MVCSLAM: Mono-Vision Corner SLAM for Autonomous Micro-Helicopters in GPS Denied Environments

Koray Celik*, Soon-Jo Chung†, and Arun K. Somani‡
Iowa State University, Ames, Iowa, 50011, USA

We present a real-time vision navigation and ranging method (VINAR) for the purpose of Simultaneous Localization and Mapping (SLAM) using monocular vision. Our navigation strategy assumes a GPS denied unknown environment, whose indoor architecture is represented via corner based feature points obtained through a monocular camera. We experiment on a case study mission of vision based SLAM through a conventional maze of corridors in a large building with an autonomous Micro Aerial Vehicle (MAV). We propose a method for gathering useful landmarks from a monocular camera for SLAM use. We make use of the corners by exploiting the architectural features of the manmade indoors.

I. Introduction

The capability of vision based SLAM in an autonomous MAV can provide vital information for situation awareness. This is particularly useful in complex urban environments which pose unique challenges and risks for the military forces to conduct urban operations. A vision-based solution does not emit light or radio signals, it is portable, compact, cost-effective and power-efficient. Such a platform has a broad range of potential military applications including navigation of robotic systems and soldier position localization with wearable or helmet-mounted devices. Moreover, an MAV with the ability to hover can play a key role in Intelligence, Surveillance and Reconnaissance missions held at GPS denied environments which are not suitable for fixed wing flight, as illustrated in Figure 1.

Nonetheless, the limitations on payload, size, and power, inherent in small MAVs, pose technological challenges due to the direct proportionality in between the quality and the weight of conventional sensors available. Under these circumstances, a theory for developing autonomous systems based on the information gathered from images is appealing, since a video-camera possesses a far better information-to-weight ratio than any other sensors available today. On the other hand, it breeds another rich kaleidoscope of computational challenges. There is no standard formulation of how a particular high level computer vision problem should be solved and, methods proposed for solving well-defined application-specific problems can seldom be generalized.

Indoor flight of a rotorcraft MAV, where no GPS coverage is available, is a collective effort of two main challenges; platform attitude management, localization and mapping. The former is straightforwardly automated via lightweight sensors such as gyroscopes and with minimal information about the environment, or even lack thereof. Whereas the latter requires gathering and aggregation of excessive amounts of information about the surroundings, particularly true for a vehicle that would be destroyed in even the most superficial impact with the surroundings. The stringent weight requirements of MAVs prevent the use of standard obstacle sensing mechanisms such as laser range-finders,[1] and parabolic cameras,[2].[3] However, the machine

*Doctoral Research Assistant, Department of Electrical and Computer Engineering, koray@iastate.edu.
†Assistant Professor of Aerospace Engineering and Electrical & Computer Engineering, sjchung@alum.mit.edu.
‡Anson Marston Distinguished Professor, Jerry R. Junkins Endowed Chair, and Department Chair, Department of Electrical and Computer Engineering, arun@iastate.edu.

vision technology has evolved so as to allow for video-cameras less than an ounce in weight with a decent picture quality, and such a video stream includes more information about the surrounding environment than other sensors alone can provide.

Nevertheless, this information comprises a surpassingly high level of abstraction and redundancy, which is particularly aggravated in cluttered environments. Even after three decades of research in machine vision, the problem with "understanding" sequences of images stands bordering on being uninfluenced, as it requires acutely specialized knowledge to interpret. Ironically, the lack of such knowledge is often the main motivation behind conducting a reconnaissance mission with an MAV. Therefore the technique used in navigating an MAV by vision alone must assume minimal a priori knowledge, while it still provides reliable results. The MAV needs to construct a collective view of its unknown environment in order to navigate through it. The contribution of this paper is a new absolute range and bearing measurement algorithm using a monocular camera, and its experimental validation. Such measurements can be used for a vision-based navigation and SLAM problem in an unknown indoor environment.

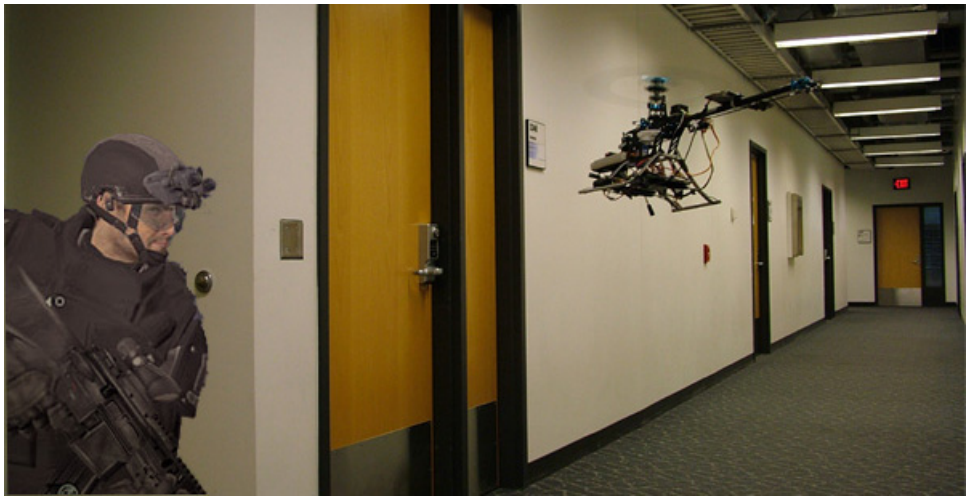

Figure 1. Potential applications of vision-based navigation in GPS denied environments.

II. Related Work

There are two main technological challenges associated with the VINAR problem: the lack of absolute depth information, and the development of robust SLAM algorithms. Since an image is a projection of a three dimensional world on a two dimensional surface, it is not essentially different than a shadow. It contains no depth information to those without comprehensive knowledge pertaining to its content. To mitigate the complications entailed by the absence of direct depth information, some alternative approaches have been tried involving the attachment of additional sensors to a camera, such as laser range finders[4] and cross validating the precise depth information provided by the laser range finder with the interpretations from the camera. However, such a technological advantage is a luxury for a rotorcraft MAV, it is more appropriate for a land based robot with no practical weight constraints. Even if a laser range finder could be designed as light as a camera, their range-to-weight ratio is much worse in comparison with a vision solution and, since they make a one dimensional slit through the scene versus the two dimensional signal created by a video-camera a complicated mechanical gimbal assembly is required to allow the laser range finder to perform a two-dimensional scan, adding to the overall weight and power consumption of the device.

