3D Mapping based VSLAM for UAVs

Xiaodong Li, Nabil Aouf and Abdelkrim Nemra

Abstract— **This paper addresses 3D texture mapping in Visual Simultaneous Localization And Mapping (VSLAM) for Unmanned Aerial Vehicle (UAV) applications. Landmark selection strategy based on feature detection methods such as Scale Invariant Feature Transform (SIFT) and Speed Up Robust Features (SURF) is adopted. The selected features are combined with additionally chosen features that are well distributed across the stereo views and refined by RANSAC in order to provide well visualized views for navigation. Experimental results are provided to demonstrate the effectiveness of our 3D mapping strategy.**

I. Introduction

Visual sensing and mapping of the environment is a fundamental issue in Unmanned Aerial Vehicle (UAV) with SLAM [1], [2], [3]. The presence of a textured map is essential for many UAV tasks and can be a powerful tool to provide enriched environmental information for both navigation and visualization. The accurate photometrical 3D model of environment allows researchers to interact with the data acquired during the mission and to understand the spatial distribution and the character of the environmental structure for the guidance of UAVs.

The textured mapping problem in UAV applications is defined as to acquire a spatial model of the UAV's environment, then to plate corresponding textures acquired from the field on the surface of 3D space. With known UAV poses and binocular vision system (cameras), local maps can be registered into a common coordinate frame to create a global map. Unfortunately, the unavoidable consequence of imprecise systems (modelling, camera calibration, INS, etc.) and noise (internal and external) require the UAV's self-localization to be filtered/fused to improve the accuracy of information. This is normally done by a serial of algorithms such as Extended Information Filter (EIF) [4] in our case.

In addition, the challenging properties of the medium and six Degree Of Freedom (Euclidian position and yaw, pitch, roll angles) involved in a UAV's dynamics has largely increased the nonlinearity of system modelling, which lowers the expected robustness. Generally transferring reconstruction methods to terrain environments based on UAV application is simply not a straight forward task. The number and quality of acquired features for texture mapping may not be good enough for enriched visualization. The limited number of correspondences extracted within VSLAM is normally not dense enough to form rich texture mapping for visualization as well. Therefore, in this research, one emphasis is on the reconstruction of sparse distinct terrain features by combining extra pixel points selected across the whole image in order to present a wider view of the scene, hence covering more field information. In addition, the mesh surface for 3D space is to be plated with corresponding texture in 2D image plane to reconstruct a 3D scene.

Xiaodong Li, Nabil Aouf and Abdelkrim Nemra are with Cranfield University, Department of Informatics and Systems Engineering, Shrivenham, SN6 8LA, United Kingdom (e-mail: x.li2@ Cranfield.Ac.uk, N.Aouf@ Cranfield.Ac.uk)

II. 3D Estimation in VSLAM with UAV Application

In VSLAM, texture mapping is the problem of integration/interpretation of the information (spatial and shade) obtained by visual sensors (cameras) into a consistent model and describing that information as a given representation. Spatial 3D reconstruction provides coordinates in 3D space for mapping environment. In our research, the estimated poses of landmarks are obtained by the VSLAM algorithm [12] with an observation model [5] based on the principle of 3D reconstruction in epipolar geometry [13].

The 3D reconstruction in our case is, therefore, under interaction between VSLAM and epipolar geometry [13], where the camera viewpoints are provided by data fusing methods of VSLAM embedded on UAVs. With a space transformation in corresponding reference frames (image → UAV→ world frame) and mapping algorithm employed in stereo camera model [13], both landmarks and UAV 3D coordinate can be reconstructed within VSLAM [12].

A. SIFT/SURF Feature Detection and Matching

Feature extraction and matching in feature based 3D reconstruction is crucial. SIFT [6] and SURF [7] is adopted to provide robust invariance to scale and camera motions in this work. SIFT features are distinctive and constitute a good option for performing reliable matching between different images of the same object or scene from varying UAV poses. It uses scale-space extrema of keypoints in the scale space, and a distinctive descriptor (128 elements) is formed by gradient histogram of that key point.

