Tackling The Scale Factor Issue In A Monocular Visual Odometry Using A 3D City Model

Paul Verlaine Gakne, Kyle O'Keefe

Tackling the Scale Factor Issue in a Monocular Visual Odometry Using a 3D City Model

Paul Verlaine Gakne and Kyle O'Keefe
Position Location and Navigation (PLAN) Group
Department of Geomatics Engineering
Schulich School of Engineering
University of Calgary
Alberta, Canada
Email: {pvgakne, kpgokeef}@ucalgary.ca

BIOGRAPHIES

Paul Verlaine Gakne received his PhD degree in Geomatics Engineering from the University of Calgary, Alberta, Canada. He received his MSc in Communication and Information Systems from Huazhong University of Science and Technology, Wuhan, People's Republic of China and BSc in Telecommunications from École Supérieure Multinationale des Télécommunications, Dakar, Senegal (Cameroon branch). His research focuses on vision and satellite-based navigation for vehicular applications.

Kyle O'Keefe is a Professor of Geomatics Engineering at the University of Calgary, in Calgary, Alberta, Canada. He has worked in positioning and navigation research since 1996 and in satellite navigation since 1998. His major research interests are GNSS system simulation and assessment, space applications of GNSS, carrier phase positioning, and local, indoor, and vehicular navigation with ground based ranging systems and other sensors.

ABSTRACT

Monocular systems are attractive because of their relatively low cost as well as their ease of calibration. However, they suffer of scale ambiguity due to the loss of one dimension when projecting the three-dimensional world onto a two-dimensional image plane. This paper presents a method of resolving the scale ambiguity and drift observed in a monocular camera-based visual odometry by using the slant distance obtained from a skyline matching between the camera and images synthesized using a 3D building model. The obtained visual odometry outputs are then combined with the solutions obtained from the skyline-based positioning for vehicular applications in Global Navigation Satellite Systems-denied/harsh environments such as deep urban canyons. Experiments conducted in downtown Calgary have shown the advantage of correcting the scale factor resulting in a 90% improvement in position solution compared to not correcting the scale drift suggesting the potential of the proposed method for critical applications such as autonomous driving or driver-assistance systems in areas where the 3D building model is available.

Figure 1: Navigating robot (credit: jpl.nasa.gov)

1 INTRODUCTION

In open sky environment with good satellite visibility, the Global Navigation Satellite Systems (GNSS) provide seamless and accurate solutions. However, their performance significantly degrades in harsh environments such as urban canyons. To overcome this limitation, various integration strategies that combine GNSS with other sensors (e.g., inertial navigation systems – INS, odometers, radar, barometer, LiDAR, altimeter, cameras etc.) to ensure satisfactory position solutions in terms of accuracy, availability and continuity in almost all environments have been thoroughly studied with more or less success. In general, the use of camera systems (thanks to the concept of the Visual Odometry – VO) is able to provide an accurate position estimation that can in turn be integrated with solutions obtained from other sensors.

In fact, the VO aims at recovering the motion (position and orientation) of a platform (e.g., vehicle, robot as

shown in Figure 1) by exploiting the images captured by a camera rigidly attached to the platform. Both monocular (i.e., one camera) and stereo (i.e., two cameras) systems can be used for VO implementation. Vision-based positioning for pedestrian (Marouane, Gutschale and Linnhoff-Popien, 2018) or vehicular applications (Aumayer, 2016) has been previously studied by exploiting various concepts such as the rigid-body motion (Ma, et al. 2003), and the use of feature points (Lowe 2004). When more than one camera is used, the VO outputs both the (change in) rotation and the true translation magnitude directly by triangulating the feature points and obtaining the depth information. However, with a single camera system, the scale cannot be directly measured. This means that the translation of the platform can be estimated accurately but up to a scale. This is one of the primary sources of error in VO-based (change of) position estimation and the main bottleneck that prevents monocular-based VO from reaching accuracies comparable to stereo-based VO. The scale factor problem in monocular VO is illustrated in Figure 2 (results for a controlled indoor area). It is clear that the VO provides an accurate solution but up to a scale only (causing the mismatch observed at the "End").

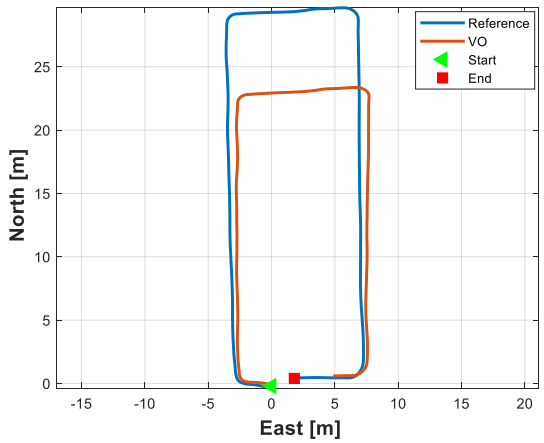


Figure 2: Illustration of the scale factor issue in monocular visual odometry (adapted from Gakne and O'Keefe (2017))

To resolve the scale ambiguity and drift, the general trend is to fuse the image measurements with other sensors. Specifically, Inertial Navigation Systems (INS) are widely used (Spaenlehauer, et al. (2017); Ji and Sanjiv (2015); Gabriel, et al. (2011)) as well as GNSS sensors (Soloviev and Venable (2010); Lim, et al. (2017); Gakne and O'Keefe (2018)).

Spaenlehauer, et al. (2017) combined inertial measurements with monocular odometry in a loosely coupled approach. Regarding their visual odometry, the tracking stage of the ORB-SLAM algorithm (Mur-Artal, Montiel and Tardos, 2015) was employed while the Euler forward integration was used for the inertial measurement processing. The main advantage of using inertial sensors is their high measurement rate that can be used to rapidly compute the pose estimate. The scale factor is computed as the ratio of $\boldsymbol{L}^2$–norm (Horn and Johnson, 1990) of translation vectors of the camera position given by the integration of INS measurements in the world coordinate frame and the $\boldsymbol{L}^2$–norm of the translation vectors of the camera position given by the VO.

