Estimating translation

The estimation scheme for unit translation is very similar to that for the rotation. The unit translation data $t_i \in S_t$, $i = 1, \ldots, n$ represent realizations of the random variable $\mathbf{t}$ with an unknown distribution defined over the 2-sphere $\mathbb{S}^2$. The 2-sphere $\mathbb{S}^2$ is divided into a number of bins of equal area B_j, $j = 1, \ldots, K_\mathrm{t}$, $K_\mathrm{t} \in \mathbb{N}$ [29]. A histogram estimate $\hat{p}^{(t)}$ of the pmf $p^{(t)}$, where $p_j^{(t)} := \mathrm{P}(\mathbf{t} \in B_j)$ is then computed by counting the number of points t_i in B_j. An estimate of the mode of the unit translation distribution, denoted by t^*, is determined by computing normalized average of t_i's corresponding to bin B_{j^*} in which the pmf takes its maximum value: $j^* = \arg\max_j \hat{p}_j^{(t)}$. The pmf of the geodesic distance $d_t(t^*, t_i)$ is obtained using $K_{t\varepsilon}$ equidistant bins of size $\varepsilon_t = \lceil \max d_t(t^*, t_i) - \min d_t(t^*, t_i) \rceil / K_{t\varepsilon}$ and let $B_{\varepsilon j^*}$ be the bin in which the pmf attains its maximum value. Once the mode t^* is identified, the low-noise data set Q_t is selected by choosing those $t_i \in S_t$ that satisfies

$$(B_{\varepsilon j^*} - 1)\varepsilon_t + \eta < d_t(t^*, t_i) < (B_{\varepsilon j^*} - 1)\varepsilon_t + \eta, \qquad (8)$$

where $\eta \in \mathbb{R}$ denotes the minimum geodesic distance $\min d_t(t^*, t_i)$. Let N_2 be the number of elements in the low-noise data set Q_t of the unit translations obtained above. The normalized arithmetic mean of the unit translations in the set Q_t given by

$$\hat{t} = \frac{\sum_{i=1}^{N_2} \frac{t_i}{N_2}}{\left|\left| \sum_{i=1}^{N_2} \frac{t_i}{N_2} \right|\right|} \qquad (9)$$

is taken as the estimate of the unit translation t.

V. Performance evaluation

The performance of the proposed algorithm is compared with RANSAC and nonlinear mean-shift (NMS) algorithms. For each algorithm, the estimation performance metric is designed in terms of deviation from the known rotation and translation, i.e., a ground truth. The true rotation and translation between the frames in an i-th image pair are denoted by $R(i)$ and $t(i)$, respectively. The rotation and translation estimation error for the i-th image pair, denoted by $e_R(i)$ and $e_t(i)$, respectively, are defined as

$$e_R(i) = \|I - R(i)^T \hat{R}(i)\|, \quad e_t(i) = \|t(i) - \hat{t}(i)\|, \quad (10)$$

where $\hat{R}(i)$ and $\hat{t}(i)$ are the estimates of rotation $R(i)$ and unit translation $t(i)$, $\|\cdot\|$ denotes the 2-norm, and I denotes a $\mathbb{R}^{3\times 3}$ identity matrix.

Oxford Corridor sequence [30] has been used to demonstrate the robustness of the presented PEGUS algorithm in comparison to RANSAC and NMS algorithms. The dataset consists of 11 images with matched feature point pairs and the ground truth in the form of camera projection matrix for each camera frame. The number of feature point matches between the first and i^{th} image, $\forall i = 2, 3, \ldots, 11$, are as follows: $[409, 409, 269, 269, 199, 199, 149, 149, 104, 104]$. Given the feature point matches, pose estimates can be obtained between the first and i^{th} camera frame using PEGUS, RANSAC, and NMS algorithms. For each algorithm and for each image pair 1000 random trials are used to obtain 1000 pose estimates, i.e., for PEGUS and NMS different feature point combinations are used for the 1000 trials whereas for RANSAC as well as NMS different initial conditions are used. The mean rotation and translation pose estimation errors are shown in Figs. 1A and 1B, respectively, and Figs. 1D and 1E show the error standard deviation based on $n = 50$ pose hypotheses used for the PEGUS and NMS algorithms. The mean and standard deviation of the processing time is shown in Figs. 1C and 1F, respectively.