SURF focuses on corners and blobs as these image characteristics are robust to various image transformations. At the feature detector stage, SURF has largely improved speed performance by using a fast approximation of the Hessian. When it comes to feature description (64 elements), SURF has been tuned towards high recognition rates with Haar Transform, which is a valid alternative to the SIFT descriptor.

By applying SIFT or SURF in both images of a stereo pair features are extracted with their descriptors. Stereo matches can be established by computing the Euclidean distance between feature descriptors found in the left and right images. The 3D location of each match is computed in

the VSLAM systems with a stereo observation camera model [13] based on the pairs of matched pointes.

With the UAV moving, scene features are tracked in subsequent images by feature matching as above in order to deal with photo-consistency in image pairs for the purpose of seamless and smooth texture mapping requirement. New landmarks from these views are aggregated in the augmented feature vector (landmark database in Figure 1) in the same fashion to keep the record of the registered terrain surface model features.

B. RANSAC Outlier Removal

RANSAC (Random Sample Consensus) [8], [9] is a general framework for model fitting in the presence of outliers in image space. It is an iterative method of finding the best model for a set of data by generating a hypothesis from random samples and verifying it to the data. While SIFT/SURF descriptor matching can be quite reliable in many situations, however, certain matching outliers will largely affect 3D triangulation and texture mapping. RANSAC is therefore required to eliminate those outliers.

In each iteration, features are matched temporally by individually comparing each feature descriptor from the pair of images using the Euclidean distance. The procedure with RANSAC runs as following:

1. A group of pair of key points (minimum four feature pairs) is randomly selected as free parameters.
2. Generate a Homography model on sample data points.
3. Test other points against this model.
4. Get inliers as points complying with the model and reject the rest as outliers.
5. Compute average error of all inliers.
6. Re-estimate model with inliers included.
7. Repeat steps 3-6 until error is tolerably small (or non-decreasing) against pre-set threshold.

Putative inliers will be stored in an independent vector to be used with VSLAM later on for both estimation and texture mapping.

C. Extended Information Filter (EIF)

Although *Extended Kalman Filter* (*EKF*) [10] is still valuable and feasible mechanism to implement *SLAM, EIF* [4] is chosen as an alternative for its benefits due to its characteristics suitable for multi-sensors data fusing. Essentially, EIF is a Kalman filter expressed in terms of measures of information about the parameters (states) of interest rather than direct state estimates and their associated covariance [4]. It is also called the inverse covariance form of the Kalman filter. Our Stereo VSLAM system is based on our previous work in [5],[12], where process state vector $x=[X,\ Y,\ Z, U,\ V,\ W,\ \phi,\ \theta,\ \psi]^T$ containing the aerial vehicle position (*X*,*Y*,*Z*), velocity (*U, V, W*), and Euler angles of attitude (φ, θ, ψ) resulting in an Inertial Navigation System (*INS*) which provide internal measurements from IMU (Inertial Measurement Unit) outputs (angular rates and accelerations). Observation model, linking the perceived visual landmarks to the SLAM state vector, is developed based on stereo camera model [13] and the coordinates transformation between binocular cameras embedded in the *UAV* and navigation frame [5]. With these reconstructed 3D landmark coordinates, the system state vector is augmented as $\boldsymbol{x} = [\,x_v\ \ x_m\,]$, where x_v is the UAV state vector defined as,

$$x_V = [X, Y, Z, U, V, W, \phi, \theta, \psi]^T$$

and x_m is the state vector of the observed landmarks given by,

$$x_m = [m_1\ m_2\ m_3 ... m_N]$$

EIF has the following form,

- Prediction

$$\hat{y}(k/k-1) = Y(k/k-1)F(k, \hat{x}(k-1/k-1), u(k-1), (k-1))$$
$$Y(k/k-1) = \left[\nabla F_x(k) Y^{-1}(k-1/k-1) \nabla F_x^{\ T}(k) + Q(k)\right]^{-1}$$

