2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

An Initialization Method for Monocular Visual Localization of Miniature Aerial Robots

Tak Kit Lau

Abstract— **This paper proposes a method to initiate the pose of an aerial robot from a glimpse of a monocular view during the rapid take-off. Unlike the existing filter-based methods that failed to estimate pose due to insufficient baseline, this paper utilizes a unified projective parametrization and a progressively suppressed refinement scheme for the non-planar homography estimations to tackle the baseline initialization problem, which has been notoriously and persistently encountered in the field of aerial robotics. Without pruning the elliptical uncertainty spheres iteratively in a filter-based framework as in the existing methods, the proposed method associates the unknown scene points and the pose of the agile aerial robot in a unified projective parametrization, and leverages a hypothesis-and-verify scheme to facilitate the decompositions of the non-planar homographies in line with a series of essential matrices. Moreover, a progressively suppressed strategy is introduced to minimize the hypothetic errors in the homographies and to minimize estimation divergence. Therefore, even with only a glance through a single camera onboard an aerial robot, a referable poses of the aerial robot can be initiated for the immediate navigation. In addition, the empirical results demonstrates that this method not only yields a consistent and referable estimation of the 6DoF pose of the robot in parallel with the agile movement of take-off, but also withstands the loop-closure evaluation in the 3D space.**

I. INTRODUCTION

Unmanned miniature[1] aerial robots belong to a particularly demanding subset of the aerial robots in view of their destined domain of deployments. The GNSS-denied, cluttered and narrow environment, such as the urban canyon or buildings, imposes significant difficulties for these robots to position themselves with the localization methods that rely on the Geographic Information Systems (GIS). Many attempted to solve this problem by utilizing the laser range finders, ultrasonic sensors, barometers, inertial sensors and magnetometers [1][2][3]. But the onboard cameras, which are commonly equipped on these robots and intuitively yield perceived information rich enough for the localization and navigation, are solely treated as a tool for the surveillance or the topological sensors against the drifting by using the optical flow method [4][5][6].

Some recent advancements in the field of vision processing for the aerial robots involve the tracking [7], obstacle avoidance [4] and geographical reconstruction [8]. For the visual localization, some attempted on the stereovision with the aid of the ranging or inertial sensors [9][10][11]. Interestingly, comparing with the binocular vision, the monocular vision is seldom employed alone on all scales of aerial platforms for the purpose of 6DoF relative positioning in spite of the obvious reasoning of being low in the payload requirement [3] in which all miniature aerial robots should favor.

Unlike the stereovision, the localization method which only utilizes the monocular vision necessitates an explicit initialization[2] which conditions on the sufficient baseline requirement [12] because the monocular vision requires a baseline for the triangulation of the map points. This condition is principally associated with the foundation of the prevalent and celebrated filter-based localization method [13][14]. This kind of filter-based frameworks, either the Extended Kalman filter (EKF) [14][15] or the Unscented Kalman filter (UKF) [16] which is a member of the Sigma-Point Kalman filters (SPKF) [17], is established on the Gaussian probability density function for all uncertainties. And with the help of the inverse depth parametrization [18], these filter-based methods can deal with the uncertain depths that exhibit the nonlinear uncertainty distributions. Although this kind of method improves significantly when comparing with the delayed initialization method [13] which relies on the Particle filter for the nonlinear depth estimations before switching to the Kalman filter framework by some cut-off variances [13], the initialization method for the monocular visual localization remains a notable and open challenge that has yet been thoroughly resolved [19][20].

This method proposes to utilize the projective parametrization to relate the scene points and the camera frame, and hence, the poses of the robot can be continuously extracted from the series of hypothesized essential matrices. Moreover, to curb the statistical error due to the random selections during the hypothesizing mechanism of the essential matrices, a progressively suppressed scheme for the bundle adjustment is introduced to efficiently offset that error and avoid the local minimums. The contribution of this paper is that, even with a glimpse of a monocular view, the proposed method yields a robust and un-delayed initialization of the 6DoF pose of the unmanned aerial robots in the rapid take-off scenario while the existing methods lose track and deliver non-referable estimations due to the limited baseline. To our best knowledge, it is the first publication to successfully and instantaneously initialize the pose of the aerial robots from only a glance of the monocular view. With no means to propose a visual localization method using a monocular vision, this paper focuses the discussion on the challenging initialization method [19][21] due to the insufficient baseline which essentially is the crux in all existing monocular visual localization methods. The video of the experiments are available at: `http://www.mae.cuhk.edu.hk/∼tklau/mvl`

This paper is closely related to the ongoing debate [20] on whether filter-based localization method is better that optimization-based ones in the setting of monocular SLAM. In this paper, we center our discussion on the initialization

This work was partially sponsored by the Hong Kong RGC under the grants 414707. T.K. Lau is with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, China. E-mail: `tklau@mae.cuhk.edu.hk`.

[1] *Miniature*: small enough to fly in indoor environment.

[2] This initialization step is often coined as *SLAM wiggle*.

978-1-61284-799-3/11/$26.00 ©2011 IEEE

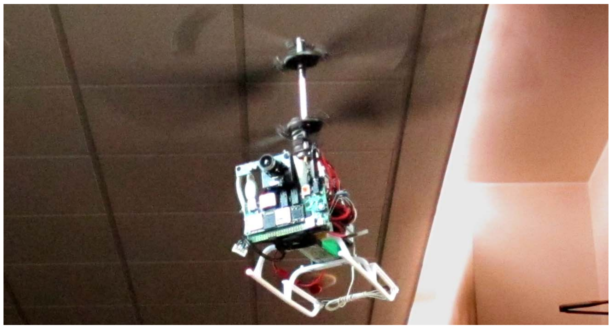


Fig. 1: The instrumented miniature co-axial helicopter in actions. The model is of EK1H-E020 from E-sky.

method for monocular visual localization, and suggests that our method, which is of an optimization-based approach, can succeed in the cases that filter-based methods failed.

