15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

A Study of Virtual Visual Servoing Sensitivity in the Context of Image/GIS Registration for Urban Environments

Hengyang Wei, Muriel Pressigout, Luce Morin
IETR - INSA Rennes
35700 Rennes Cedex 7, France
{hwei, mpressig, lmorin}@insa-rennes.fr

Myriam Servières, Guillaume Moreau
CRENAU-AAU / Ecole Centrale Nantes
44321 Nantes Cedex 3, France
{myriam.servieres, guillaume.moreau}@ec-nantes.fr

Abstract

This paper studies the sensitivity of pose estimation to the 2D measure noise when using virtual visual servoing. Attempting to apply virtual visual servoing to image/Geographic Information System (GIS) registration, the robustness to the noise in images is an important factor to the accuracy of estimation. To analyze the impact of different levels of noise, a series of image/GIS registration tests based on synthetic input image are studied. Also, RANSAC is introduced to improve the robustness of the method. We also compare some different strategies in choosing geometrical features and in the treatment of projection error vector in virtual visual servoing, providing a guide for parametrization.

1 Introduction

Image/GIS (Geographic Information System) registration in urban environments is an important step for outdoor augmented reality. It establishes the relationship between the 2D objects in the acquired image/video with the 3D models in the GIS. Registration is closely related to camera pose estimation: it comes to estimate the position and orientation of the camera in the GIS world frame.

Several methods have been proposed to solve the registration problem. Some methods use fiducial markers for outdoor tracking [6]. In marker-less circumstance, model-based algorithms such as [1] are often used in pose estimation. Other approaches construct the model of the scene at the same time as estimating the camera pose, based on SLAM [7][8] or structure from motion [9].

Virtual visual servoing [1][2] is a framework for real-time registration. Its advantage is that it can combine different geometrical features for tracking. Virtual visual servoing is quite accurate when the tracked features are well extracted from images. However, few works use virtual visual servoing in outdoor pose estimation [5] with large scale building models.

Indeed, buildings contours extraction is difficult because of the uncontrollability of urban environments. This may cause inaccuracy of the extracted contours, bringing noise and error to pose estimation. Thus the sensitivity of virtual visual servoing estimation to noise on observed image features is a key factor. Considering that the contours of buildings are often occluded by objects on the ground such as people, cars and trees, we focus on the buildings skyline as proposed by [4]. The contribution of this paper consists of assessing the impact of noise on virtual visual servoing pose estimation using skyline.

2 Image/GIS Registration

This section reminds the principle of image/GIS registration based on virtual visual servoing and the used features and parameterization.

2.1 Virtual visual servoing

In virtual visual servoing [1][2], any kind of geometrical feature $\mathbf{p}$ can be used as long as the corresponding interaction matrix is computed. The interaction matrix related to the projection in the image plane $\mathbf{p_m}$ is noted as $\mathbf{L_{p_m}}$ and defined by

$$\mathbf{L_{p_m}} = \frac{\partial \mathbf{p_m}}{\partial \mathbf{r}} \tag{1}$$

where $\mathbf{r}$ is the extrinsic parameters matrix of the camera.

In our experiments, the skyline is modeled as a series of segments, the straight lines holding these segments will be used as the input geometric features. Each 2D line from the skyline is matched with its 3D counterpart among the GIS building contours. These 2D and 3D lines are provided as input data to virtual visual servoing, together with an initial camera pose. The process then estimates the camera pose corresponding to the input image using virtual visual servoing.

As mentioned in [2], the different kinds of geometrical features correspond to different interaction matrices. For a straight line which is defined as

$$\begin{cases} A_1X + B_1Y + C_1Z + D_1 = 0 \\ A_2X + B_2Y + C_2Z + D_2 = 0 \end{cases} \tag{2}$$

where D_1 and D_2 are not both zero, and its projection in image plane defined as

$$x\cos\theta + y\sin\theta - \rho = 0 \tag{3}$$

the interaction matrix related to (θ, ρ) is defined by

$$\mathbf{L_{p_m}} = \left(\begin{array}{cccccc} \lambda_\theta \cos\theta & \lambda_\theta \sin\theta & -\lambda_\theta \rho & \rho\cos\theta & -\rho\sin\theta & -1 \\ \lambda_\rho \cos\theta & \lambda_\rho \sin\theta & -\lambda_\rho \rho & (1+\rho^2)\sin\theta & -(1+\rho^2)\cos\theta & 0 \end{array} \right) \tag{4}$$

where $\lambda_\theta = (A_i \sin\theta - B_i \cos\theta)/D_i$ and $\lambda_\rho = (A_i\rho\cos\theta + B_i\rho\sin\theta + C_i)/D_i$, with $i = 1$ or 2 for which $D_i \neq 0$.