- Estimation

$$\hat{y}(k/k) = \hat{y}(k/k-1) + i(k)$$
$$Y(k/k) = Y(k/k-1) + I(k)$$

Where $\hat{y}(k/k)$ and $\hat{y}(k/k-1)$ are information state vectors, with information state contribution

$$i(k) = \nabla H_x^{\ T}(k) R^{-1}(k) [v(k) + \nabla H_x(k) \hat{x}(k/k-1)]$$

and innovation,

$$v(k) = z(k) - H\ (\hat{x}(k/k-1))$$

$Y(k/k)$ and $Y(k/k-1)$ are information matrices with

$$I(k)\ = \nabla H_x^{\ T}(k) R^{-1}(k) \nabla H_x(k)$$

where $\hat{y}(k/k) = \hat{Y}(k/k)\hat{x}(k/k)$ and $\hat{Y}(k/k) = \hat{P}(k)^{-1}$ are the connection between *EIF* and *EKF*. The prediction of the state variable (output $\hat{y}(k/k), \hat{Y}(k/k), \hat{x}(k/k), \hat{P}(k/k)$) at time instant *k* based on the state variable (input $\hat{y}(k/k-1)$, $\hat{Y}(k/k-1)$, $\hat{x}(k/k-1)$, $\hat{P}(k/k-1)$) at time *k-1* is denoted by subscript *k/k-1*.

Y_k: Information Matrix

y_k: Information state vector

P_k: State covariance

$\hat{x}_k$: State vector

z_k: Current measurement

Q_k: Processing noise covariance matrix

R_k: Measurement noise covariance matrix

The estimation equation of EIF is computationally simpler than EKF and the updating form of EIF is more suitable than EKF for multi-sensor fusion systems.

III. TEXTURE MAPPING WITHIN VSLAM

A realistic appearance of the constructed 3D map is implemented by integrating triangulation and plating of texture within VSLAM. Fig. 1 presents the flowchart which depicts this procedure in our project. It is summarized as following:

A. Feature Detection/Extraction

Binocular system – two calibrated cameras are used to capture UAV environment scenes from different perspectives. Features are extracted by SIFT or SURF crossing multiple images of the same object. These key points (features) can then be registered in 2D coordinate frame.

In order to avoid being a texture less surface map where no features are extracted by SIFT/SURF from the image plane, an additional number of 2D pixel points are selected across the whole image at constant pixel interval. With these extra features, a large-scale of the scene can be covered and texture extracted.

B. Cloud Points Creation for Landmarks

With calibrated onboard stereo cameras and their estimated positions in VSLAM, a set of vertices in 3D space are then generated based on the correspondences obtained through homography transformation from those additional extracted features combined with those SIFT or SURF features extracted in the VSLAM process, which are used as input of inverse observation model [5]. This combination of features is also necessary to produce quasi-dense 3D coordinates for mapping as it is impossible to have dense reconstructed 3D points with SIFT/SURF only due to the real time requirement and limited view of the cameras in VSLAM.

Indeed, as the SIFT/SURF features are extracted mainly from distinct areas, plating texture with these features will result in very limited vision of the field ground, and typically produces unsatisfactory results with the general *3D* reconstruction methods applied to this challenging UAV environment, especially, under the requirement of wide view.

C. Surface Reconstruction

To have textured 3D mapping model based on the sparse cloud points obtained during VSLAM process, the Delaunay triangulation (DT) [11] for the surface interpolation is engaged to take the 3D points and create numerous polygons or faces, namely, a mesh, which can provide a good approximation for the surface structure.