II. Formulation of the Method

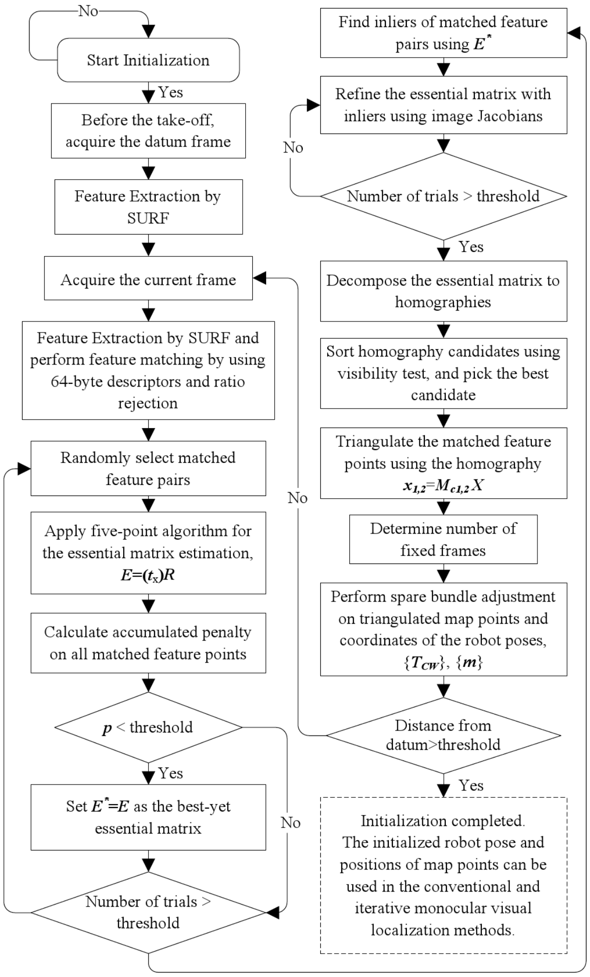
The monocular visual localization has long been confronted by the probabilistic approaches. Consider a fat sheaf of frames taken by a monocular vision in a scene, the scattered image patches, which are selectively picked by the detection schemes like the Harris corner [22] or the FAST corner [23], are matched frame-by-frame to prune the elliptical spheres of the uncertain depths of the feature patches through triangulations in some iterating filters, mostly the EKF [24]. However, even with the inverse depth parametrization [18] that linearizes the uncertainty of depths and hence enforces the covariances to follow the Gaussian mixture model, these probabilistic methods are still fundamentally a frame-by-frame pruning method and yields their best guess only if the uncertainties nicely follow some pre-defined distributions which mimic from the natural ambiguousness, or, when the elliptical spheres, which represent the uncertainty distribution of the feature positions, are pruned enough to faithfully satisfy the Gaussian distribution on a long run. However, it is not the case for the aerial robots which need to instantaneously initialize their 6DoF pose estimation in order to facilitate the navigation during the rapid motion of the take-off. While the aforementioned methods are doubtful [20] in dealing with the initialization problem which is significantly critical for the aerial robots during their first movement for the deployments, we address this issue from a new perspective by a global projective method than the filter-based approach. The idea of the proposed method is found on the obvious fact that scene points are projected to the camera frames based on the projection matrix and the inter-frame homography. For a camera pose C_1 and a scene point k,

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix}_{C_i} = \mathit{Project}\left(\boldsymbol{T}_{C_i,W}\begin{pmatrix} X_k \\ Y_k \\ Z_k \end{pmatrix}_W\right) \tag{1}$$

And equivalently for a camera pose C_2 in the same scene,

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix}_{C_j} = \mathit{Project}\left(\boldsymbol{T}_{C_j,W}\begin{pmatrix} X_k \\ Y_k \\ Z_k \end{pmatrix}_W\right) \tag{2}$$

where *Project* $(\cdot)$ is the pin-hold projection function, and is further elaborated in (4). $(u_k\ v_k)_{C_i}{}^T$ and $(u_k\ v_k)_{C_j}{}^T$ are the distorted image coordinates observed from the two camera frames. $(X_k\ Y_k\ Z_k)_W{}^T$ is the 3D position of the scene point k with respect to the world frame. $\boldsymbol{T}_{C_i,W}$ and $\boldsymbol{T}_{C_j,W}$ are the transformation matrix from the world frame to the camera pose C_i and C_j. While taking the first camera frame as a datum frame for the relative positioning, $\boldsymbol{T}_{C_j,W}$ will then describe the pose of the camera frame C_j. Without pruning the uncertainty elliptical spheres as in the aforementioned filter-based method, the camera pose can be theoretically solved by finding the projection matrix and the position of the scene point. But here are the pros and cons. The downside is that, (2) is sensitive to the noisy measurements of the matched feature pairs. And we address this issue and discuss the solution in the Section III. On the other hand, the advantage of this method is that, it can resolve the relative pose of the aerial robot even with only two frames, while the previous un-delayed initialization method, which involves a frame-by-frame uncertainty pruning operation, can hardly enforce the depth uncertainty to follow the Gaussian mixture model in such a glimpse of the view. But for the aerial robots, when they are deployed to actions in the unknown environment, that glimpse of a view is exactly all they observe to initialize their poses to facilitate the immediate navigation. The proposed method precisely addresses this initialization problem in this critical scenario in which the