Gabriel, et al. (2011) fused the visual and inertial data to determine the scale factor unobserved in the visual framework. Two approaches were used to estimate the scale factor in a monocular visual Simultaneous Localization and Mapping (SLAM). The first method consisted of making use of the spline fitting task (Jung and Taylor, 2011) and the second made use of a multi-rate Extended Kalman Filter (EKF) that was embedded in a Parallel Tracking and Mapping (PTAM – Klein and Murray, 2007). Both approaches provided accurate results at reasonable processing times.

Ji and Sanjiv (2015) proposed a method aiming at operating aircraft in GPS-denied environments using visual odometry (with a camera pointed downward) and a high-accuracy INS. It was observed that the noise contained in the INS measurements can significantly affect the platform motion estimation resulting drift in the aircraft position estimate. To reduce the position estimation error, their method parametrizes feature points with their depth direction perpendicular to the ground. The depth of feature points is obtained in two different ways (by using an altimeter and a 2D laser). This leads to a slower drift because the position error resulting from the INS orientation noise is partially cancel by this process.

Although methods employing inertial sensors are effective and have advantages of lightweight, high measurement rate, immunity to RF interference etc. they are subject to the inertial sensors errors. INS drift rapidly when they do not get updates from other sensors such as GNSS. Other research focused on directly combining GNSS with vision to solve for the scale factor.

Soloviev and Venable (2010) resolved the scale factor ambiguity by integrating the Global Positioning System (GPS) carrier phase measurements with a vision system. Their method consisted of integrating GPS/vision in order to estimate the position as well as the orientation changes of the camera. The integration is realized by combining GPS carrier phase measurements with feature points that are extracted from images. The state relating the position vector with the feature point ranges is first defined, then the changes in the carrier phase measurements between two consecutive images are defined as function of the position change vector. The combination of these two terms allowed to unambiguously resolve the position changes, as well as the range estimates. A similar approach was presented in Gakne and O'Keefe (2018) but the pseudorange measurements are used in lieu of the carrier phase.

Lim, et al. (2017) proposed a method for augmenting GPS with a monocular camera for accurate and reliable positioning by combining the GPS measurements and the relative attitude obtained from the vision system (by the use of the vanishing point). They show that their method performs well in various environments and has results in better solutions than loosely coupled GPS/INS integration approach.

However, in urban environments, GNSS measurements are subject to multiple source of error (mainly multipath) that can lead to position errors up to hundreds of meters. In the Vision/INS case, the scale factor correction can be heavily degraded. For this reason, some researchers have integrated GNSS/Vision/INS or more sensors.

Chu, et al. (2012) integrated a monocular camera measurement with IMU and GNSS for land vehicle navigation in harsh environments based on an EKF. In this case, the translation magnitude is determined from GNSS measurements using differential GNSS technique. The acquired image data is synchronized with GNSS and the baseline of the position change between two GNSS epochs coincides with the camera translation magnitude.

Ben Afia, Escher and Macabiau (2015) integrated even more sensors namely GNSS, IMU, vision and wheel speed sensor (WSS) for vehicular applications. In their implementation the INS estimation errors are corrected by the mean of GNSS, vision and WSS as well as motion constraints for land vehicles (using an error state EKF based on a closed loop configuration). The inertial measurement scale factor was modelled as a Gauss-Markov process.

Finally, another path of research has dealt with the monocular scale factor ambiguity without adding any other sensors by defining the road as a plane and using the vehicle height to obtain the scale information (Choi, Park and Yu (2013); Kitt, et al. (2011)). However, these approaches still contain errors because the assumption that the road is a plane is only true up to an extent.

The objective of this paper is twofold: (*i*). to combine the information obtained from a 3D building model by matching synthesized 3D building model images and camera images and also with visual odometry; (*ii*). To calculate the scale factor from the 3D building model to increase the translation magnitude accuracy for monocular VO.

The remainder of this paper is organized as follows: Section 2 briefly presents the existing works related to this paper; Section 3 details the methodology employed in this paper. Section 4 presents the experiments, results and analysis. Finally, Section 5 concludes the paper.

2 RELATED WORK

A 3D city model can be defined as a digital representation of areas that depicts buildings, terrain surfaces and other features of a given city. Nowadays, digital representation of most of the major cities around the world is available whether commercially (e.g., from 3dcadbrowser.com) or freely (e.g., google.com/earth). Several researchers have made prior attempts to combine the VO with a 3D building model and/or with other sensors for navigation purposes. This works can be classified into several categories.

The first category uses the vision sensor alone to compute the scale factor by making assumptions regarding the camera height relative to the ground, defined as a plane.

Kitt, et al. (2011) compensated the scale factor drift in a monocular VO by making use of constraints and using the knowledge of the camera height (assumed constant during the data collection) as well as by making assumptions about the environment and the camera orientation. The road is assumed to be planar and the camera pitch and roll equal to zero. The tracked feature points are assumed lying on the road plane and then the translation's magnitude is obtained by considering the height of the camera above the ground. A similar approach has been used later by Zhang, Singh and Kantor (2012); Choi, Park and Yu (2013); Song, Chandraker and Guest (2016). However, it is obvious that in real-world scenarios, the road is not necessarily planar and the pitch and roll are non-zero in high dynamic maneuver scenarios for example (which can be a typical case for vehicular applications). Moreover, in the high dynamics case, even the height of the camera can significantly vary. However, it should be acknowledged that errors caused by such assumptions can be smaller for robots driven in indoor environment scenarios for example. Aqel, et al. (2017) proposed a method dealing with the camera height variation in order to improve the accuracy of the scale factor estimation. A downward pointing camera installed under the vehicle was used. Their method consists of setting reference points on the images by the mean of two laser pointers. The scale variation is then estimated by monitoring the changes of distance between the two reference points after obtaining the height variation of the vehicle.