DT is adopted due to its advantage on contiguous, non-overlapping triangles which are as equi-angular as possible. Thus, it reduces potential numerical precision problems due to long skinny triangles. DT is also independent of the order the points that are processed ensuring that any point on the surface is as close as possible to a node. With these advantages, we can then approximate the peaks and valleys of a surface by connecting an irregular point set to construct a mesh of triangles with each satisfying the Delaunay property. Using DT, a limited number of sample points can efficiently produce something that appears natural.

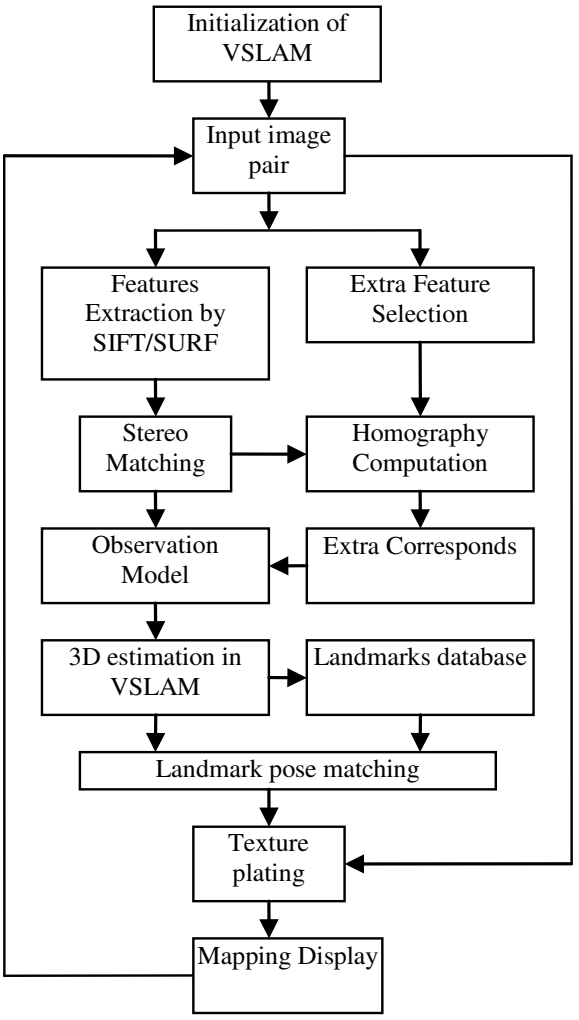

Figure 1. Texture Mapping in VSLAM

D. Texture Mapping/Rendering on Triangle Surface

Texture Mapping/Rendering or Plating is a method for adding surface texture (images) or colour to each face on the surface mesh.

In this research, the texture is image based and Delaunay triangulation is used to create the faces of the objects. A 2D Delaunay triangulation is applied in the x-y plane in the 3D space to have 2D mesh faces. Texture from the real image based on the triangulated feature points of the realistic scene in image plane, which correspond to the 3D points triangulated in 3D space is clipped. These textures are then plated on the 2D mesh faces of 3D space generated earlier, Fig. 2.

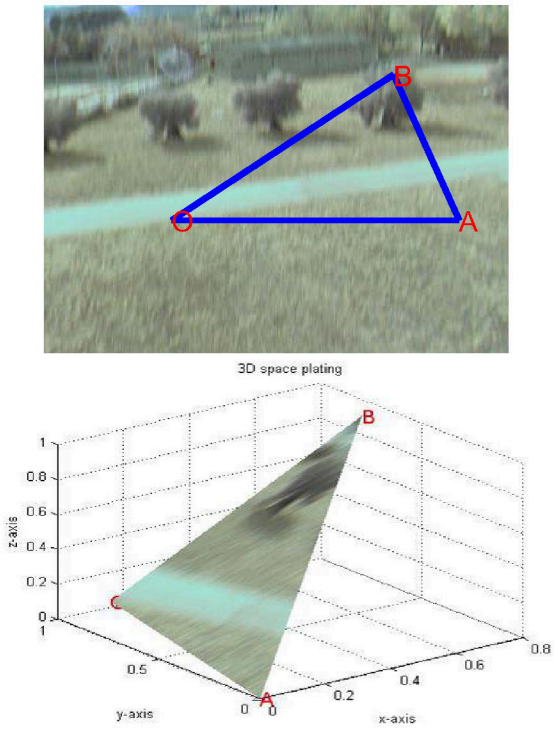

Figure 2. Texure (triangulated area in left image) Plating on 3D Triangle surface with coordinates O:(0.1,0.9,0.1),A(0.0,0.0,0.0), B:(0.8,0.8,0.9)