Using a video-camera with an adjustable focus via moving lenses has been discussed in the literature[5] owing to the depth-of-field effect and the Scheimpflug Principle,[6] in which the distance in front of and beyond the particular subject in front of a video-camera appearing to be out of focus when the lens axis is perpen-

dicular to the image plane. Therefore, the distance of a particular area in an image where the video-camera has the sharpest focus can be acquired. Nonetheless, the focus of interest may not be an useful feature to begin with. The other reasons rendering these methods far from practical for our application include calibration issues specific to different cameras and lenses, and limitations of cameras currently available that are suitable for MAV use. In addition, unless the lenses can be moved at a very high frequency, this approach will significantly reduce the sensor bandwidth.

Binocular cameras for stereo-vision have been promising tools for range measurement for purposes of path planning and obstacle avoidance where the computer compares the images while shifting the two images together over top of each other to find the parts that match. The disparity at which objects in the image best match is used by the computer to calculate their distance. Be that as it may, binocular cameras are heavier and more expensive than their monocular counterparts, and stereo-vision has intrinsic limitations in its ability to measure the range,[28] particularly when large regions of the image contain a homogeneous texture such as a wall or a carpet. Furthermore, human eyes change their angle according to the distance to the observed object to detect different ranges, which represents a significant mechanical complexity for a lens assembly and a considerable challenge in the geometrical calculations for a computer.

The literature recognizes MonoSLAM[7] an elegant approach to vision based SLAM with minimum assumptions about the free movement of the camera, it may be summarized in three steps as follows, detect and match feature points, predict motion via analyzing feature points with error estimates, and update a map with locations of feature points. This results in "a probabilistic feature-based map, representing at any instant a snapshot of the current estimates of the state of the camera and all features of interest and, crucially, also the uncertainty in these estimates".[7] These error estimates mentioned, along with the map containing all known feature points, allow the algorithm to correct for drift when a feature point is rediscovered, providing a precise tracking system in which other than a standard initialization target defining the origin and orientation of the world coordinate frame. However, MonoSLAM assumes an extensive feature initialization procedure, and is not meant to leave the immediate vicinity of the starting position. In case it did so, with so many new features being introduced, using "image patches" as feature points, in which, small portions of the scene are stored in the memory for later reference via correlation, the increasing number of possible feature points would quickly become overwhelming for the computer mounted on an MAV with limited power and system resources. In addition, corners are better (i.e. more rigid) features in overall, considering the features used in MonoSLAM experiments. Moreover, MonoSLAM assumes a hand-held camera with a limited range but in contrast, the camera in this paper should be able to move through, for instance, an entire floor of a building, and still be able to maintain the track of its relative position. We address these issues in the following sections.

III. The Platform

Our test platform for the VINAR is an electric powered rotorcraft MAV with a maximum payload of 2lbs and 700mm rotor-span. With the most current battery technology available at the time of this paper the endurance is approximately 10 minutes, with up to a mile of communications range. Mechanically identical to its full-size counterparts, the MAV features true-to-life collective pitch helicopter flight dynamics. The flight stability and control is handled via an on-board IMU and autopilot unit from Micropilot[10] that uses two yaw gyroscopes, two roll and pitch gyroscopes, three-axis accelerometers, barometric and ultrasonic altimeters, and a magnetic compass to achieve flight control. The on-board digital magnetic compass measures the bearing of the MAV in degrees along the yaw axis with respect to the magnetic North. The vision-computed heading of the MAV is cross validated with the true heading from the compass.

Since the MAV does not have any GPS reception indoors, the autopilot is merely responsible for governing the PID loops aileron from roll, elevator from pitch, and collective (mixed with throttle via a quadratic function specific to the aircraft) from altitude to keep the MAV at hover. In other words the autopilot is *flying* the MAV but not *navigating* it, since it has no way of measuring the consequential results of its actions. Navigation, including obstacle avoidance and Simultaneous Localization and Mapping (SLAM) are to be performed by vision, via a wireless light-weight on-board video camera. In other words, VINAR based SLAM is replacing the GPS.

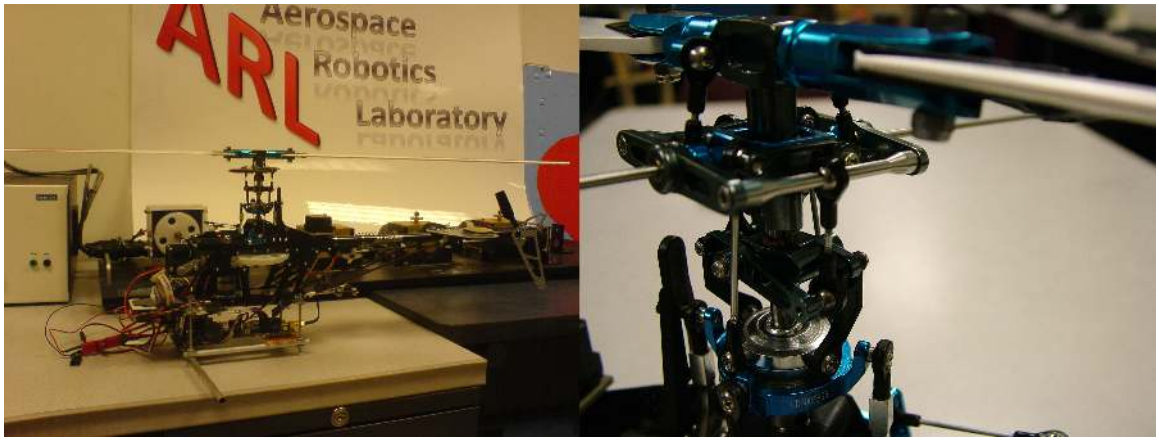

Figure 2. The rotorcraft MAV used in our experiments.

IV. Landmark Extraction from Live Video Stream

A dependable solution for tracking motion and trajectory of a vehicle, and the ability to extract landmarks in a reliable manner complement each other. Maintaining a set of features large enough to allow for accurate motion estimations, yet sparse enough so as not to produce a negative impact on the system performance is imperative. The main difficulty for challenge of a vision based approach to landmark extraction is the common similitude of landmarks and other features. When ranking the landmarks, the ones that neither vanish nor shift positions with respect to a stationary observer dynamically, but only with respect to the moving observer, are considered superordinate. Shi and al.[12] propose a criteria to what constitutes to a "better" feature in a two-dimensional image, via properties such as texturedness, dissimilarity, and convergence. According to Shi and al. sections of an image with large eigenvalues are to be considered "good" features.