Manolis and Xenophon (2013) also note that measuring and/or estimating the height of the vehicle in order to estimate the scale factor is not sufficiently accurate and proposed three alternative techniques for estimating the translation magnitude pertaining to a three-dimensional reconstruction (two for stereo systems and one for monocular systems). In their approach, the camera pose is estimated from a single image by matching sets of 2D and 3D feature points and making the pose estimation refinement iteratively. To do this, they embedded a P3P solver (Xiao-Shan, et al. 2003) into the random sample consensus (RANSAC – Fischer and Bolles (1981)) framework.

Gräter, Schwarze and Lauer (2015) associated the vanishing point (VP) concept with the road definition as a plane (estimated using structure from motion techniques) to correct the scale drift for advanced driver assistance systems. The proposed method fits the reconstructed feature points and VPs and uses the least-squares optimization to refine the plane and obtain the scale.

A second group of research deals with the scale estimation in monocular systems by combining the vision system with GNSS. Gakne (2018); Gakne and O'Keefe (2018) modelled the scale in a tightly-coupled GNSS/Vision system and show that the scale factor drift can be reasonably modelled with a Gaussian random walk. The scale was observed from the previous and actual position estimates from VO. Li, et al. (2018) later proposed an integration of GNSS and a monocular system for SLAM. Their approach fuses the measurements of each system using an optimization-based scheme. It

outputs the absolute position and attitude of the driven car as well as the translation's scale factor.

The fusion of inertial systems and vision can be classified as the third research path for scale estimation in monocular applications. This is presented in Gabriel, et al. (2011) as described in the previous section. Weiss and Siegwart (2011) combined a monocular vision system with an inertial sensor equipped with a three-axis gyroscope and accelerometer. Their algorithm is built to operate in real time and is independent of the vision algorithm that is used to estimate the camera poses instead.

The fourth category includes works where more than two sensors are added. Ben Afia, Escher and Macabiau (2015) integrated the GNSS, IMU, monocular vision and the wheel speed sensors for navigation in urban canyons. The WSS velocity was particularly affected by an unknown scale that was then modelled as a constant (scale factor error). Won, et al. (2014) evaluated the navigation performance of GNSS/INS/vision integration simulated data, where the number of observed satellites was decreased from three to one. Their method directly outputs absolute (scaled) solutions.

It is also worth mentioning here that a stereo system combined with GNSS sensors that allow direct observation of the scale factor has been extensively studied as presented in Fei, Yashar and Yang (2015) and Aumayer (2016). Also, RGB-D cameras have been used for direct scale determination such as presented in Mur-Artal, Montiel and Tardos, (2015).

Finally, any other sensor or digital map that can provide useful information can be combined with a camera in order to address the scale factor issue. Zhang, et al. (2018) proposed an integration strategy that fuses a monocular visual SLAM with a 1D-laser range finder to obtain the scale estimation and the drift correction. An analytical feasibility for estimating the scale factor by fusing the vision system outputs (derived based on the local dense reconstruction of image sequences) and the laser information is described in detail. Kai, et al. (2014) proposed an integrated monocular VO/laser system that corrects the scale factor drift for astronaut navigation. The laser distance/range finder provides a distance to a single point that can be used for 3D scene reconstruction or scale estimation. The work presented in this paper is close to this category but instead uses an existing 3D city model to obtain the distance from a driven vehicle to a point.

One aspect presented in this paper is the skyline-based positioning. It consists of matching images obtained from a camera with images synthesized from a 3D city model. This technique has been presented in our previous works (Petovello and He, 2016; Gakne and O'Keefe, 2017). In this paper particular attention is given to a combination of a VO (realized from a special setup with a sky-pointing camera) and an existing 3D building model. The next section provides details on the method used for the scale factor drift correction proposed in this paper.

3 METHODOLOGY

This section elaborates the proposed method for scale factor estimation and drift correction. This is principally based on the use of a 3D city model. The proposed work is mainly subdivided in two subsections: the visual odometry and the skyline-based positioning.

3.1 Monocular Visual Odometry

VO can be defined as the computation of the camera motion from monocular or stereo image sequences. Herein, a feature points-based VO is developed from image sequences from an upward-facing camera and follows the steps bellow:

- Image acquisition and correction: this step consists of acquiring images from a monocular camera and rectifying the image distortion via the calibration process. The calibration process also allows the determination of the camera intrinsic parameters such as the focal length, the principal point coordinates and the skew coefficient between the image axes;
- Feature detection, description and matching: feature points are detected and described by using the Oriented FAST and Rotated BRIEF (ORB – Rublee, et al., 2011). The description of this algorithm as used in this paper is presented in Gakne and O'Keefe (2018);
- Motion estimation: from the tracked feature points and their displacement on the images, the relative motion of the camera is computed.

An example of feature detection, matching and outlier rejection is given in Figure 3 where the red circles represent feature points that are detected but not used (outliers); Green circles are those that are detected properly matched between consecutive image frames. The established matches are illustrated by the green lines.

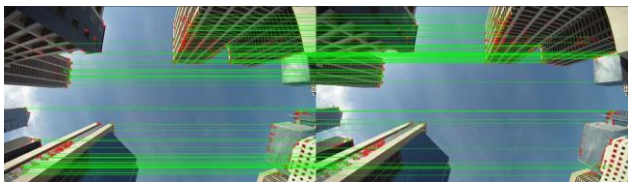


Figure 3: Feature point detection, description (ORB algorithm), matching and outlier removal (RANSAC). Left: image frame at time t; Right: image frame at time $t + \Delta t$.