No
Start Initialization
Yes
Before the take-off, acquire the datum frame
Feature Extraction by SURF
Acquire the current frame
Feature Extraction by SURF and perform feature matching by using 64-byte descriptors and ratio rejection
Randomly select matched feature pairs
Apply five-point algorithm for the essential matrix estimation, $E=(t_\times)R$
Calculate accumulated penalty on all matched feature points
p < threshold
Yes
No
Set $E^*=E$ as the best-yet essential matrix
Number of trials > threshold
Find inliers of matched feature pairs using E^*
Refine the essential matrix with inliers using image Jacobians
No
Number of trials > threshold
Yes
Decompose the essential matrix to homographies
Sort homography candidates using visibility test, and pick the best candidate
Triangulate the matched feature points using the homography $x_{1,2}=M_{c1,2}X$
No
Determine number of fixed frames
Perform spare bundle adjustment on triangulated map points and coordinates of the robot poses, $\{T_{CW}\}$, $\{m\}$
Distance from datum>threshold
Yes
Initialization completed. The initialized robot pose and positions of map points can be used in the conventional and iterative monocular visual localization methods.

Fig. 2: The proposed initialization method for the aerial robots. In view of that the first motion to be taken by an aerial robot in deployment is take-off, this method is especially designed to cope with this scenario in order to yield an immediate pose estimation in parallel with the departing motion. This instantaneous initialization can replace the existing delayed or un-delayed initialization methods and yield a robust pose estimation under the rapid movement of the take-off. The details of the method are further elaborated in the sections.

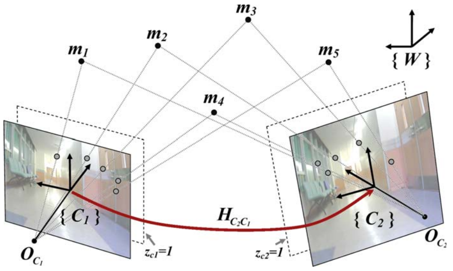


Fig. 3: The unified projective parametrization for the estimations of the non-planar homographies from the noisy measurements of matched features.

observations are limited and the robot motion is rapidly altering.

III. Estimation of Non-planar Homography from Noisy Observations

The problem of estimating the initial pose of an aerial robot which carries a single camera in parallel with the motion when it takes off, a theoretically guaranteed method is to retrieve $\boldsymbol{T}_{C_j,W}$ in (2). Although the non-planar homography can be used for estimating the pose [25], due to the seemingly overwhelming and noisy measurements of the feature matches between the frames, it is not practical to use the homography to solve for the robot pose in the closed-form solutions. In this section we will discuss a series of mechanisms to best estimate the homography from the noisy measurements of the feature matches. Consider that at $t = 0$, the first frame, which is also known as the datum frame, is acquired at a pose $\mathbf{T}_{RW}(0)$ with respect to the world frame from the monocular vision, and is represented by the SE(3) transformation matrix in the Lie groups [26]. It transforms from the world frame to the body frame of the robot, which is essentially the camera frame, at $t = 0$. And at $t = t_o$, another frame is acquired at $\mathbf{T}_{RW}(t_o)$, where $\mathbf{T}_{RW}(0) \neq \mathbf{T}_{RW}(t_o)$. Regardless of the motion the robot experiences, the corresponding feature pairs between these two frames must satisfy the constraints associated with the essential matrix [27], such that,

$$(\hat{\mathbf{x}}_c|_{t=t_o})^T \boldsymbol{E} (\hat{\mathbf{x}}_c|_{t=0}) = 0 \tag{3}$$

where $\hat{\mathbf{x}}_c\big|_{t=0}$ and $\hat{\mathbf{x}}_c|_{t=t_o}$ are the normalized image coordinates of a scene point on the frames taken at $t = 0$ and $t = t_o$, respectively. These normalized image coordinates can be represented by taking a pin-hole camera model with the distortions [28][29],

$$\begin{aligned}\begin{pmatrix} u \\ v \end{pmatrix} &= \mathit{Project}\left(\mathbf{T_{RW}}\begin{pmatrix} x_w \\ y_w \\ z_w \end{pmatrix}\right) = \mathit{Project}\begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} \\ &= \begin{pmatrix} u_o \\ v_o \end{pmatrix} + \begin{pmatrix} f_u & \alpha f_u \\ 0 & f_v \end{pmatrix} r' \begin{pmatrix} \frac{x_c}{z_c} \\ \frac{y_c}{z_c} \end{pmatrix} + \begin{pmatrix} d_{x1} \\ d_{x2} \end{pmatrix}\end{aligned} \tag{4}$$

where $(u\ v)^T$ is the distorted image coordinate which can be directly observed on the image frame. $(x_W\ y_W\ z_W)^T$ and $(x_c\ y_c\ z_c)^T$ are the 3D position of a feature point in the scene with respect to the world frame and the camera frame which is equivalent to the body frame of the aerial robot. $(u_o\ v_o)^T$ is the principal point. $f_{u,v}$ are the focal lengths expressed in units of horizontal and vertical pixels. α is the skew coefficient for non-rectangular pixels. r' is the distorted radius due to the radial distortion. $d_{x1,2}$ are the tangential distortions. The radial distortion can be expressed in a definite order, such that,