It can be seen from Fig. 1A that the rotation estimation performance of the three pose estimation algorithms is comparable to each other with PEGUS showing a marginal improvement over the other two methods. However, PEGUS demonstrates a significant improvement in the translation estimation performance compared to both RANSAC and NMS algorithms. The presented PEGUS algorithm shows visible translation estimation error for the last four frames when the number of matching features and inliers have decreased. The processing time plots shown in Figs. 1C and 1F indicate predictable processing times for the presented PEGUS and NMS algorithms, whereas the processing time for RANSAC increases with reduction in the number of inliers. Visual servo control methods relying on the rotation and translation estimates to design a stabilizing control low for a physical system are inherently sensitive to the processing time delay. Therefore, a pose estimation algorithm, such as PEGUS, ensuring bounded and predictable processing time can improve the performance and stability of such control systems.

VI. Synthetic Experiments

Synthetic data is produced to analyze the robustness, scalability, and behavior of pose estimation error with respect to the image noise. A random cloud of 100 Euclidean points was generated and projected on the image plane using a pinhole camera model. The Euclidean point cloud is viewed from two distinct camera positions with the known relative rotation and translation serving as a ground truth. 25% feature outliers were added by corrupting the 25 projected image point matches and a zero-mean Gaussian noise of standard deviation $\sigma = 0.1$ pixels was added independently to the x and y coordinates of the point cloud. The number of hypotheses for pose estimation using the PEGUS and NMS algorithms are assumed to be $n = 50$. The parameter values specified above are used throughout this section unless otherwise specified. In the subsequent results, the notation $\zeta = \{x : \delta : y\}$ implies that the value of parameter ζ is varied from x to y with an increment of δ.

A. Robustness to outliers

Robustness of the three pose estimation algorithms, PEGUS, RANSAC, and NMS, is analyzed by varying the percentage of outliers $\mathcal{P}_o$ from 0% to 95% with an increment of 5% and the outliers were added randomly to the synthetic data. For each case, the experiment was repeated 1000 times and pose estimates were obtained using PEGUS, RANSAC, and NMS algorithms.

The plots of mean rotation and translation estimation errors as a function of the percentage feature outliers are shown in Figs. 2A and 2B, respectively. As per the expectation, the performance of all the algorithms deteriorate as the number of outliers increase. From Fig. 2B it can be seen that the mean translation estimation error using PEGUS is minimum among the three algorithms, while RANSAC performs better in rotation estimation (see Fig. 2A) for the percentage outliers above 60%.

B. Scalability

Scalability of PEGUS is studied through synthetic experiments by varying the number of feature points M from 10 to 500 with an increment of 10. For each experiment 1000 random trials were conducted to obtain the mean rotation and translation estimation errors as shown in Fig. 3. It is evident that the presented algorithm incurs a large pose estimation error for less number of available feature points. This is due to the presence of a large number of correlated hypotheses that are corrupted by outliers. However, it can be seen that the pose estimation error remains steady even for a large number of feature points thus demonstrating suitability of the hypothesis generation scheme presented in Section IV-A, Step 1.