Given two sets of M feature points represented by $\mathbf{f}_{\mathbf{p}_1} = \{\mathbf{p}_0, \mathbf{p}_1, ..., \mathbf{p}_{M-1}\}$ and $\mathbf{f}_{\mathbf{p}_2} = \{\mathbf{p}_0', \mathbf{p}_1', ..., \mathbf{p}_{M-1}'\}$, the rotation $\mathbf{r}_c$ and the translation $\mathbf{t}_c$ can be determined using least-squares and the singular value decomposition (SVD) such as we have:

$$(\mathbf{r}_c, \mathbf{t}_c) = \arg\min_{R_c, T_c} \sum_{i=0}^{M-1} \omega_i \left\| (R_c \mathbf{p}_i + T_c) - \mathbf{p}_i' \right\|^2 \quad (1)$$

where $\omega_i > 0$ is the weight of each feature point pair.

The step-by-step VO algorithm formulation is given in Gakne (2018) and summarized in Figure 4.

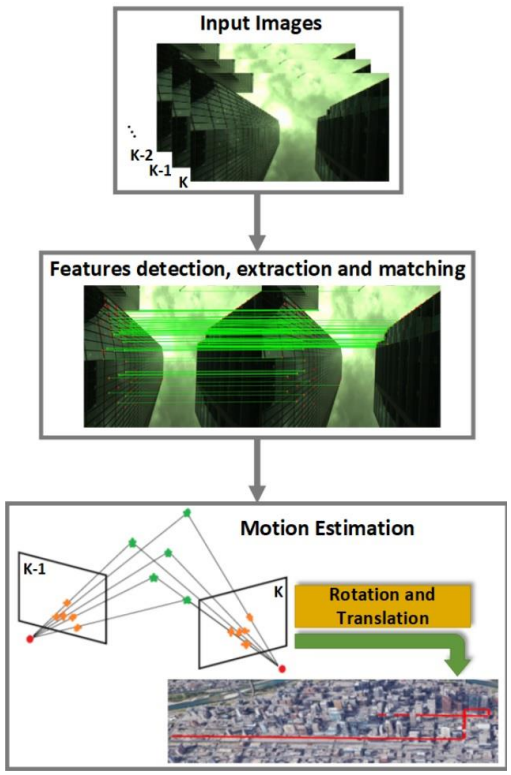


Figure 4: Feature points-based visual odometry

The first box corresponds to the image acquisition and rectification. The second box depicts the feature points detection, description and matching. Finally, the third box shows the rotation and translation estimation and the (change in) position of the camera rigidly mounted on the vehicle.

Having explained the visual odometry, the next section presents the 3D building model based positioning.

3.2 3D Building Model-based Positioning

A skyline is literally defined as an outline of land and buildings defined against the sky. If accurately generated, such information can be used as a fingerprint that can uniquely describe a given city/area. The skyline-based positioning used in this paper is presented in detail in Gakne and O'Keefe (2017) and summarized in Figure 5. The localization problem in this case is solved by using the skylines obtained from a 3D city model and from an upward-facing camera.

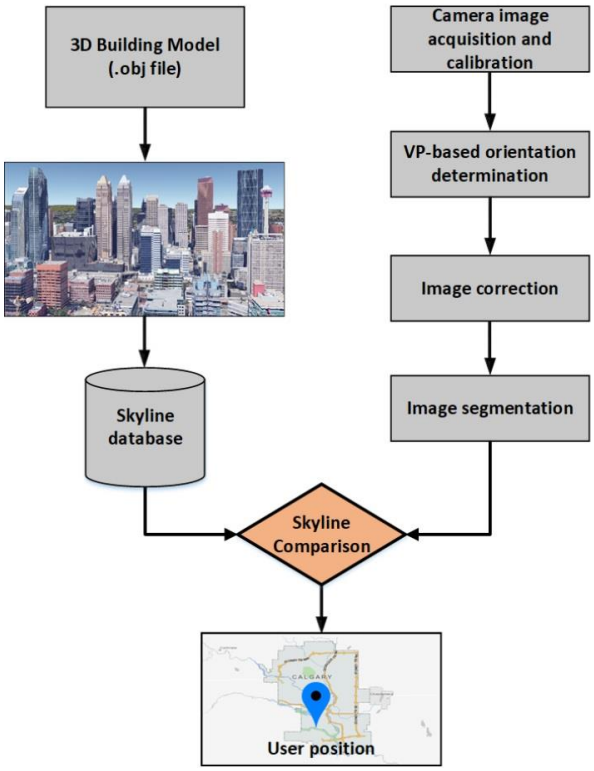


Figure 5: Skyline-based positioning flowchart

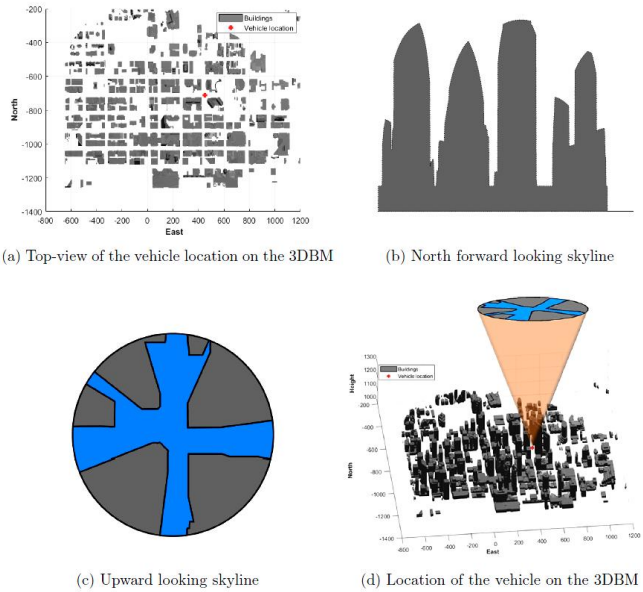
A visual spectrum narrow field-of-view camera similar to that available on most mobile phones is used. The images are then segmented into sky and non-sky areas. The obtained binary images are compared with the ideal images synthetized from the 3D building model (stored in a database). The position solution estimate corresponds to the best match obtained between the observed and synthetized images.