IV. Experimental Results

In this experiment, 3D reconstruction with texture mapping models is conducted within the procedure of VSLAM. Models are constructed from synchronized images collected using a calibrated onboard binocular camera pair on the UAV together with estimated UAV navigation data. Images are colour with a tiny resolution of (240x320) suffering from the distortion of motion blur caused by the moving flight. This later has side-effect on the feature detection and matching and consequently on 3D triangulated estimation and texture mapping. Figure 3 shows the example of 3D cloud points of landmarks obtained in VSLAM using SURF. Figure 4 and Figure 5 present the 3D mapping results obtained using our mapping strategy using SIFT and SURF feature detectors respectively. These figures show that both SIFT and SURF based 3D mapping provide acceptable results although it appears that the performance obtained using SURF is better. This later is may be due to the stability of the SURF extracted features and the higher matching rate results of SURF as confirmed in the analysis of [14]. The reconstructed visual scenario presents in general a good view of the environment crossed through by the UAV in the air. The reconstructed structure and character of the fields are consistent with the images taken with the onboard stereo cameras.

Figure 6 presents the results of the 3D mapping with SURF extracted features in the VSLAM process and without the additional distributed features/pixels used. As opposite to the results of Figure 5 where these additional distributed features are used, the final 3D mapping in Figure 6 is less informative and of less quality.

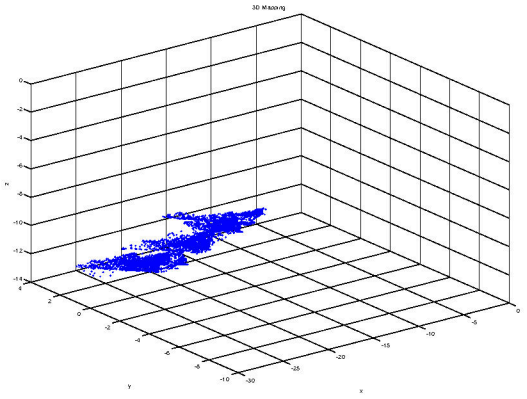

Figure 3. 3D cloud points of landmarks

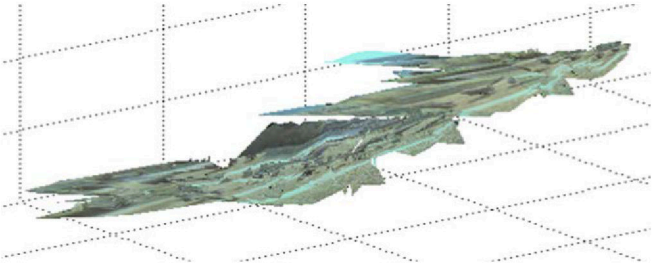

Figure 4. Texture 3D mapping in VSLAM with SIFT(*35 steps + 5181 points +10171 faces*)

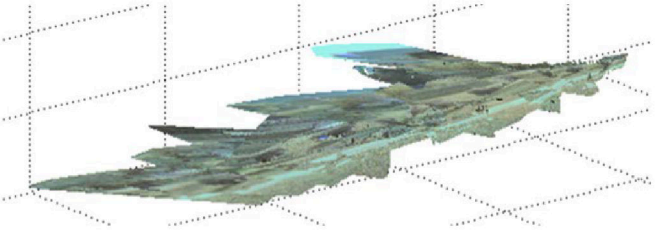

Figure 5. Texture 3D mapping in VSLAM with SURF(*35 steps + 5670 points +11162 faces*)