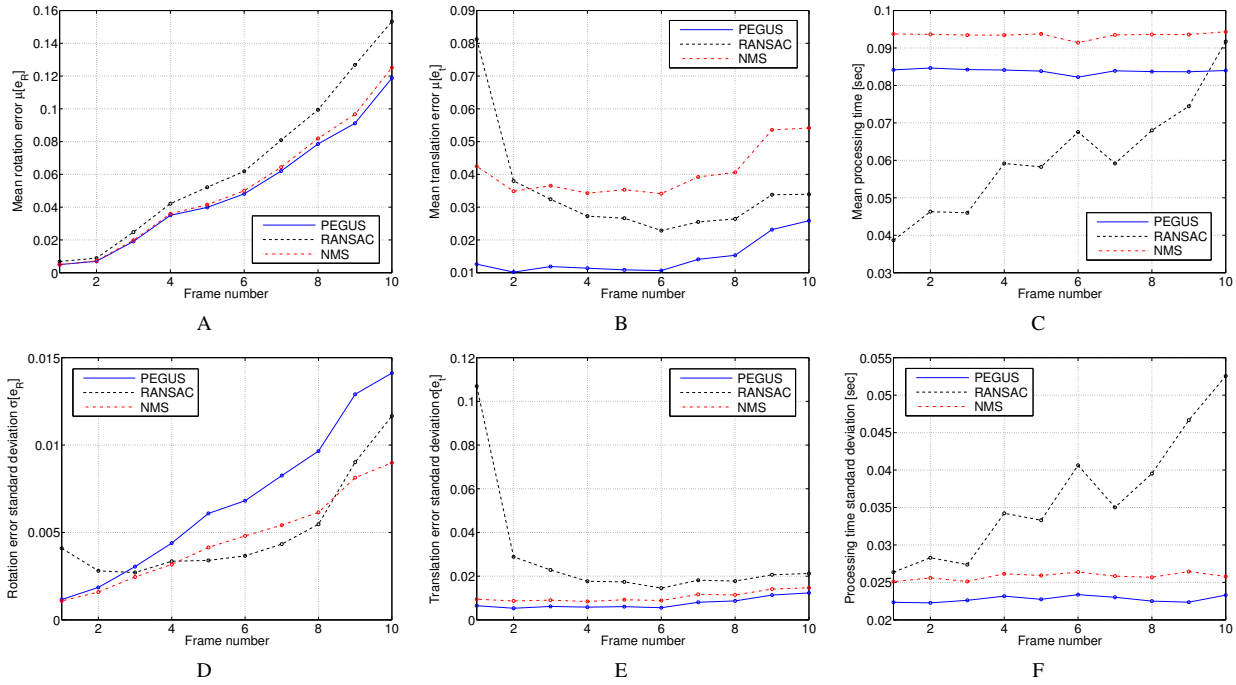

Figure 1. Oxford Corridor sequence: Mean and standard deviation of the rotation estimation error (A and D), translation estimation error (B and E), and processing time (C and F) exhibited by PEGUS, RANSAC, and NMS algorithms using 1000 random trials.

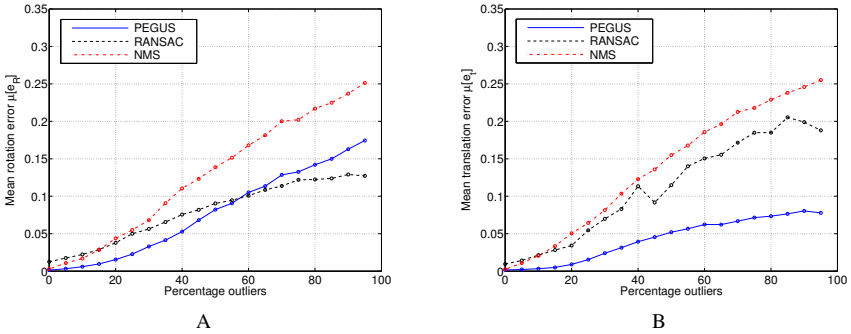

Figure 2. Robustness analysis: Mean of the rotation and translation estimation errors for PEGUS, RANSAC, and NMS algorithms using 1000 random trials of the synthetic data for the percentage feature outliers $\mathcal{P}_o = \{0{:}5{:}95\}$.

C. Sensitivity to noise

A set of synthetic experiments were carried out to compare the performance of PEGUS with RANSAC and NMS algorithms in the presence of zero mean Gaussian image noise. The noise standard deviation is varied from 0 to 4 pixels in the increments of 0.1 pixel. Figs. 4A and 4B show the performance comparison in terms of the mean rotation and translation estimation errors, respectively, as a result of 1000 random trials. The estimation performance of all the methods deteriorates with increase in the image noise. Although PEGUS may not show significant improvement over RANSAC and NMS algorithms in terms of noise rejection, the estimation performance of PEGUS is comparable to these methods.

VII. Conclusion

A robust two-view relative pose estimation algorithm is presented. Hypothesize-and-test methods such as RANSAC ignore all but one of the good hypotheses, whereas the proposed algorithm identifies a set of "low-noise" pose hypotheses among the large number of possible hypotheses to obtain a coarse pose estimate. Identification of the "low-noise" set of hypotheses is simplified by expressing the rotations as unit-quaternions and constructing a pmf by grid-

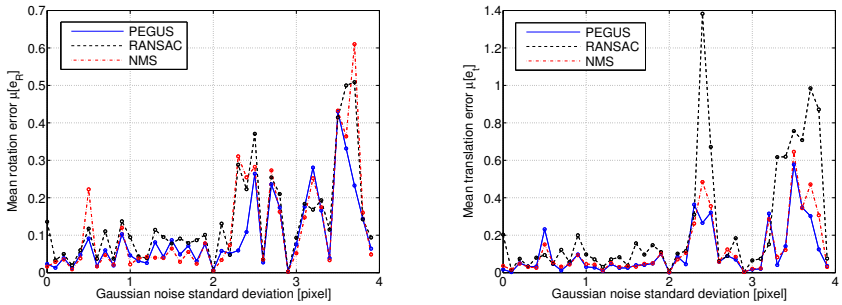

Figure 4. Effect of the zero-mean Gaussian image noise on the (A) mean rotation estimation error and (B) mean translation estimation error for PEGUS, RANSAC, and NMS algorithms based on 1000 random trials of the synthetic data for the noise standard deviation $\sigma_n = \{0{:}0.1{:}4\}$ pixels.