At each vehicle location, the skyline is computed from the 3D building model (see Figure 6) over a range of azimuth. The skyline is defined as:

$$\varsigma_p = \left\{ \left(\boldsymbol{\varepsilon}_{p,j}, \mathbf{h}_{p,j} \right) \right\}, \; j = 0..N-1 \quad (2)$$

where j and $N = 720$ are the azimuth index and number respectively (i.e., an azimuth resolution of 0.5° is used). $\boldsymbol{\varepsilon}$ is the highest elevation angle of the obstructing surface and $\mathbf{h}$ the corresponding height.

One of the main challenges of the skyline-based positioning is to accurately segment the camera image. The difficulty arises because of the lighting variations, weather conditions and the facade of the buildings (e.g., buildings with glass). In this work, the Flood-fill algorithm described in Gakne (2018) is used for the camera image segmentation. Readers can refer to the provided reference for details the image synthesization from the 3D city model.



(a) Top-view of the vehicle location on the 3DBM (b) North forward looking skyline

(c) Upward looking skyline (d) Location of the vehicle on the 3DBM

Figure 6: Location of the vehicle indicated on the 3DBM as well as the ideal skyline synthesized forward and upward skylines at the same location

The comparison/matching between the camera images and the database images (obtained from the 3D building model) is done by the mean of a similarity metric. Herein, the cross-correlation coefficient ($\mathbf{CC}$) is used. It is defined as:

$$\mathbf{CC}_{\mathbf{I}_{b(p)}}(u,v) = \frac{1}{n}\sum_{u,v}\left(\left[\mathbf{I}_b\right]_{cam}(u,v) \circ \left[\mathbf{I}_b\right]_{3D_{(p)}}(u,v)\right) \quad (3)$$

where n is the total number of pixel per image; p is the database image's position; $\circ$ is the Hadamard product of two matrices; $\mathbf{CC}_{\mathbf{I}_b}$ is the cross-correlation coefficient of the binary images; $\left[\mathbf{I}_b\right]_{cam}$ represents the binary/segmented image obtained from the camera; $\left[\mathbf{I}_b\right]_{3D}$ is the binary image synthesized from the 3D city model.

The similarity metric obtained as in Figure 7 gives the location of the vehicle on the travelled path. In order to improve the matching accuracy between the observed (camera) and the 3D building model images the vanishing points have been used to determine the camera pitch and roll and to rectify images accordingly as in Gallagher, (2005) and detailed in Gakne and O'Keefe (2017). In order to be succinct, this process will not be repeated in this paper.

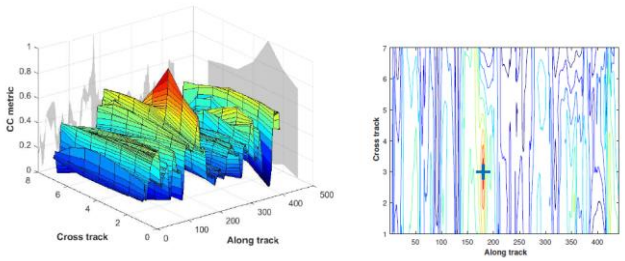


Figure 7: Similarity metric indicating the location of the vehicle on the road

To summarize, the skyline-based positioning follows the three main steps bellow (Figure 5):

- Generate the georeferenced ideal skylines from the 3D building model to populate a database;
- Collect and segment the camera images (to obtain the observed skyline);
- Compare/match the camera images with the ideal images to compute the vehicle position.

Having the skyline defined as in Equation (2) the next step is to compute the slant distance that will be used to compute the scale factor.

3.3 Slant Distance and Scale Factor Computation

In this paper the highest point in view from the vehicle location is chosen and the slant distance to this point from the vehicle is computed. From Equation (2), the slant distance is computed as:

$$\mathbf{d}_{\text{slant}} = \frac{\mathbf{h}}{\sin(\boldsymbol{\varepsilon})} \quad (4)$$

where the parameters $\mathbf{h}$ and $\boldsymbol{\varepsilon}$ are defined as in Equation (2). The computation of these parameters follows that presented in Petovello and He (2015) as implemented in Gakne (2018).

The slant distance computed in this way is similar to the distance obtained from LiDAR data for example. As such, if carefully used, this information can be used to compute the scale factor ambiguity observed from a monocular system. The global scale factor is thus computed as:

$$\mathbf{s}_g = \frac{\mathbf{d}_{\text{slant}}}{\left\|\mathbf{P}_{\text{3DBM}} - \mathbf{P}_{\text{VO}}\right\|} \quad (5)$$

where $\mathbf{P}_{\text{VO}}$ stands for the previous position obtained from the visual odometry and $\mathbf{P}_{\text{3DBM}}$ represents the position obtained from the 3D building model.

Whenever the 3D building model solution is available, $\mathbf{s}_g$ is used to correct the scale drift that is introduced in the relative scale factor over a certain number of image

frames. The relative scale factor is estimated based in the actual and previous estimated platform position (Gakne and O'Keefe, 2018). This is given as:

$$\hat{s}_k = \sqrt{\left(\hat{x}_k - \hat{x}_{k-1}\right)^2 + \left(\hat{y}_k - \hat{y}_{k-1}\right)^2 + \left(\hat{z}_k - \hat{z}_{k-1}\right)^2} \quad (6)$$

where $\left(\hat{x}, \hat{y}, \hat{z}\right)$ represent the estimated platform's position and k is the image frame number.

The steps for correcting the scale factor drift are given as follow:

- Compute the scale factor as given in Equation (6) from frame-to-frame;
- Check if the 3D building model solution is available. If yes, compute the slant distance as defined in Equation (4) then use this information to compute the global scale factor as in Equation (5)

These steps are summarized in the Algorithm 1.