$$r' = 1 + k_1 r^2 + k_2 r^4 + k_5 r^6 \tag{5}$$

where $r = \frac{\sqrt{x_c^2+y_c^2}}{z_c}$. And, the tangential distortions can be written as,

$$\mathbf{d_x} = \begin{pmatrix} d_{x1} \\ d_{x2} \end{pmatrix} = \begin{pmatrix} 2k_3 \frac{x_c y_c}{z_c^2} + k_4 \left(r^2 + 2\left(\frac{x_c}{z_c}\right)^2\right) \\ k_3 \left(r^2 + 2\left(\frac{y_c}{z_c}\right)^2\right) + 2k_4 \frac{x_c y_c}{z_c^2} \end{pmatrix} \tag{6}$$

where $k_{1,2,5}$ are distortion coefficients which can be calibrated effortlessly from the conventional camera calibration toolboxes along with other intrinsic parameters. As the pixel is seldom non-rectangular, and hence the normalized image coordinates can be represented as followed,

$$\hat{\boldsymbol{x}}_c = \frac{1}{Z_c}\begin{pmatrix} X_c \\ Y_c \end{pmatrix} = \begin{pmatrix} \frac{u - u_o - d_{x1}}{f_u r'} \\ \frac{v - v_o - d_{x2}}{f_v r'} \end{pmatrix} \tag{7}$$

The essential matrix in (3) can be efficiently obtained from the five-point algorithm [30] if the corresponding feature matches are noise-free, which can hardly be realistic. Therefore, a RANSAC-type (Random Sample Consensus) scheme is implemented to iteratively refine and achieve an optimal essential matrix. To minimize the disturbance of outliers to the estimation of the essential matrix, the whole hypothesis-and-verify process [31] runs iteratively and randomly picks different matched feature pairs for the estimation of the essential matrix by the five-point algorithm [30]. Each essential matrix is penalized in terms of accuracy using the squared error [32], such that,

$$p = \begin{cases} \boldsymbol{e}^T \boldsymbol{e} & \text{, for } \|\boldsymbol{e}\| < \sigma_{\text{threshold}} \\ p_{max} & \text{, for } \|\boldsymbol{e}\| \geq \sigma_{\text{threshold}} \end{cases} \tag{8}$$

where p is the penalty. p_{max} is the maximum penalty to separate the hypothesis that does not fit the measurements. The error $\boldsymbol{e}$ is the projection error in the image coordinates. Taking a unified projection on the camera frame, such that the scene points are represented on a plane at $z_c = 1$, the error for the i^{th} scene point, which is observed in the j^{th} camera pose can be written as,

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{J}\left(\begin{pmatrix} x_{i,c_j} \\ y_{i,c_j} \end{pmatrix} - \begin{pmatrix} \hat{x}_{i,c_j} \\ \hat{y}_{i,c_j} \end{pmatrix}\right) \\ &= \begin{pmatrix} \frac{\partial u}{\partial x_c} & \frac{\partial u}{\partial y_c} \\ \frac{\partial v}{\partial x_c} & \frac{\partial v}{\partial y_c} \end{pmatrix}\begin{pmatrix} x_{i,c_j} - \hat{x}_{i,c_j} \\ y_{i,c_j} - \hat{y}_{i,c_j} \end{pmatrix}\end{aligned} \tag{9}$$

where $\boldsymbol{J}$ is the Jacobian matrix that establishes the relation of differential motion between the scene positions in the camera frame and the corresponding image coordinates. The Jacobian matrix can be derived by recalling (4) and (7). And for the ease of notation, $(\cdot)_{i,C_j}$ is written as $(\cdot)_C$. For the image coordinates,

$$u = u_o + f_u r' \left(\frac{x_c}{z_c}\right) + d_{x1},\ v = v_o + f_v r' \left(\frac{y_c}{z_c}\right) + d_{x2} \tag{10}$$

The derivative of the image coordinates in (10) with respect to the scene position at the unified camera frame can be derived,

$$\frac{\partial u}{\partial x_c} = \frac{\partial f_u r' x_c}{\partial x_c} + \frac{\partial d_{x1}}{\partial x_c} = f_u r' + f_u x_c \frac{\partial r'}{\partial x_c} + \frac{\partial d_{x1}}{\partial x_c} \tag{11}$$

where the partial differentiation of the radial and tangential distortions can be written using (5) and (6),

$$\frac{\partial d_{x1}}{x_c} = 2k_3 y_c + \frac{2k_4 x_c}{\sqrt{x_c^2 + y_c^2}} + 4k_4 x_c \tag{12}$$

$$\frac{\partial r'}{\partial x_c} = 2k_1 x_c + 4k_2 x_c \left({x_c}^2 + {y_c}^2\right) + 6k_5 x_c \left({x_c}^2 + {y_c}^2\right)^2 \tag{13}$$

Similarly,

$$\frac{\partial u}{\partial y_c} = f_u x_c \frac{\partial r'}{\partial y_c} + \frac{\partial d_{x1}}{\partial y_c} \tag{14}$$

$$\frac{\partial v}{\partial x_c} = f_v y_c \frac{\partial r'}{\partial x_c} + \frac{\partial d_{x2}}{\partial x_c} \tag{15}$$

$$\frac{\partial v}{\partial y_c} = f_v r' + f_v y_c \frac{\partial r'}{\partial y_c} + \frac{\partial d_{x2}}{\partial y_c} \tag{16}$$

where,

$$\frac{\partial d_{x1}}{\partial y_c} = 2k_3 x_c + \frac{2k_4 y_c}{\sqrt{x_c^2 + y_c^2}} \tag{17}$$

$$\frac{\partial d_{x2}}{\partial x_c} = \frac{2k_3 x_c}{\sqrt{x_c^2 + y_c^2}} + 2k_4 y_c \tag{18}$$

$$\frac{\partial d_{x2}}{\partial y_c} = \frac{2k_3 y_c}{\sqrt{x_c^2 + y_c^2}} + 4k_3 y_c + 2k_4 x_c \tag{19}$$

$$\frac{\partial r'}{\partial y_c} = 2k_1 y_c + \left[4k_2 y_c + 6k_5 y_c ({x_c}^2 + {y_c}^2)\right] \left({x_c}^2 + {y_c}^2\right) \tag{20}$$