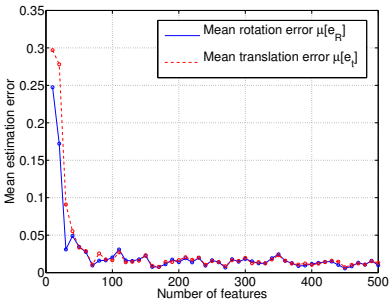

Figure 3. Scalability analysis: Effect of the number of feature points M on the mean rotation and translation estimation errors based on 1000 random trials of the synthetic data for the number of feature points $M = \{10{:}10{:}500\}$.

ding $\mathbb{S}^3$. A refined pose estimate is obtained by constructing another pmf of the geodesic distance on $\mathbb{S}^3$. An identical scheme is used for unit-translations, except that the hypotheses lie on a unit sphere in 3-dimensions. Experimental results demonstrate improved performance of the proposed method against RANSAC as well as non-linear mean shift, in terms of both the estimation accuracy and computation time. Since the proposed method does not involve any iterative search, its computation time is more predictable than that of RANSAC. Also, robustness analysis using synthetic data demonstrated improved translation estimation behavior of the presented algorithm over RANSAC and NMS.

APPENDIX

The accuracy of the pose estimates obtained by PEGUS algorithm depends on the n pose hypotheses generated in Step 1. In this section we describe some of the properties of the hypotheses generation scheme used in the algorithm.

Each distinct P pairs of point correspondence leads to a distinct hypothesis of q and t. Let $\mathbf{h}$ be the random variable representing the hypothesis that is obtained when the Simple Random Sampling With Replacement (SRSWR) scheme is executed. The possible values that $\mathbf{h}$ can take are denoted by $h_i, i = 1, \ldots, N_{\max}$. Each h_i represents a pair q_i, t_i since there exists a map from each set of P feature point pairs to hypotheses h_i, for instance, using the 8-point algorithm.

Proposition 1: The SRSWR scheme for hypotheses generation ensures that each possible hypothesis is obtained with equal probability, i.e., $P(\mathbf{h} = h_i) = \frac{1}{N_{\max}}$.

Proof: A hypothesis $\mathbf{h}$ is uniquely defined by the P point correspondence used to generate it, which are denoted by $\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^P$. We assume that the all feature point pairs are sorted to have increasing index from 1 through M.

$$\begin{aligned}\mathrm{P}(\mathbf{h} = h_i) &= \mathrm{P}(\mathbf{f}^1 = h_i^1, \mathbf{f}^2 = h_i^2, \ldots, \mathbf{f}^P = h_i^P) \\ &= \prod_{k=2}^{8} \mathrm{P}(\mathbf{f}^k = h_i^k | \mathbf{f}^{k-1} = h_i^{k-1}, \ldots, \mathbf{f}^1 = h_i^1) \\ &\quad \times \mathrm{P}(\mathbf{f}^1 = h_i^1) \\ &= \frac{1}{M-(P-1)} \frac{1}{M-(P-2)} \cdots \frac{1}{M} \end{aligned} \tag{11}$$

where the second equality follows from the chain rule of conditional probability. The third equality follows from the fact that once the first k point correspondence are picked, the probability of picking the next correspondence among the remaining points is $1/(M-k)$. Further, the generated hypothesis h_i is independent of the order of feature point correspondence. Therefore, the probability $\mathrm{P}(\mathbf{h} = h_i)$ in (11) can be re-written as

$$\begin{aligned}\mathrm{P}(\mathbf{h} = h_i) &= \frac{1}{M-(P-1)} \frac{1}{M-(P-2)} \cdots \frac{1}{M} P! \\ &= \frac{(M-P)!P!}{M!} = \frac{1}{N_{\max}}. \end{aligned} \tag{12}$$

where the definition of P-combination of set M presented in Section IV-A is used. From (12), it can be seen that a hypothesis h_i consisting of P pairs of point correspondence is sampled with a probability of $1/N_{\max}$ and replaced before the next draw is taken. Due to sampling with replacement,

the probability of selection of next hypothesis remains unchanged. Therefore in the presented algorithm, a subset $h_j \subset \mathbf{h}, j = 1, \ldots, n$ of hypotheses from the total possible hypotheses is selected with a uniform probability of $1/N_{\max}$. ■

REFERENCES

[1] R. Haralick, H. Joo, C. Lee, X. Zhuang, V. Vaidya, and M. Kim, "Pose estimation from corresponding point data," *Systems, Man and Cybernetics, IEEE Trans. on*, vol. 19, no. 6, pp. 1426–1446, Nov/Dec 1989.