Algorithm 1 Scale factor computation and correction

Input: ς

Output: s

Initialization: $s_r \leftarrow 1,\ s \leftarrow 1$

for $(k \leftarrow 0 : \texttt{NumberFrame} - 1)\{$

$$\hat{s}_r = \sqrt{\left(\hat{x}_k - \hat{x}_{k-1}\right)^2 + \left(\hat{y}_k - \hat{y}_{k-1}\right)^2 + \left(\hat{z}_k - \hat{z}_{k-1}\right)^2};$$

if (3DBM_Solution_available) $\{$

$$\mathbf{d}_{\text{slant}} = \frac{\mathbf{h}}{\sin(\boldsymbol{\varepsilon})};$$

$$\mathbf{s}_g = \frac{\mathbf{d}_{\text{slant}}}{\left\|\mathbf{P}_{\text{3DBM}} - \mathbf{P}_{\text{VO}}\right\|};$$

$$\hat{s}_r \leftarrow \mathbf{s}_g;$$

$\}$

$\}$

$$s \leftarrow \hat{s}_r;$$

return s

4 EXPERIMENT AND RESULTS

The performance of the proposed scale drift correction is evaluated in this section.

4.1 Experiment Setup

An experiment consisting on a monocular system rigidly mounted on the top of the car driven in downtown Calgary, AB, Canada was conducted to test this other method (and is the same as used in Gakne and O'Keefe (2018)). A sky-pointing camera and a reference system (a SPAN LCI from NovAtel) are both mounted on the vehicle roof as depicted in Figure 8.

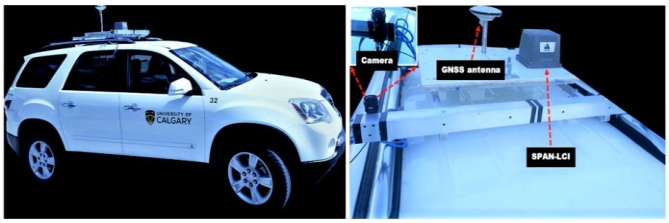


Figure 8: Experiment Setup. Left: the vehicle; Right: top view of the vehicle

4.2 Results and Analysis

With this experimental data, the performance of VO when aided by a 3D building model can now be compared to the case where the building model is not available. For the comparison, the scale factor is first computed as in Equation (6) after estimating the current position for every system (reference, VO and VO aided 3DBM). Then, the scale factor error is computed by subtracting each from the reference scale factor. The time series results are depicted in Figure 9.

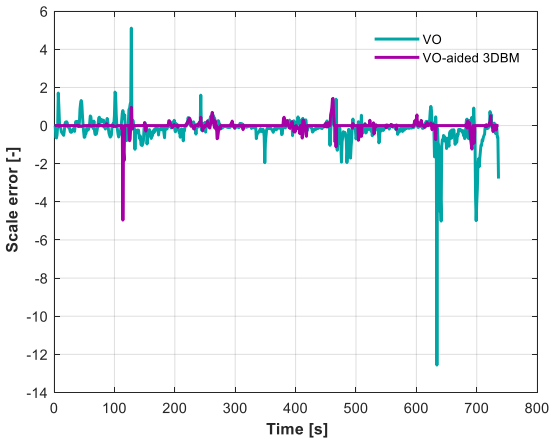


Figure 9: Scale errors

The scale factor error obtained from the VO is overall noisier and larger than the VO aided by the 3DBM. However, it is still clear that there are peaks observed in the aided VO scale. These arise when the 3DBM solution is not available or significantly degraded.
The final position output by each system (VO aided by the 3D building model and the pure VO) is depicted in Figure 10.

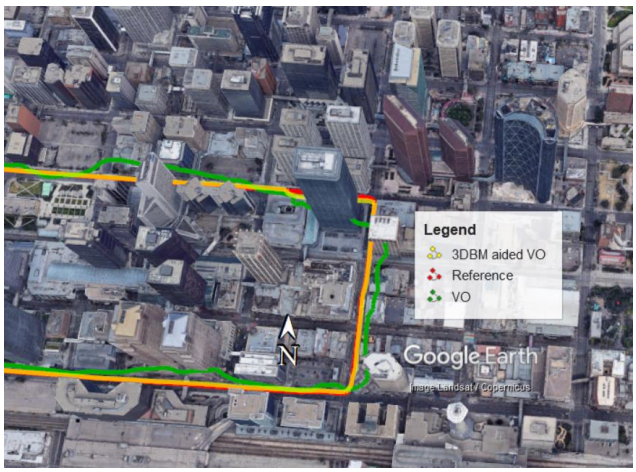


Figure 10: Trajectories comparison (3DBM stands for 3D building model)

The figure shows that the trajectory obtained from the VO aided by the 3D city model closely follows the reference solution in terms of scale except for areas where the 3D building model solution is not available for long time and/or in areas where buildings are missing in the 3D model (especially the upper part of Figure 10. The 3D building model used in this research was created in March 2013, and new buildings have since been constructed). This justifies the degraded solution at the upper turn (east-west) for the VO aided 3D building model. This issue can be solved by employing a recently created/updated 3DBM. Also, due to the scale inaccuracy obtained from the VO-only case, it can be seen that at the turns, the VO solution drifts (went over the reference at the lower turn) while the translation magnitude is smaller than the reference at the upper corner.

The cumulative distribution function (CDF) of the horizontal position error is depicted in Figure 11.