Once the essential matrix is estimated through this scheme, the homography which is unsusceptible to the planar constraint can be obtained from this matrix. The extraction of the homography [25] can be performed by the method of the singular value decomposition (SVD). Consider that, the essential matrix is composed of the translation $\boldsymbol{t}_{3\times1}$, and the rotation matrix $\boldsymbol{R}_{3\times3}$, we have,

$$\boldsymbol{E} = (\boldsymbol{t}_\times)\, \boldsymbol{R} \tag{21}$$

Then, using the Lie groups representations [26], the homography can be written as,

$$\begin{aligned} \boldsymbol{H} &= (\boldsymbol{R}\,|\boldsymbol{t}) \\ &= \left(\boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^T \mid \mathit{Unskew}\left(\sigma \boldsymbol{U}\boldsymbol{A}\boldsymbol{U}^T\right)\right) \end{aligned} \tag{22}$$

in which,

$$\boldsymbol{A} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \boldsymbol{B} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{23}$$

where,

$$\sigma = \mathit{det}\,(\boldsymbol{U})\, \mathit{det}\,(\boldsymbol{V}) \tag{24}$$

The steps of the mechanism to choose the best estimations of the essential matrix and the homography are further elaborated in the Algorithm 1. Although this hypothetic and partly statistical approach rejects outliers, it still subjects to the randomness in the process of the random selection. Recalling (1) and (2), while the projection is calibrated and the inter-frame homography is now approximated, a refinement based on such a constraint shall be applied on the observations of the scene points in order to yield the estimation of the robot pose with the maximum likelihood. This part is introduced in the next section.

Algorithm 1 Estimation of the non-planar homography $\boldsymbol{H}^*$ and the essential matrix $\boldsymbol{E}$ from noisy measurement of matched feature pairs

```
INPUTS:
    $m$ {all matched feature pairs between two frames}
    $nTrial_1$ {number of trial for the main iteration}
    $nTrial_2$ {number of trial for refinements}
    $nRandomPair$ {number of pairs to choose randomly}
OUTPUT:
    $\boldsymbol{H}^*$ {Homography from non-planar geometry.
        It transforms from the first frame to the second
        frame}
for $i = 0$ to $nTrial_1$ do
    $index = random\_select(nRandomPair, m)$
    $\boldsymbol{E} = apply\_five\_points\_algorithm(m, index)$
    $score = 0$
    for all matched feature pairs do
        $penalty+ =$penalty_by_MLESAC$(\boldsymbol{E}, m)$
        {A maximum likelihood estimation method [32].
        Inliers are penalized by the degree of how bad they
        approximate the hypothesis.}
    end for
    if $penalty < penalty_{best}$ then
        $\boldsymbol{E}^* = \boldsymbol{E}$
        $penalty_{best} = penalty$
    end if
end for
for all cases of homography decompositions do
    $\boldsymbol{E} = (\boldsymbol{t}_\times)\,\boldsymbol{R}$
    $\sigma = det(\boldsymbol{U})det(\boldsymbol{V})$
    $\boldsymbol{H}.translation = \sigma\boldsymbol{U}\boldsymbol{A}\boldsymbol{U}^T$
    $\boldsymbol{H}.rotation = \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^T$
    Put $\boldsymbol{H}$ into a stack, $H\_stack$, as a candidate
end for
for $i = 0$ to $nTrial_2$ do
    refine_homography_by_image_jacobians$(m$,$H\_stack\{i\})$
end for
for $i = 0$ to $size(H\_stack)$ do
    $score = 0$
    for $j = 0$ to $size(m)$ do
        $(x\ y\ z)^T$ =apply_homography$(H\_stack\{i\}, m\{j\})$
        if $z > 0$ then
            $score$++
        end if
        if $score > score_{best}$ then
            $score_{best} = score$
            $\boldsymbol{H}^* = H\_stack\{i\}$
        end if
    end for
end for
```

IV. Refinement of Pose Estimations Based on Observations

From the noisy measurements of the matched feature pairs, a considerable effort is applied to estimate the essential matrix and the inter-frame homographies and hence the robot pose can be evaluated without pruning the uncertainty elliptical spheres to follow the Gaussian mixture model as in the previous initialization method. Now, although the hypothesis-and-verify scheme in the previous section is effective against the outlying measurements of the matched features, it still yields a degree of inaccuracy which is statistically associated with its random selection. Hence a bundle adjustment and a suppression scheme for the efficiency consideration are applied on the estimations of the homographies and the triangulated map points. Essentially, this refinement is to minimize the re-projection error, such that,

$$\begin{aligned} &\{\{\boldsymbol{T}_{C_1W}, ..., \boldsymbol{T}_{C_NW}\}, \{\boldsymbol{m}_0, ..., \boldsymbol{m}_K\}\} \\ &= \underset{\{\{\boldsymbol{T_{CW}}\},\{\boldsymbol{m}\}\}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=0}^{K} \left(\begin{pmatrix} u_{j,i} \\ v_{j,i} \end{pmatrix} - \mathit{Reproject}\left(\boldsymbol{m_j}, \boldsymbol{T_{C_iW}}\right)\right) \end{aligned} \tag{25}$$

where $\boldsymbol{m}_j$ is the 3D position of the j^{th} map point in the scene with respect to the world frame. $(u_{j,i}\ v_{j,i})^T$ is the image coordinate of the j^{th} map point and is observed in the camera frame C_i. The *Reproject*$(\cdot)$ is the re-projection function which projects a map point to a calibrated camera.