[2] B. M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nlle, "Review and analysis of solutions of the three point perspective pose estimation problem," *Int. J. of Comput. Vision*, vol. 13, pp. 331–356, 1994.

[3] T. Q. Phong, R. Horaud, A. Yassine, and P. D. Tao, "Object pose from 2-D to 3-D point and line correspondences," *Int. J. of Comput. Vision*, vol. 15, pp. 225–243, July 1995.

[4] F. Dornaika and C. Garcia, "Pose estimation using point and line correspondences," *Real-Time Imaging*, vol. 5, pp. 215–230, June 1999.

[5] A. Ansar and K. Daniilidis, "Linear pose estimation from points or lines," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 25, no. 5, pp. 578–589, May 2003.

[6] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," Carnegie Mellon University, Tech. Rep. CMU-CS-91-132, Apr. 1991.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, November 2004.

[8] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *Comput. Vision, Proc. European Conf. on*, 2006, pp. 404–417.

[9] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 5828, pp. 133–135, 1981.

[10] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[11] D. Nistér, "An efficient solution to the five-point relative pose problem," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 26, no. 6, pp. 756–770, June 2004.

[12] R. Kumar and A. Hanson, "Robust methods for estimating pose and a sensitivity analysis," *CVGIP: Image Understanding*, vol. 60, no. 3, pp. 313–342, 1994.

[13] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 87–119, 1995.

[14] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim, "Robust regression methods for computer vision: A review," *Int. J. of Comput. Vision*, vol. 6, pp. 59–70, 1991, 10.1007/BF00127126.

[15] P. Rosin, "Robust pose estimation," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on*, vol. 29, no. 2, pp. 297–303, Apr. 1999.

[16] A. F. Siegel, "Robust regression using repeated medians," *Biometrika*, vol. 69, no. 1, pp. 242–244, Apr. 1982.

[17] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection (Wiley Series in Probability and Statistics)*. Wiley.

[18] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, pp. 381–395, 1981.

[19] P. H. S. Torr and A. Zisserman, "MLESAC: a new robust estimator with application to estimating image geometry," *Comput. Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[20] O. Chum and J. Matas, "Matching with PROSAC - progressive sample consensus," in *Comput. Vision and Pattern Recognition, Proc. IEEE Comput. Society Conf. on*, vol. 1, 2005, pp. 220–226.

[21] E. Michaelsen, W. V. Hansen, M. Kirchhof, J. Meidow, and U. Stilla, "Estimating the essential matrix: GOODSAC versus RANSAC," in *Photogrammetric Comput. Vision (PCV), ISPRS Symposium on*, 2006, pp. 220–226.

[22] D. Nistér, "Preemptive RANSAC for live structure and motion estimation," *J. of Machine Vision and Applications*, vol. 16, pp. 321–329, December 2005.

[23] J. Matas and O. Chum, "Randomized RANSAC with $T_{d,d}$ test," *Image and Vision Computing*, vol. 22, no. 10, pp. 837–842, Sept. 2004.

[24] O. Chum and J. Matas, "Optimal randomized RANSAC," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 30, no. 8, pp. 1472–1482, Aug. 2008.

[25] T. Sattler, B. Leibe, and L. Kobbelt, "SCRAMSAC: Improving RANSAC's efficiency with a spatial consistency filter," in *Comput. Vision, Proc. IEEE Int. Conf. on*, Oct. 2009, pp. 2090–2097.

[26] G. Stockman, S. Kopstein, and S. Benett, "Matching images to models for registration and object detection via clustering," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 4, no. 3, pp. 229–241, May 1982.

[27] R. Subbarao, Y. Genc, and P. Meer, "Nonlinear mean shift for robust pose estimation," in *Applications of Comput. Vision, Proc. IEEE Workshop on*. Washington, DC, USA: IEEE Comput. Society, 2007, p. 6.

[28] M. Moakher, "Means and averaging in the group of rotations," *SIAM J. on Matrix Analysis and Applications*, vol. 24, 2002.

[29] P. Leopardi, "A partition of the unit sphere into regions of equal area and small diameter," *Numerical Analysis, Electronic Trans. on*, vol. 25, pp. 309–327, 2006.

[30] Visual Geometry Group (VGG), University of Oxford, http://www.robots.ox.ac.uk/~vgg/data1.html.