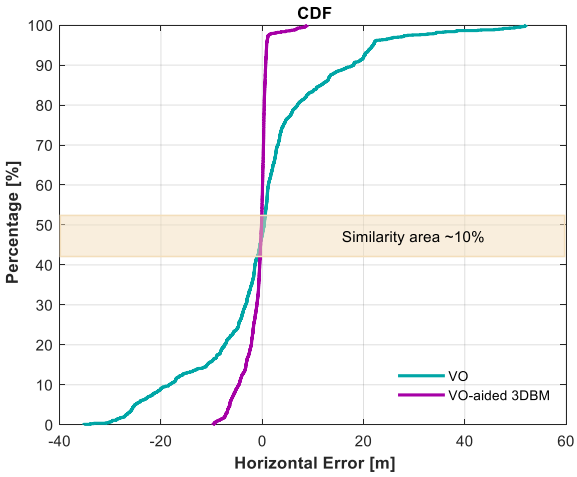


Figure 11: CDF of the horizontal errors

This figure shows that position error for the VO-only scenario is bounded between -30 m to 50 m while for the VO-aided 3DBM, the position error is bounded between -10 m and 15 m. This clearly suggests that the position error introduced by the scale drift degrades the VO position solutions compared to the VO aided by the 3DBM. Results are similar about 10% of the time, suggesting that the proposed method improved the final solution up to 90% of the time.

The assessment of the influence of the scale factor drift on other metrics such as the velocity are left for future work. It is also worth mentioning that the skyline-based positioning performs better in environments with tall buildings. It is thus important to consider a camera field-of-view (FOV) such as it is possible to capture as much as possible surrounding buildings.

5 CONCLUSIONS

This paper presents a method of resolving the scale factor drift observed in monocular visual odometry systems by using a 3D city model to obtain the slant distance, similar to the information provided by LiDAR data. The proposed method was assessed by using data collected in downtown Calgary and compared with monocular visual odometry alone. The results have shown that the proposed approach improved the final solution 90% of the time compared to the VO-only.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Mark Petovello for his suggestions during the early development of this work.

REFERENCES

Aqel, M. O., M. H. Marhaban, M. I. Saripan, and N. B. Ismail. 2017. "Estimation of image scale variations in monocular visual odometry systems." *IEEJ Transaction Electrical and Electronic Engineering* 12: 228-243. doi:10.1002/tee.22370.

Aumayer, B. M. 2016. *Ultra-tightly Coupled Vision/GNSS for Automotive Applications.* PhD Thesis, Department of Geomatics Engineering, University of Calgary. doi:dx.doi.org/10.5072/PRISM/28546.

Ben Afia, A., A. C. Escher, and C. Macabiau. 2015. "A Low-cost GNSS/IMU/Visual monoSLAM/WSS Integration Based on Federated Kalman Filtering for Navigation in Urban Environments." *ION GNSS+ 2015, Proceedings of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation.* Tampa, FL, USA. 618-628.

Choi, S., J. Park, and W. Yu. 2013. "Resolving Scale Ambiguity for Monocular Visual Odometry." *10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI).* Jeju, Korea. 604-608.

Chu, T., N. Guo, S. Backén, and D. Akos. 2012. "Monocular Camera/IMU/GNSS Integration for Ground Vehicle Navigation in Challenging GNSS Environments." *Sensors* 12: 3162-3185. doi:10.3390/s120303162.

Fei, Liu, Balazadegan Sarvrood Yashar, and Gao Yang. 2015. "Tightly Coupled Stereo Vision Aided Inertial

Navigation Using Continuously Tracked Features for Land Vehicles." *Proceedings of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2015).* Tampa, Florida. 2127 - 2133.

Fischer, M.A., and R.C Bolles. 1981. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography." *Communications of the ACM* (ACM) 24 (6): 381-395. doi:10.1145/358669.358692.

Gabriel, Nützi, Weiss Stephan, Scaramuzza Davide, and Siegwart Roland. 2011. "Fusion of IMU and Vision for Absolute Scale Estimation in Monocular SLAM." *Journal of Intelligent & Robotic Systems* (Springer Netherlands) 61 (1-4): 287–299. doi:https://doi.org/10.1007/s10846-010-9490-z.

Gakne, Paul Verlaine. 2018. "Improving the Accuracy of GNSS Receivers in Urban Canyons using an Upward-Facing Camera." PhD Thesis, Geomatics Engineering, University of Calgary. doi:dx.doi.org/10.5072/PRISM/32274.

Gakne, Paul Verlaine, and Kyle O'Keefe. 2017. "Monocular-based pose estimation using vanishing points for indoor image correction." *2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN).* Sapporo, Japan. doi:10.1109/IPIN.2017.8115954.

—. 2017. "Skyline-based Positioning in Urban Canyons Using a Narrow FOV Upward-Facing Camera." *Proceedings of the 30th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2017).* Portland, Oregon. 2574-2586.

Gakne, Paul Verlaine, and Kyle O'Keefe. 2018. "Tightly-Coupled GNSS/Vision Using a Sky-Pointing Camera for Vehicle Navigation in Urban Areas." *Sensors* 18 (4): 1244. doi:10.3390/s18041244.

Gallagher, A. C. 2005. "Using vanishing points to correct camera rotation in images." *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05).* Victoria, BC, Canada. 460-467. doi:10.1109/CRV.2005.84.

Gräter, J., T. Schwarze, and M. Lauer. 2015. "Robust scale estimation for monocular visual odometry using structure from motion and vanishing points." *2015 IEEE Intelligent Vehicles Symposium (IV).* Seoul. 475-480. doi:10.1109/IVS.2015.7225730.

Horn, R. A., and C. R. Johnson. 1990. "Norms for Vectors and Matrices Ch. 5 in Matrix Analysis." Cambridge, England: Cambridge University Press.

Ji, Zhang, and Singh Sanjiv. 2015. "Visual–Inertial Combined Odometry System for Aerial Vehicles." *Journal of Field Robotics* 32 (8): 1043-1055.

Jung, Sang-Hack, and Camillo J. Taylor. 2011. "Camera trajectory estimation using inertial sensor measurements and structure from motion results." *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.* Kauai, HI, USA. II-732-II-737. doi:10.1109/CVPR.2001.991037.