As the video frames advance in time, changes between two frames is described as $I_1(\vec{x}_f) = I_0(\vec{x}_f + \delta(\vec{x}_f))$ which denotes that by moving the points from the frame $I_0(\vec{x}_f)$ by $\delta(\vec{x}_f)$, the new frame $I_1(\vec{x}_f)$ is reconstructed. The vector $\vec{x}_f = [x_f \; y_f]^T$ is a representation of the cartesian coordinates of the two-dimensional video frame f. The image motion model in between f and $f+1$ is then given by 1.

$$\vec{d} = \vec{\delta}(\vec{x}_f) = \begin{bmatrix} d_x \\ d_y \end{bmatrix} \tag{1}$$

The general method involves calculating the minimal eigenvalue for every source image pixel in the image, followed by a non-maxima suppression in $3x3$ neighborhood remain. The features with minimal eigenvalue less than a threshold value are rejected, leaving only stronger features. The basis of the problem is mathematically expressed as finding the A and d that minimizes the standard measure of dissimilarity in 2 which denotes summing over all the image pixels within the patch where $w(x)$ is a weighting function and W represents the window of the given feature patch.

$$\varepsilon = \iint_W [J(Ax + d) - I(x)]^2 w(x) dx \tag{2}$$

We make the following assumptions: the motion in the video corresponds to real world 3D motion projected on the frame, which is the case in our practical application, and the optical flow is the same everywhere, which is reasonable when the patch under consideration is small enough, (2) can be written as (3):

$$J(\vec{d}) = \iint_W [I_1(\vec{x}_f) - I_0(\vec{x}_f + \vec{d})]^2 d\vec{x}_f \tag{3}$$

Linearizing the equation (2) with respect to $\vec{d}$ using the Taylor expansion:

$$I_0(\vec{x}_f + \vec{d}) = I_0(\vec{x}_f) + \vec{g}(\vec{x}_f)^T \vec{d} \tag{4}$$

In (4), $\vec{g}_x(\vec{x}_f)$ and $\vec{g}_y(\vec{x}_f)$ are the derivatives of the frame in x_f and y_f direction at the point $\vec{x}_f$, where $\vec{g}(\vec{x}_f) = [\vec{g}_x(\vec{x}_f)\ \vec{g}_y(\vec{x}_f)]^T$. The dissimilarity that minimizes 3 is the solution of $Z\vec{d} = \vec{e}$, in which $\vec{e} = \iint_W (I_0 - I_1)[g_x\ g_y]^T d\vec{x}_f$ and,

$$Z = \iint_W \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix} d\vec{x}_f \tag{5}$$

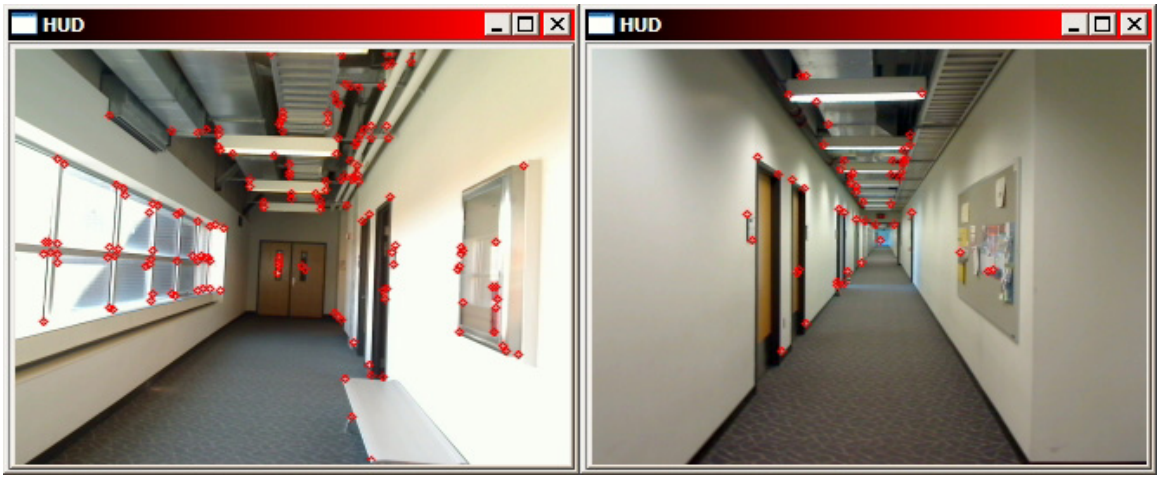

Figure 3. Feature extraction from live video stream using Shi and Tomasi based methods.

Albeit these methods are capable of creating a rich set of features, when landmarks need to be extracted from that set, some pitfalls to its operation appear due to the deceptive nature of vision. For instance, the method will get attracted to a bright spot on a glossy surface, which could be the reflection of ambient lightning, therefore an inconsistent, or deceptive feature. Therefore, a rich set of features does not necessarily mean a set that is capable of yielding the same or compatible results in different statistical trials. In SLAM, a sparse set of reliable landmarks is preferable over a populated set of questionable ones. Authors in[13] present a more recent feature *goodness* measure claimed to be the successor method to Shi and al, in which a variation of the approach is proposed to estimate the size of the tracking procedure convergence region for each feature, based on the Lucas-Kanade tracker[14] performance. The method selects a large number of features based on the criteria set forth by Shi and Tomasi and then removes the ones with small convergence region. Although this improves the consistency of the earlier method, it is still probabilistic and therefore, it cannot make an educated distinction in between a feature and a landmark.

Harris - Stephens - Plessey Corner Detection Algorithm[8] is another method based on the local auto-correlation function of a two-dimensional signal; a measure of the local changes in the signal with small image patches shifted by a small amount in different directions. If a small window is placed over an image, and if that window is placed on a corner-like feature, then if it is moved in any direction there will be a large change in intensity. If the window is over a flat area of the image then there will be no intensity change when the window moves. If the window is over an edge there will only be an intensity change if the window moves in one direction. If the window is over a corner then there will be a change in all directions. Harris method will provide a more sparse, yet stronger and more consistent set of corner-like features due to its immunity to rotation, scale, illumination variation and image noise.