A. Linear Triangulation for Scene Points

From (25) the refinement of the homography relies on the 3D positions of the scene points which scatter across the frames. Based on the preliminary estimation of the homography, the rough estimations of the positions of the

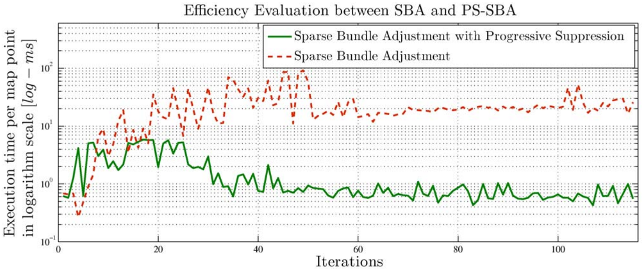


Fig. 4: The scheme of progressive suppression not only improves the execution time, but it also avoids the erroneous local minimum.

scene points can be obtained. Although the linear triangulation is susceptible to outliers and cannot yield an accurate estimations of the positions from the noisy measurements, the sparse bundle adjustment is scheduled to take place in the next step with the progressive suppression, hence the efficient linear triangulation can be employed for the 3D position estimations, such that,

$$\left(\hat{\boldsymbol{X}}_{C_1}\times\right)\boldsymbol{T}_{C_1W}\boldsymbol{X}_W=\boldsymbol{0},\quad \left(\hat{\boldsymbol{X}}_{C_2}\times\right)\boldsymbol{T}_{C_2W}\boldsymbol{X}_W=\boldsymbol{0} \tag{26}$$

Hence,

$$\begin{pmatrix} \hat{y}_{C_1}{}^3\boldsymbol{T}_{C_1W}-{}^2\boldsymbol{T}_{C_1W} \\ \hat{x}_{C_1}{}^3\boldsymbol{T}_{C_1W}-{}^1\boldsymbol{T}_{C_1W} \\ \hat{y}_{C_2}{}^3\boldsymbol{T}_{C_2W}-{}^2\boldsymbol{T}_{C_2W} \\ \hat{x}_{C_2}{}^3\boldsymbol{T}_{C_2W}-{}^1\boldsymbol{T}_{C_2W} \end{pmatrix}\boldsymbol{X}_W=\boldsymbol{M}_{\boldsymbol{C_1},\boldsymbol{C_2}}\boldsymbol{X}_W=\boldsymbol{0} \tag{27}$$

where ${}^i\boldsymbol{T}_{C_1W}, {}^i\boldsymbol{T}_{C_2W}$ are the i^{th} row vectors of the SE(3) transformation matrices $\boldsymbol{T}_{C_1W}, \boldsymbol{T}_{C_2W}$. $(\hat{x}_{C_1}\ \hat{y}_{C_1})^T$ and $(\hat{x}_{C_2}\ \hat{y}_{C_2})^T$ are the normalized image coordinates of a scene point observed and represented in the camera frames C_1, C_2. This linear combination in $\boldsymbol{M}_{\boldsymbol{C_1},\boldsymbol{C_2}}$ can then be solved by SVD.

B. Progressive Suppression on Robot Poses and Map Points

Upon triangulating the scene points, a refinement is needed to offset the inaccurate results of the feature matching. However, even with the most efficient bundle adjustment which makes use of the Levenberg-Marquardt method [33], the bundle adjustment can hardly be a timely measure for the refinement as the processing time jumps exponentially with the number of observations and the camera poses. Nevertheless, unlike the iterative RANSAC-type hypothetic methods, the bundle adjustment produces the optimal results in terms of the maximum likelihood in each run. Therefore, instead of firing the bundle adjustment to refine the robot poses along with the previously bundle-adjusted poses and scene points, a progressively suppression scheme is enforced to fix the previously refined robot poses for each operation of the bundle adjustment. Hence an accurate estimation of the relative pose can be delivered without losing the efficiency. On the other hand, the suppression can avoid the erroneous local minimum [34] which often appears on the robot poses that are repeatedly refined through the bundle adjustment. The dramatic improvement in the efficiency is shown in Fig. 4.

V. Verification

To verify the proposed method, the experiments were extensively carried out on an instrumented miniature coaxial helicopter which weighs 470g and spans with two pairs of blades that is 460mm long in diameter each. It carries a CMOS camera with a lens of 4mm in the focal length. This aerial robot supports onboard recording by a single-board-computer (SBC) manufactured by Gumstix and can wire the vision data to a ground station. The videos taken on the aerial robot are recorded and processed on a ground computer after the flights. From the experiments, the co-axial helicopter took off with the forward-looking camera. Using the proposed method, the robot pose was instantaneously estimated in parallel with the motion as opposed to the previous un-delayed initialization method which failed in the same sets of experimental data. On the other hand, a manual loop-closure test was performed on the proposed method. As a whole, the experimental results demonstrated that the proposed initialization method can not only simultaneously initialize the pose of the aerial robot regardless of the limited baselines and the rapid motion during the take-off, but also keep tracking the camera motion continuously and consistently close the loop in the loop-closure evaluation. The performances are shown in Fig. 5 (a, b) and Fig. 6.