Kai, Wu, Di Kaichang, Sun Xun, Wan Wenhui, and Liu Zhaoqin. 2014. "Enhanced Monocular Visual Odometry Integrated with Laser Distance Meter for Astronaut Navigation." *Sensors* 14 (3): 4981–5003. doi:10.3390/s140304981.

Kitt, B., J. Rehder, A. Chambers, M. Schönbein, H. Lategahn, and S. Singh. 2011. "Monocular Visual Odometry using a Planar Road Model to Solve Scale Ambiguity." *Proceedings of the 5th European Conference on Mobile Robots (ECMR 2011).* Örebro, Sweden: Ed.: A. J. Lilienthal.

Klein, Georg, and David Murray. 2007. "Parallel Tracking and Mapping for Small AR Workspaces." Nara, Japan. Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07).

Li, Xiao Chen, Hu Weidong, Zhang Lefeng, Shi Zhiguang, and Maisi. 2018. "Integration of Low-Cost GNSS and Monocular Cameras for Simultaneous Localization and Mapping." *Sensors* 18 (7): 2193. doi:10.3390/s18072193.

Lim, J. H., K. H. Choi, J. Cho, and H. K. Lee. 2017. "Integration of GPS and monocular vision for land vehicle navigation in urban area." *International Journal of Automotive Technology* 18 (2): 345–356. doi:10.1007/s12239-017-0035-3 .

Lowe, D. G. 2004. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60 (2): 91-110. doi:10.1023/B:VISI.0000029664.99615.94.

Ma, Y., S Soatto, J. Kosecka, and S. S. Sastry. 2003. *An Invitation to 3-D Vision: From Images to Geometric Models.* SpringerVerlag.

Manolis, Lourakis, and Zabulis Xenophon. 2013. "Accurate Scale Factor Estimation in 3D Reconstruction." *In: Wilson R., Hancock E., Bors A., Smith W. (eds) Computer Analysis of Images and Patterns. CAIP 2013. Lecture Notes in Computer Science.* Springer, Berlin, Heidelberg. 498-506. doi:10.1007/978-3-642-40261-6_60.

Marouane, C., R. Gutschale, and C. Linnhoff-Popien. 2018. "Visual Odometry for Pedestrians Based on Orientation Attributes of SURF." *In: Bi Y., Kapoor S., Bhatia R. (eds) Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016. IntelliSys 2016. Lecture Notes in Networks and Systems.* Springer, Cham. 153-174. doi:10.1007/978-3-319-56991-8_13.

Mur-Artal, R., and J. D. Tardós. 2017. "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras." *in IEEE Transactions on Robotics* 33 (5): 1255-1262. doi:10.1109/TRO.2017.2705103.

Mur-Artal, R., J. M. M. Montiel, and J. D. Tardos. 2015. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System." *IEEE Transactions on Robotics* 31 (5): 1147–1163.

Petovello, M. G., and Z. He. 2015. "Assessment of Skyline Variability for Positioning in Urban Canyons." *Proceedings of the ION 2015 Pacific PNT Meeting.* Honolulu, Hawaii. 1059-1068.

Petovello, M., and Z He. 2016. "Skyline Positioning in Urban Areas Using a Low-cost Infrared Camera." *2016 European Navigation Conference (ENC).* 1-8.

Rublee, E., V. Rabaud, K. Konolige, and G. Bradski. 2011. "ORB: An Efficient Alternative to SIFT or SURF." *In Proceedings of the 2011 International Conference on Computer Vision (ICCV'11).* Barcelona, Spain. 2564–2571. doi:10.1109/ICCV.2011.6126544.

Soloviev, Andrey, and Donald Venable. 2010. "Integration of GPS and vision measurements for navigation in GPS challenged environments." *IEEE/ION Position, Location and Navigation Symposium.* Indian Wells, CA. 826-833. doi:10.1109/PLANS.2010.5507322.

Song, S., M. Chandraker, and C. C. Guest. 2016. "High Accuracy Monocular SFM and Scale Correction for Autonomous Driving." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (4): 730-743. doi:10.1109/TPAMI.2015.2469274.

Spaenlehauer, Ariane, Frémont Vincent, Sekercioglu Y. Ahmet, and Fantoni Isabelle. 2017. "A Loosely-Coupled Approach for Metric Scale Estimation in Monocular Vision-Inertial Systems." *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI).* 137-143.

Weiss, S., and R. Siegwart. 2011. "Real-time metric state estimation for modular vision-inertial systems." *2011 IEEE International Conference on Robotics and Automation.* Shanghai. 4531-4537. doi:10.1109/ICRA.2011.5979982.

Won, D.H., E. Lee, M. Heo, S.-W. Lee, J. Lee, J. Kim, S. Sung, and Y.J. Lee. 2014. "Selective Integration of GNSS, Vision Sensor, and INS Using Weighted DOP Under GNSS-Challenged Environments." *in IEEE Transactions on Instrumentation and Measurement* 63 (9): 2288-2298. doi:10.1109/TIM.2014.2304365.

Xiao-Shan, Gao, Hou Xiao-Rong, Tang Jianliang, and Cheng Hang-Fei. 2003. "Complete Solution Classification for the Perspective-Three-Point Problem." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (8): 930-943. doi:10.1109/TPAMI.2003.1217599.

Zhang, Ji, Sanjiv Singh, and George A. Kantor. 2012. "Robust Monocular Visual Odometry for a Ground Vehicle in Undulating Terrain." *The 8th International Conference on Field and Service Robots (FSR 2012).*

Zhang, Z., R. Zhao, E. Liu, K. Yan, and Y. Ma. 2018. "Scale Estimation and Correction of the Monocular Simultaneous Localization and Mapping (SLAM) Based on Fusion of 1D Laser Range Finder and Vision Data." *Sensors* 1948.