Consider $I_x y$ to be a 2D gray-scale image. Assuming $I(x_i + \Delta x, y_i + \Delta y)$ is the image function, (x_i, y_i) represent the points in the small window W centered on the point (x, y) the auto-correlation function $c(x, y)$

is defined as:

$$c(x,y) = \sum_{W} \left[I(x_i, y_i) - I(x_i + \Delta x, y_i + \Delta y) \right]^2 \tag{6}$$

After the image patch over the area is shifted by (x, y), sum of square difference between (u, v) and (x, y) is calculated and the shifted image is approximated with a 2nd. order Taylor series expansion in 7 cropped to the first order terms, where I_x and I_y are partial derivatives which express the approximation:

$$I(x_i + \Delta x, y_i + \Delta y,) \approx I(x_i y_i) + \left[I_x(x_i y_i) I_y(x_i y_i) \right] \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{7}$$

By substitution of 7 into 6,

$$c(x,y) = \sum_{W} \left(\left[I_x(x_i y_i) I_y(x_i y_i) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \right)^2 \tag{8}$$

$$c(x,y) = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}^T \begin{bmatrix} \sum\limits_{W} \left([I_x(x_i y_i)) \right)^2 & \sum\limits_{W} \left([I_x(x_i y_i) I_y(x_i y_i)) \right) \\ \sum\limits_{W} \left([I_x(x_i y_i) I_y(x_i y_i)) \right) & \sum\limits_{W} \left([I_y(x_i y_i)) \right)^2 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}^T P(x,y) \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{9}$$

where $P(x, y)$ represents the intensity structure of the local neighborhood. Assume the eigenvalues of this matrix are λ_1 and λ_2. Then, the Harris corners are selected as follows:

	Algorithm: Harris Corner Selection Criteria
1	**If**($\lambda_1 \simeq 0$ **AND** $\lambda_2 \simeq 0$)
2	**Then**, There is no feature of interest at this pixel.
3	**If**(eigenvalue[1] $\simeq 0$ AND eigenvalue[2] $\gg 0$)
4	**Then**, An edge is found.
5	**If**($\lambda_1 > 0$ **and** $\lambda_2 > 0$)
6	**Then**, A corner is found.

Although it is not trivial to generalize why one feature extraction method is better than the other, one that can better segregate exceptional features is more preferable. Considering indoor environments of man-made architectures, architectural corners comprise extraordinary landmarks. The level of consistency they can offer for the purpose of understanding the surrounding environment is superior to detecting a larger set of corner-like features. Architectural corners are unlikely to change their position in three dimensional space and the positions of corners can be further exploited to infer about the location and size of surrounding walls, providing rigid points of identification which makes the comparison of features scalable. Harris algorithm is more likely to get attracted to those corners. However, unless architectural corners are clearly contrasting the background (e.g. textured), Shi and Tomasi based approaches have a higher chance of capturing them.

V. Motion and Mapping Estimation

After the corner extraction process is complete for the current set of potential features within the field of view of the video-camera, the next step is to exploit the relative point locations in response to the movement of the MAV and attempt to determine the structure of the elements that may lay ahead, such as walls and corridors. Other relevant information to be extracted in this step includes the speed, direction and degree of motion.

Nevertheless, this set of corners is likely to have redundant features in it, such as corners that do not belong to the boundaries of the architecture, therefore do not produce good landmarks. We are more interested in corners that provide hints about the shape and structure of the architecture the MAV is flying

through, particularly those that indicate a wall or an opening (e.g. door) approximately perpendicular to the pitch axis of the MAV, so it can be ensured that the MAV is always flying at the middle point of the hallway.

A dynamic thresholding of the frame is used to segment lighter foreground objects from their background, which helps normalize the visual changes as a consequence of uneven ambient lighting, and otherwise changing lightning conditions. Considering one each pixel at a time, dynamic thresholding algorithm creates a binary frame as a result of the analysis of each pixel with respect to the mean of its local neighborhood whose outer limits are determined by a window size. The resulting frame is processed using an edge detector to emphasize the architectural lines and suppress noise. The lines are extracted using Hough Transform[9] and sorted according to their slope, length, and intersection properties. Resulting architectural lines reveal the likelihood of detected corners being a landmark.

A. Range Measurement in a Hall Way

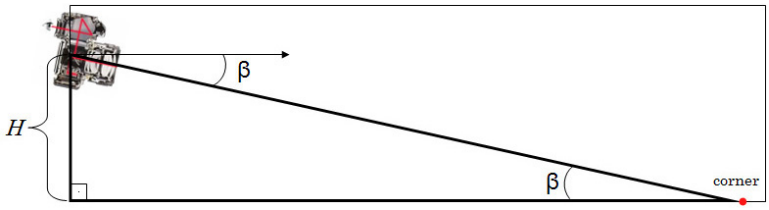

Figure 4. The image shows a conceptual cutaway of the corridor from the left. The angle β represents the angle at which the camera is pointing down.

We introduce a new range and bearing measurement strategy using a monocular camera in this section. We begin the range and bearing measurement by assuming that the height of the camera from the ground, H, is known a priori. Note that H, which equals the altitude of the MAV, can precisely be measured using the ultrasonic altimeter onboard. For another application, where a human carries a helmet-mounted monocular camera, obtaining such height information is trivial. The camera is pointed at the far end of the corridor, but slightly tilted down with an angle β, which is measured by the tilt sensor on the MAV. This incorporation of the downward tilt angle of the camera was inspired by the human perception system, where humans accurately judge the absolute distance by integrating local patches of the ground information into a global surface reference frame.[30] Note that X denotes the distance from the normal of the camera with the ground, to the first detected corner (see Figure 4). The two lines that define the ground plane of the corridor are of particular interest, indicated by blue arrows in Figure 5. By applying successive rotational and translational transformations[27,29] among the camera image frame, the camera frame, and the target corner frame, we can compute the slope angles for these lines, denoted by ϕ in Figure 6.

$$\tan\phi_1 = \frac{H}{W_l \cos\beta} = L_1, \quad \tan\phi_2 = \frac{H}{W_r \cos\beta} = L_2 \tag{10}$$

From (10), we can determine the individual slopes, L_1 and L_2. If the left and right corners coincidentally have the same relative distance X and the orientation of the vehicle is aligned with the corridor, $W_r + W_l$ gives the width of the corridor as shown in Figure 5. Equation (11) shows how these coordinates are obtained for the left side of the hallway.