VI. Concluding Remarks

This paper suggests that by utilizing a simple, textbook-type projective parametrization for the estimations of the inter-frame homography under a progressively suppressed refinement scheme, the miniature aerial robots can initiate the scene points and localize themselves in parallel with the rapid movement during the take-offs. As opposed to the previous un-delayed initialization method that yields the best estimation only when the elliptical uncertainty spheres for the depths are sufficiently pruned through the enough frames in order to enforce the uncertainties to converge from a Gaussian mixture model to a single Gaussian probabilistic distribution, the proposed method solves for the poses by the extractions of the non-planar homographies from the noisy measurements of the matched features. Therefore, even with only a glance of the view, the pose of the aerial robot and the positions of the scene points can be accurately estimated without using the inverse depth parametrization.

Acknowledgment

The author gives warm thanks to Yun-Hui Liu for his helpful discussion and support, and to Raymond Kai-Wun Lin for his help in the development and setup of the embedded camera systems on the coaxial aerial robot.

References

[1] A. Hyslop and J. Humbert, "Autonomous Navigation in Three-Dimensional Urban Environments Using Wide-Field Integration of Optic Flow," *AIAA Journal of Guidance, Control, and Dynamics*, pp. 147–159, January 2010.

[2] M. Achtelika, A. Bachrachb, R. Heb, S. Prenticeb, and N. Royb, "Autonomous navigation and exploration of a quadrotor helicopter in GPS-denied indoor environments," in *Proceedings of the Robotics: Science and Systems Conference*, 2008.

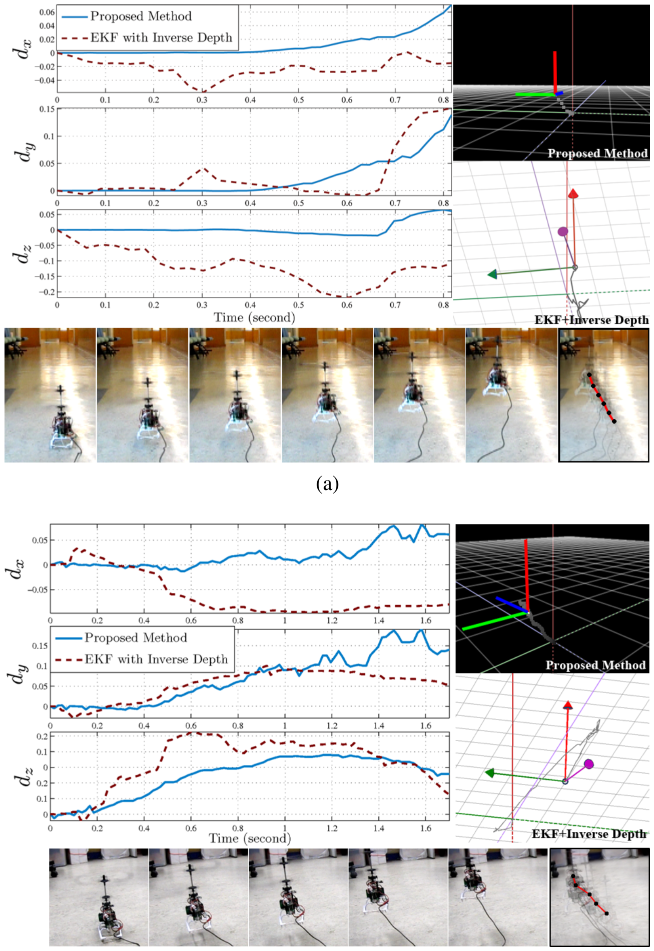
[3] S. Ahrens, D. Levine, G. Andrews, and J. How, "Vision-Based Guidance and Control of a Hovering Vehicle in Unknown, GPS-denied Environments," in *Proceedings of the IEEE Int'l Conference on Robotics and Automation*, 2009, pp. 2643–2648.

[4] AR Drone from Parrot SA, "http://ardrone.parrot.com/parrot-ar-drone/en/," Internet, 2009.

[5] W. A. C. R. Kehoe, J. and R. Lind, "State Estimation using Optical Flow from Parallax-Weighted Feature Tracking," in *Proceedings of the AIAA Guidance, Navigation, and Control Conference*, 2006.

[6] D. Lee, P. Merrell, Z. Wei, and B. Nelson, "Two-frame structure from motion using optical flow probability distributions for unmanned air vehicle obstacle avoidance," *Journal of Machine Vision and Applications*, pp. 1432–1769.

[7] C. Jones, J. Heyder-Bruckner, T. Richardson, and C. Jones, "Vision-based Control for Unmanned Rotorcraft," in *Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2006.



(a)

(b)

Fig. 5: ***Rapid take-off motion:*** In these trials, the co-axial helicopter was commanded to take-off with inclination to the left (direction of positive x-axis). The EKF-based method could only sensibly estimate the vertical displacement which experienced the most significant movement. For the motion with shorter baselines, such as the horizontal motions, the results by the previous initialization method were not even correct in the signs. However, the proposed was sustained to yield a referable pose under the rapid movement of the take-off. See the accompanying video for the trials#1, 2.

[8] T. Templeton, D. Shim, C. Geyer, and S. Sastry, "Autonomous vision-based landing and terrain mapping using an MPC-controlled unmanned rotorcraft," in *Proceedings of the IEEE Int'l Conference on Robotics and Automation, Roma, Italy*, 2007, pp. 1349–1356.