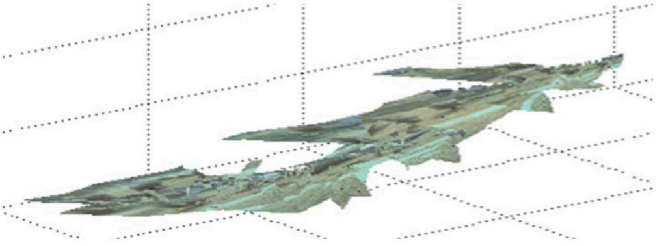

Figure 6. Texture 3D mapping in VSLAM with features extracted from image only(*35 steps + 3741 points +7305 faces*)

V. Conclusion

This paper contributes to the development of 3D texture mapping models to provide a visualized VSLAM results for UAV applications. Our technique computes locations of 3D points in the environment of landmarks based on UAV position estimated with VSLAM in global frame. It takes as input pairs of stereo images obtained by calibrated stereo digital camera. By adding a number of extra pixel points selected across texture image and consitent by pose-

matching, the smothed and extended view of the field is reconstructed. This vivid scenario is produced dynamically and synchronically with the movement of the UAV.

Texture mapping technique presented here provides a convincing performance in VLALM for UAV application with trade-off of sparse 3D points and sufficient estimaton accuracy yielding substantial computational savings to meet the requirement of real time application.

The experimental results are encouraging although the quality of images taken by the UAV platform are not of high resolution and the distortion caused by vibration are always severe, and the estimation errors in VSLAM generally worsen the quality of 3D reconstruction of landmarks.

REFERENCES

[1] M.W.M.G. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte and M. Csorba, "A Solution to the Simultaneous Localization and Map Building (SLAM) Problem," in *IEEE Trans. on Robotics and Automation*, vol:17, pp:229-241, 2001.

[2] H. Durrant-Whyte, "Uncertain geometry in robotics," in *IEEE Trans. Robot and Automation*, Vol4, pp:23-31, 1988.

[3] R. Smith, M. Self, P. Cheeseman. "A stochastic map for uncertain spatial Relationships" in *International Symposium of Robotics Research*, pp: 467–474, 1987.

[4] Arthur G.O. Mutambara, "Decentralized Estimation and Control for Multisensor System", CRC Press LLC, 1998

[5] A. Nemra and N. Aouf "Experimental airborne NH∞ vision-based simultaneous localization and mapping in unknown environments" in *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, Vol:224, no: 12, 2010

[6] D. G. Lowe. "Distinctive image features from scale-invariant keypoints" in *International Journal of Computer Vision*, Vol: 60, pp:91–110, 2004.

[7] Bay H, Tuytelaars T,Van Gool L. "SURF: Speeded up robust features" in *Proceedings of the European Conference on Computer Vision*, Austria, 2006.

[8] Martin A. Fischler and Robert C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography" in *Comm. of the ACM 24*, pp: 381–395 1981.

[9] Elan Dubrofsky "Homography Estimation". MASTER Thesis, University of British Columbia, Canada, 2009

[10] Dan Simon, "Optimal State Estimation", John Wiley & Sons, 2006

[11] Raphael, Kuate. Automatique de maillage de type Delaunay en 3D. Paris, France : universite Pierre et Marie Curie, juin 2005.

[12] Xiaodong Li, Nabil Aouf, "Estimation analysis in VSLAM based on UAV application", Research Report, Department of Informatics and Systems Engineering, Cranfield University, 2012

[13] E Trucco,"Introductory techniques for 3-D computer vision", ISBN-13: 978-0132611084 March 16,1998

[14] Xiaodong Li, Nibal Aouf, "SIFT and SURF Feature Analysis for UAV's Visual Navigation Based on Visible and Infrared Data", research report, Department of Informatics and Systems Engineering, Cranfield University, 2011