$$u_L = u_o + \frac{\alpha(W_l)}{\cos\beta x + \sin\beta H} \qquad v_L = v_o + \frac{\cos\beta H - \sin\beta x}{\cos\beta x + \sin\beta H} \tag{11}$$

where (u_L, v_L) and (u_R, v_R) denote the perspective-projected coordinates of the two corners at the left and right side of the corridor. It should be mentioned that we wrap (u, v) with a radial distortion[7] to find a more accurate location of the corners on the image frame. In addition, the ratio α of the camera focal length (f) to the camera pixel size (d) is given by

$$\alpha = \frac{f}{d} \tag{12}$$

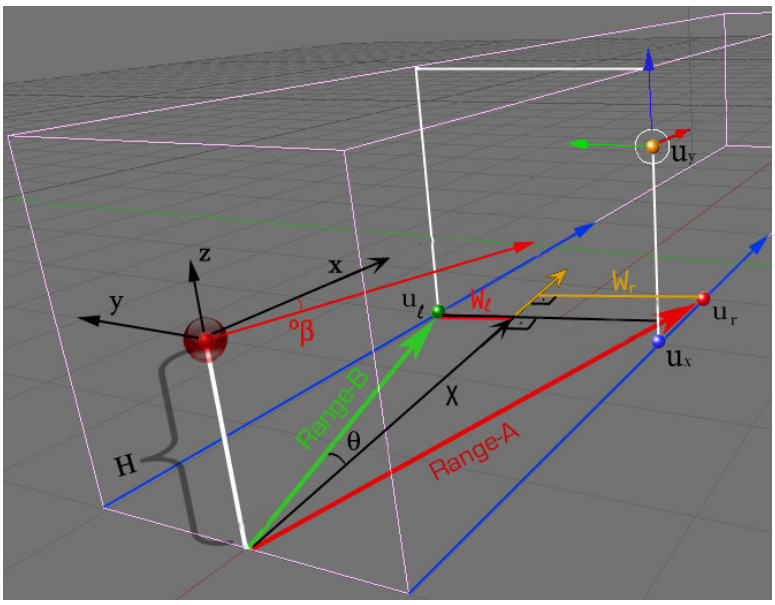

Figure 5. A three dimensional representation of the corridor, and the MAV. *Range-B* $= \sqrt{W_l^2 + X^2}$ represents the range to the landmark u_l, where $\theta = \tan^{-1}(W_l/X)$ is the bearing to that landmark. *Range-A* is range to another independent landmark whose parameters are not shown. At any time there may be multiple such landmarks in question. If by coincidence, two different landmarks on two different walls have the same range, then $W_l + W$ gives the width of the corridor.

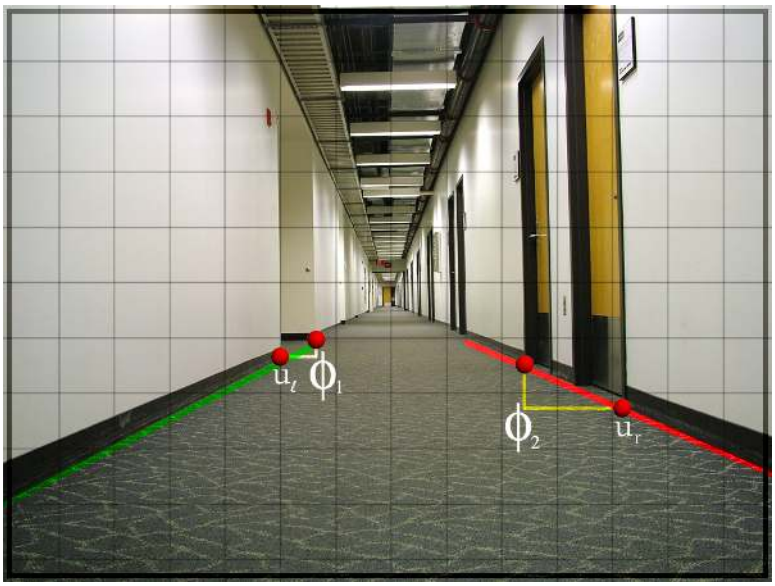

Figure 6. The image plane of the camera.

From the two equations given in (11), we can solve for H in (13).

$$H = \frac{\alpha W_l}{u_L - u_o} \sin\beta + \left(\frac{v_L - v_o}{u_L - u_o}\right) W_l \cos\beta \tag{13}$$

We can rewrite (13) using $\cos\beta = \frac{H}{L_1 W_l}$ from (10).

$$CH = \frac{\alpha W_l}{u_L - u_o} \sin\beta, \quad C = 1 - \frac{v_L - v_o}{u_L - u_o} \frac{1}{L_1} \tag{14}$$

Finally, we solve for the longitudinal distance X and the transverse distance W_l, by combining the preceding equations:

$$\begin{aligned} W_l &= \frac{(u_L - u_o)H}{\alpha} \sqrt{C^2 + \frac{\alpha^2}{(u_L - u_o)^2 L_1^2}} \\ &\text{assuming that, } u_L > u_o \\ \cos\beta &= \frac{H}{W_l L_1} \\ X &= \left(\frac{\alpha W_l}{u_L - u_o} - \sin\beta H\right) \frac{1}{\cos\beta} \end{aligned} \tag{15}$$

The same process can be repeated for any number of corners, including the corners on the other side. In essence, exploiting the certain geometry of the corners present in the corridor, we can compute the absolute range and bearing of the features (corners) needed for the SLAM formulation. Again, this formulation requires only one camera. Results of empirical tests suggest that the preceding equations indeed accurately measure the range and bearing angles of the corner, given that the height of the camera H and the focal ratio α are accurate. Its precision depends of the resolution of the camera frame, i.e., the number of pixels per frame.