[9] J. Langelaan, "State estimation for autonomous flight in cluttered environments," *AIAA Journal of Guidance Control and Dynamics*, vol. 30, no. 5, pp. 1414–1426, 2007.

[10] F. Caballero, L. Merino, J. Ferruz, and A. Ollero, "Vision-Based Odometry and SLAM for Medium and High Altitude Flying UAVs," *in Journal of Intelligent and Robotic Systems*, vol. 54, no. 1, pp. 137–161, 2009.

[11] Beau Tippetts, Dah-Jye Lee, Spencer Fowers, James Archibald, "Real-Time Vision Sensor for an Autonomous Hovering Micro Unmanned Aerial Vehicle," *Journal of Aerospace Computing, Information, and Communication*, vol. 6, no. 10, pp. 570–584, 2009.

[12] M. Tomono, "Monocular slam using a Rao-Blackwellised particle filter with exhaustive pose space search," in *Proceedings of the IEEE Int'l Conference on Robotics and Automation, Roma, Italy*, 2007, pp. 2421–2426.

[13] A. Davison, I. Reid, N. Molton, and O. Stasse, "MonoSLAM: Real-time Single Camera SLAM," *in the IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, p. 1052, 2007.

[14] J. Civera, O. Grasa, A. Davison, and J. Montiel, "1-Point RANSAC for EKF-Based Structure from Motion," in *Proceedings of the IEEE Int'l Conference on Intelligent Robots and Systems*, 2009.

[15] J. Montiel, J. Civera, and A. Davison, "Unified inverse depth parametrization for monocular slam," in *in Proceedings of Robotics: Science and Systems*, 2006.

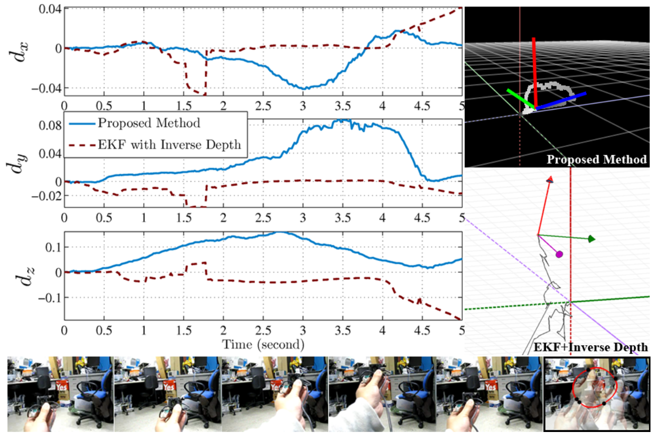


Fig. 6: ***Loop-closure evaluation:*** To test the performance of the proposed method in terms of the consistency, a manual loop-closure test was performed. In the trial, the camera was manually actuated to follow a closed-loop trajectory in a cluttered scene. It shows that the proposed method could immediately initialize its pose from only a glimpse of view, and withstand the test and close the loop not just on the topological levels but also in the 3D space. On the contrary, the EKF-based positioning method failed to initialize its pose even with the inverse depth parametrization. See the accompanying video for the trial#3.

[16] S. Holmes, G. Klein, and D. Murray, "A Square Root Unscented Kalman Filter for visual monoSLAM ," in *Proceedings of the IEEE Int'l Conference on Robotics and Automation, Pasadena, CA*, 2008, pp. 3710–3716.

[17] S. Julier and J. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Proceedings of the Int'l Symposium Aerospace/Defense Sensing, Simulation and Controls*, vol. 3, 1997, p. 26.

[18] J. Civera, A. Davison, and J. Montiel, "Inverse Depth Parametrization for Monocular SLAM," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 932–945, 2008.

[19] D. Kragic and M. Vincze, "Vision for Robotics," *Foundations and Trends in Robotics*, vol. 1, no. 1, pp. 1–78, 2010.

[20] H. Strasdat, J. Montiel, and A. Davison, "Real-Time Monocular SLAM: Why Filter?" in *Proceedings of the IEEE Int'l Conference on Robotics and Automation, Anchorage, AK*, 2010.

[21] D. Kragic, *Unifying Perspectives in Computational and Robot Vision*. Springer, 2008.

[22] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proceedings of the 4th Alvey Vision Conference*, vol. 15, 1988, pp. 147–151.

[23] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *Lecture Notes in Computer Science*, vol. 3951, p. 430, 2006.

[24] F. Bonin-Font, A. Ortiz, and G. Oliver, "Visual Navigation for Mobile Robots: A Survey," *in Journal of Intelligent and Robotic Systems*, vol. 53, pp. 263–296, 2008.

[25] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.

[26] B. Hall, *Lie groups, Lie algebras, and representations: an elementary introduction*. Springer, 2003.

[27] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, p. 133, 1981.

[28] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[29] J. Heikkila and O. Silven, "A Four-step Camera Calibration Procedure with Implicit Image Correction," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997, pp. 1106–1112.

[30] H. Stewénius, C. Engels, and D. Nistér, "Recent developments on direct relative orientation," *in ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 60, no. 4, pp. 284–294, 2006.

[31] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 2007.

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

[33] K. Levenberg, "A method for the solution of certain nonlinear problems in least squares," *The Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944.

[34] B. Triggs, Z. A., and R. Szeliski, *Vision Algorithms: Theory and Practice : International Workshop on Vision Algorithms, Corfu, Greece, September 21-22, 1999 : Proceedings*. Springer, 2000.