B. Corner SLAM formulation

Let us consider one instantaneous field of view of the camera, shown in Figure 7, in which the center of the four corners (shown in red) is shifted. Note that the proposed method currently uses only the ground corners (corners 3 and 4 in the figure), although other corners can be used with their height information. From the distance measurements in (15), we can derive the relative range and bearing of a corner of interest (index i) as follows

$$\mathbf{y}_i = \begin{pmatrix} \sqrt{X_i^2 + W_i^2} \\ \tan^{-1}\left[\frac{\pm W_i}{X_i}\right] \end{pmatrix}, \tag{16}$$

where X_i and W_i of the i-th corner are defined in (15) and Figure 5. This measurement equation can be related with the states of the vehicle and the i-th corner (landmark) at each time stamp (k) as follows

$$\mathbf{y}_i(k) = \mathbf{h}_i(\mathbf{x}(k)) = \begin{pmatrix} \sqrt{(x_r(k) - x_{ci}(k))^2 + (y_r(k) - y_{ci}(k))^2} \\ \tan^{-1}(\frac{y_r(k) - y_{ci}(k)}{x_r(k) - x_{ci}(k)}) - \theta_r(k) \end{pmatrix} + w(k) \tag{17}$$

where $\mathbf{x}_v(k) = (x_r(k), y_r(k), \theta_r(k))^T$ is the vehicle state vector of the 2D vehicle kinematic model, while $x_{ci}(k)$ and y_{ci} represent the x and y coordinates of the i-th landmark (corner). Note that $w(k)$ denotes the measurement noise.

The system states $\mathbf{x}(k)$ consists of the vehicle state vector $\mathbf{x}_v(k)$ and the positions of the corners such as

$$\mathbf{x}(k) = \left(\mathbf{x}_v(k)^T, x_{c1}(k), y_{c1}(k), \cdots, x_{cn}(k), y_{cn}(k)\right)^T \tag{18}$$

where n is the total number of the existing corners in the feature map.

We can also incorporate the three-dimensional vehicle model in lieu of (17). For simplicity, we focus on the two-dimensional car-like vehicle model[17,19] as our vehicle dynamics. Indeed, the dynamic model of MAVs with an autopilot controller, which carries out an altitude hold, resembles a car-like kinematic model.

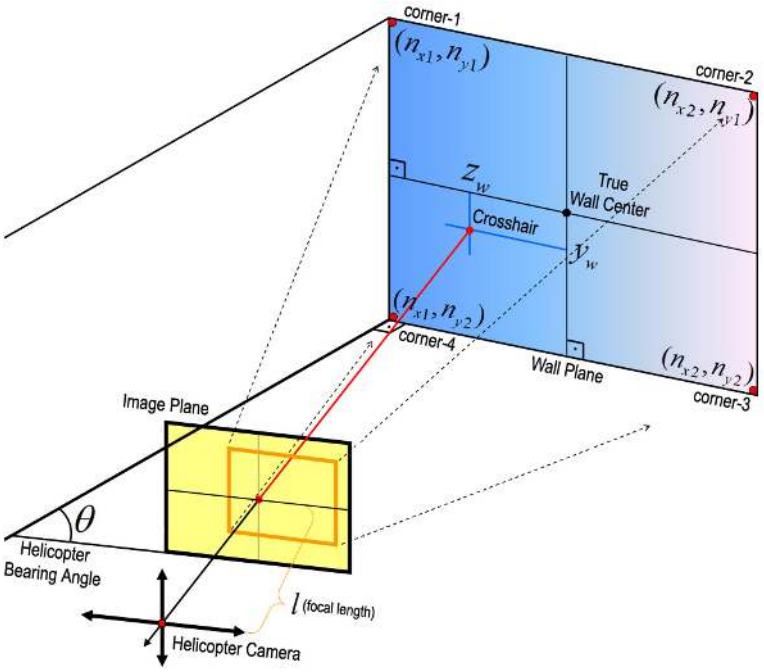

Figure 7. 3D representation of an instantaneous shot of the MAV-camera flying through a corridor towards a wall, with bearing angle θ.

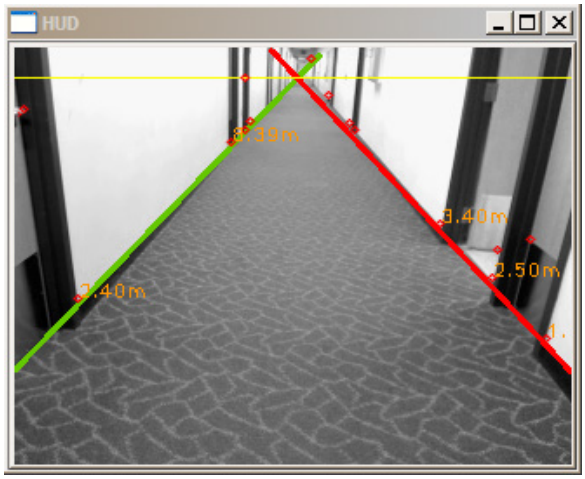

Figure 8. The world as seen from the MAV, with accurate range measurements to landmarks. The red and green lines represent the slope measurements. The yellow line represents the artificial horizon.

Figure 9. The visual radar that displays the MAV and the world map.

The extended Kalman filter (EKF)[17,19] based SLAM formulations can be used to simultaneously estimate the vehicle pose and the location of the corners. In particular, we make use of the compressed EKF SLAM algorithm,[26] that can significantly reduce the computation requirements when the vehicle navigates for a long period of time. A more sophisticated method such as FastSLAM[19] is a subject of the future work. Since the Kalman filters[18] of the autopilot system, which incorporates the IMU/gyroscope and electronic compass measurements, output the heading information, we are only concerned about providing the global metrology system in the absence of the GPS signal. We have to simultaneously locate the landmarks (corners shown in red in Fig 6), as well as the vehicle states x_r, y_r, θ_r described by

$$\mathbf{x}_v(k+1) = \begin{pmatrix} \cos\theta_r(k)u_1(k) + x_r(k) \\ \sin\theta_r(k)u_1(k) + y_r(k) \\ u_2(k) + \theta_r(k) \end{pmatrix} + \gamma(k) \tag{19}$$

where the linearized input signal noise $\gamma(k)$ can be represented by [26]

$$\gamma(k) = \frac{\partial F}{\partial u}|_k \gamma_u(k) + \gamma_f(k) \tag{20}$$

Notice that we have the two control inputs for this vehicle kinematic model: the linear vehicle velocity $u_1(k)$ and the angular rate $u_2(k)$. The low-pass filtered velocity estimates from the Kalman filter can also be used, as in our empirical tests.

The standard EKF routines iterate the prediction step and measurement update step using the Jacobian matrices obtained from (17) and (19). Care must be taken to determine if the detected corners exist and can be associated with the existing corners in the map. An acute component that empowers VINAR is the mechanism that associates range and bearing measurements with landmarks; as a prerequisite for the method to function comme il faut, each measurement must correspond to the correct landmark. The MAV is assumed to take off with uncertainty about its position. The measurements obtained by VINAR are with respect to the location of the MAV, which incrementally becomes the navigation map. VINAR treats new landmarks differently; a new landmark is given a high level of uncertainty, as illustrated in Figure 9, and it has to prove its consistency in order for the uncertainty to decrease. Only then, the landmark is considered eligible to be incorporated into the state vector and consequently becomes a part of the navigation map. Otherwise, the map would be populated with a vast number of high-uncertainty landmarks which presumably do not contribute to SLAM.

This data association problem also decides if a new corner is sufficiently different from the existing ones to warrant a new landmark. For the data association, the measure of the innovation is written as:

$$I_v = \left(\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))\right)^T \mathbf{S}^{-1} \left(\mathbf{y}(k) - \mathbf{h}(\mathbf{x}(k))\right) \tag{21}$$

and the innovation covariance $\mathbf{S}$ is given by

$$\mathbf{S} = \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \mathbf{P} \frac{\partial \mathbf{h}}{\partial \mathbf{x}}^T + \mathbf{R} \tag{22}$$

where $\mathbf{P}$ is the error covariance matrix, and $\mathbf{R}$ is the covariance matrix of the measurement noise. The $\mathbf{S}$ is checked for the two different threshold values, which determine whether to associate with the existing corners or augment as a new corner. In order to improve the computational efficiency, we compute the data association value $\mathbf{S}$ only for the corners in front of the vehicle (within the camera field of view).

The preceding formulation solves the SLAM problem, thereby simultaneously estimating the pose and orientation of the MAV with respect to the corners as well as the location of the corners. Our preliminary experiments with a 2 mega-pixel web camera with $\alpha = 281.49$ achieved accurate range measurement as depicted in Figures 8 and 9.

VI. Experimental Results

As depicted in Figure 9, our VINAR based SLAM correctly locates the corner locations and builds a top-down map of its environment. The red circle with the tangent yellow dot represents the MAV and its heading. Red and blue dots represents the landmarks in which, red landmarks are the first few good ones that

were detected when the mission started. The MAV assumes it is at $(0, 0)$ cartesian coordinates at mission start, and this initial position is marked by four colored pixels around the origin. The maroon and green lines are x and y axes, respectively. The black plot represents the trail of the MAV. Gray lines are virtual laser lines which represent the range in between the MAV and the landmarks. Orange lines represent the doors, or other similar openings. An orange elliptical figure around a landmark represents the uncertainty for that particular landmark with respect to the ellipse axes. A large ellipse axis represents an inconsistent feature in that direction which might have been introduced when external disturbances are present, for example, a person walking in front of the MAV. Our system is robust to such transient disturbances since the corner-like features that might have been introduced by the walking person will have very high uncertainty. The range of the visual radar is adjusted based on the resolution of the camera, such that features at very far distances where the resolution is inadequate for a high quality feature detection, will be disregarded.

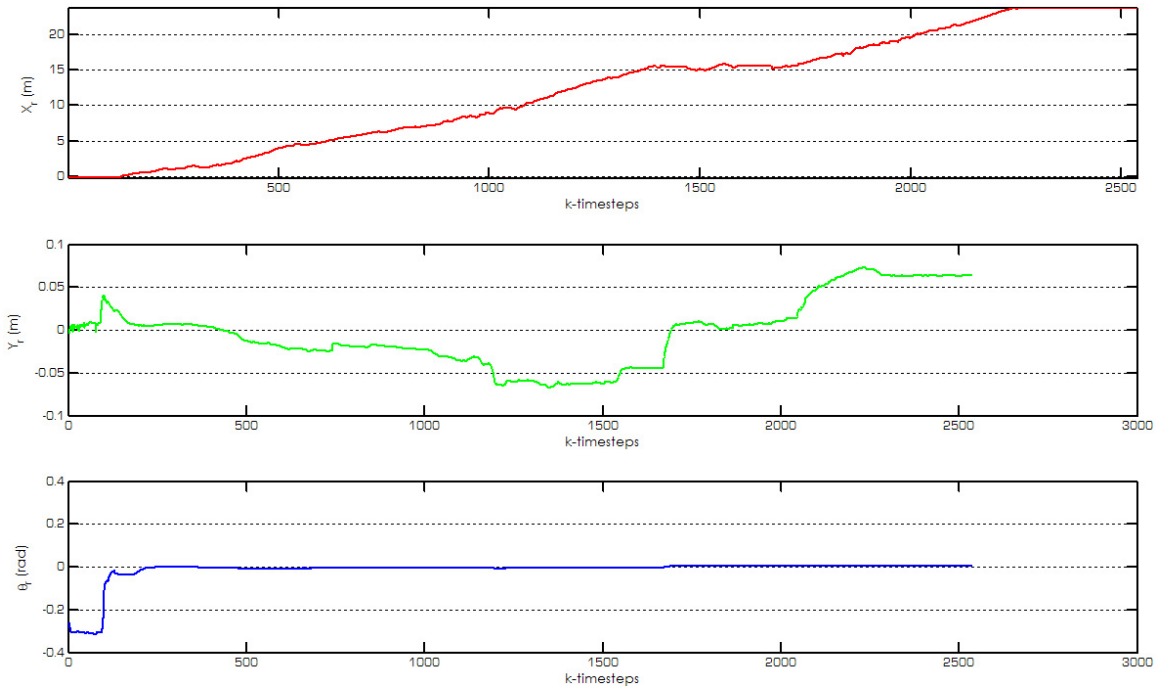

Figure 10. Plots of the system states over time.

Plots of the system states with respect to time in Figure 10 further shed light upon the system behavior in Figure 9. All distance measurements are given in meters, and angular measurements in radians. The *k-timesteps* is the unit of time such that $k = 1/f$, where f represents the frame-rate of the video-camera in frames-per-second.

Acknowledgement

The research reported in this article was in part supported by National Science Foundation (Grant ECCS-0428040), and Information Infrastructure Institute (I^3). The authors thank Dr. David Jensen at Rockwell Collins Inc. for technical discussions and support.

VII. Conclusion

This paper introduced the implementation of a vision based SLAM and navigation strategy for an autonomous indoor MAV, and its experimental validation in a vision based mission through hallways of a building as an indoor airborne navigation and mapping system. Since our system uses a light-weight monocular camera, able to measure ranges to good features, and does not depend on GPS coverage, a practical solution is born for autonomous indoor flight and navigation. Our design is also robust in the sense it does

not depend on extensive feature finalization procedures. Airborne SLAM is still at its infancy, nevertheless its capabilities over conventional sensors stimulates future research. Our system is only limited by the capabilities of the camera and the availability of good corners. Many of the current limitations in airborne SLAM are also governed by the computational power-per-ounce ratio of computers which affects the real-time quality of service of any algorithms involved. This problem can be addressed by by removing the computer from the helicopter and processing video on the ground control center, which brings a limit to the effective range of the aircraft as a compromise.

References

[1]Uijt de Haag, M., Venable, D., and Smearcheck, M., "Use of 3D laser radar for navigation of unmanned aerial and ground vehicles in urban and indoor environments," Proc. of SPIE Vol. 6550, 65500C, 2007.

[2]Vandapel, N., Kuffner, J. and Amidi, O., "Planning 3-D Path Networks in Unstructured Environments," Proc. of International Conference on Robotics and Automation (ICRA), Barcelona, Spain, April 2005.

[3]Kim, S., and Oh, S., "SLAM in Indoor Environments using Omni-directional Vertical and Horizontal Line Features," Journal of Intelligent and Robotic Systems, Vol. 51, Issue 1, pp. 31-43, ISSN:0921-0296, Jan. 2008.

[4]Harati, A., and Siegwart, R., "Orthogonal 3D-SLAM for Indoor Environments Using Right Angle Corners," Proceedings of the IEEE/RSJ Intenational Conference on Intelligent Robots and Systems, IROS, Beijing, China, 2006.

[5]Isoda, N., Terada, K., Oe, S., and IKaida, K., "Improvement of Accuracy for Distance Measurement Method by using Movable CCD," pp. 29-31, Sensing Instrument Control Engineering (SICE), Tokushima, July 29-31 1997.

[6]http://en.wikipedia.org, "Scheimpflug Principle".

[7]Davison, A., Nicholas, M., and Olivier, S., "MonoSLAM: Real-Time Single Camera SLAM," Pattern Analysis and Machine Intelligence (PAMI)(29), No. 6, pp. 1052-1067, 2007.

[8]Harris, C., and Stephens, M., "A combined corner and edge detector," Proc. of the 4th. Alvey Vision Conference, pp. 147-151, 1988.

[9]Duda, R. O., and Hart, P. E., "Use of the Hough Transformation to Detect Lines and Curves in Pictures," Comm. Association for Computing Machinery (ACM), Vol. 15, pp. 1115, Jan. 1972.

[10]http://www.micropilot.com

[11]Kima, J., Sukkariehb, S., "Real-time implementation of airborne inertial-SLAM," Robotics and Autonomous Systems 55, 2007, pp. 6271.

[12]Shi, J., and Tomasi, C., "Good features to track," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 593-600, June 1994.

[13]Zivkovic, Z., and Heijden, F., "Better Features to Track by Estimating the Tracking Convergence Region," IEEE International Conference on Pattern Recognition (ICPR), Vol. 2, p. 20635, 2002.

[14]Lucas, B., and Kanade, T., "An iterative image registration technique with an application to stereo vision," International Joint Conferences on Artificial Intelligence (IJCAI), pp. 674679, 1981.

[15]Yuen, D. C. K., and MacDonald, B. A., "Vision-Based Localization Algorithm Based on Landmark Matching, Triangulation, Reconstruction, and Comparison," IEEE Transactions on Robotics, Vol. 21, No. 2, pp. 217, 2005.

[16]Chekhlov, D., Pupilli, M., Mayol, W., and Calway, A., "Robust Real-Time Visual SLAM Using Scale Prediction and Exemplar Based Feature Description," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-7, 2007.

[17]Choset, H., Lynch, K. M., Hutchinson, S., Kantor, G., Burgard, W., Kavraki, L. E., and Thrun., S., *Principles of Robot Motion - Theory, Algorithms, and Implementations*, The MIT Press.

[18]Kalman, R., "A New Approach to Linear Filtering and Prediction Problems," Transactions of the American Society Of Mechanical Engineers (ASME), Journal of Basic Engineering, Volume 82, Series D, pp 35-45, 1960.

[19]Burgard, W., Thrun, S., and Fox, D., *Probabilistic Robotics*, The MIT Press.

[20]DeSouza, G.N., Kak, A.C., "Vision for Mobile Robot Navigation: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 24, No. 2, pp.237-267, February 2002.

[21]Jensfelt, P., Kragic, D., Folkesson, J., and Bjorkman, M., "A Framework for Vision Based Bearing Only 3D SLAM," Proc. of International Conference on Robotics and Automation (ICRA) 2006, pp.1944-1950, May 15-19, 2006.

[22]Langelaan, J., and Rock, S., "Passive GPS-Free Navigation for Small UAVs," IEEE Aerospace Conf., March 5-12, 2005.

[23]Lemaire, T., S. Lacroix, and J. Sola, "A Practical 3D Bearing-Only SLAM Algorithm," Proc. 2005 IEEE/RSJ IROS 2005, pp.2449-2454, August 2-6, 2005.

[24]Se, S., Lowe, D., and Little, J., "Vision-based Global Localization and Mapping for Mobile Robots," IEEE Trans. on Robotics, Vol. 21, No. 3, pp.364-375, June 2005.

[25]Thrun, S., "Robotic Mapping: A Survey", *Exploring Artificial Intelligence in the New Millenium.* Morgan Kaufmann, 2002.

[26]Guivant, J. E., and Nebot, E. M., "Optimization of the Simultaneous Localization and Map-Building Algorithm for Real-Time Implementation," IEEE Transactions on Robotics and Automation, VOL. 17, NO. 3, 2001.

[27]Boussard, C., Hauti'ere, N., and dAndrea-Novel, B., "Vision Guided by Vehicle Dynamics for Onboard Estimation of the Visibility Range," International Federation of Automatic Control (IFAC) Symposium on Autonomous Vehicles, IAV Sept. 3-5, 2007.

[28]Saxena, A., Schulte, J., and Andrew, Y. Ng, "Depth Estimation using Monocular and Stereo Cues," Proc. of International Joint Conferences on Artificial Intelligence (IJCAI), pp 2197-2203, 2007.

[29]Davis, E. R., *Machine Vision : Theory, Algorithms, Practicalities*, 3rd ed., Morgan Kaufmann, 2004.

[30]Wu, B., Ooi, T. L., and He, Z. J., Perceiving distance accurately by a directional process of integrating ground information, Nature Vol. 428, pp. 73-77, March 2004.