2.2 Pixel parameterization

In order to evaluate projection error in pixels (and not in meters as in previous equations), the projection of the straight line of equation (2) in image plane should be expressed in pixel:

$$x_{px}\cos\theta_{px} + y_{px}\sin\theta_{px} - \rho_{px} = 0 \tag{5}$$

and the interaction matrix related to (θ_{px}, ρ_{px}) should then be calculated as

$$\mathbf{L}_{\mathbf{p}_{\mathbf{m}_{px}}} = \frac{\partial \mathbf{p}_{\mathbf{m}_{px}}}{\partial \mathbf{r}} = \frac{\partial \mathbf{p}_{\mathbf{m}_{px}}}{\partial \mathbf{p}_{\mathbf{m}}}\frac{\partial \mathbf{p}_{\mathbf{m}}}{\partial \mathbf{r}} = \mathbf{J}_{px}\mathbf{L}_{\mathbf{p}_{\mathbf{m}}} \tag{6}$$

where $\mathbf{J}_{px}$ is the Jacobian matrix written as

$$\mathbf{J}_{px} = \begin{pmatrix} A(\theta_{px},\theta) & 0 \\ A(\theta_{px},\theta)B(\theta_{px},\theta,\rho) & C(\theta_{px},\theta) \end{pmatrix} \tag{7}$$

with

$$A(\theta_{px},\theta) = \frac{\alpha_x}{\alpha_y}\left(\frac{\cos\theta_{px}}{\cos\theta}\right)^2 \tag{8}$$

$$\begin{aligned} B(\theta_{px},\theta,\rho) = \; & y_0\cos\theta_{px} - x_0\sin\theta_{px} \\ & + \alpha_y\rho\frac{\sin\theta}{\cos\theta_{px}} - \alpha_x\rho\frac{\sin\theta_{px}}{\cos\theta} \end{aligned} \tag{9}$$

$$C(\theta_{px},\theta) = \alpha_x\left(\frac{\cos\theta_{px}}{\cos\theta}\right) \tag{10}$$

where α_x, α_y and s are the camera intrinsic parameters, α_x and α_y are the scale factors in x and y directions, and s is the skew of the camera.

2.3 Normalization

Additionally, a normalization between the errors on ρ and on θ may be added to take into account their different order of magnitude. The projection error vector is normalized by dividing fixed coefficients, which are the coordinates by the maximum initial error on ρ and on θ respectively.

2.4 Robust estimation with RANSAC

In our experiments, we want to estimate the benefit of a robust estimator such as RANSAC [3] for reducing the impact of noise in the estimation of camera pose.

After the matching between straight lines forming the skyline in both model and image spaces, RANSAC is launched on the set $\mathtt{S}$ of matched features. For each trial, a minimum set $\mathbf{s}_i$ of three features will be randomly sampled from $\mathtt{S}$, then virtual visual servoing will be applied on $\mathbf{s}_i$ to estimate the camera pose, noted as the transform matrix ${}^c\mathbf{M}_o^*(\mathbf{s}_i)$, from the world coordinate system to the camera coordinate system.

For each estimation ${}^c\mathbf{M}_o^*(\mathbf{s}_i)$, other features in $\mathtt{S}$ are tested by calculating the distance between image skyline and projected segment:

$$d(\mathbf{p}_{\mathbf{m_d}}, \mathbf{p}_{\mathbf{m}}({}^c\mathbf{M}_o^*(\mathbf{s}_i))) < \epsilon_1 \tag{11}$$

Here the distance between segments is defined as

$$d(s_1, s_2) = \max(d(m_1, s_2), d(m_2, s_1)) \tag{12}$$

where s_1, s_2 represent two segments, and m_1, m_2 are the midpoints of the two segments respectively. The distance $d(m, s)$ is defined as the distance from point m to the straight line holding the segment s.

A feature which passes the test in equation (11) with the estimation ${}^c\mathbf{M}_o^*(\mathbf{s}_i)$ is noted as an inlier to this estimation. The estimation ${}^c\mathbf{M}_o^*(\mathbf{s}_i)$ showing the higher number of inliers will be chosen as the estimation of RANSAC.

3 Experimental Setup

In order to control the level of noise on the data, all tests use synthetic images of 800 × 800 pixels as inputs (Figure 1). The ground-truth image skyline is generated from the building models using the known ground-truth camera pose. Then, the skyline segments in 3D models are computed by back-projection from the ground-truth image skyline. The matching between 2D lines of image skyline and 3D lines of GIS model is also done by using the ground-truth camera pose, and it is thus guaranteed to be correct in these experiments. The input initial pose is obtained by adding uniform random shift to the ground-truth camera pose, with an amplitude of 5 degrees and 5 meters respectively for rotation and translation, which is similar to the error provided by real sensors in an urban environment[10][11].

Two types of tests will be launched. In the first test, no noise is added on image skyline, and we aim to study the accuracy of pose estimation for different parameterization choices. In the second test, the parameterization is fixed, and noise is added to 2D image segments end points for studying the sensitivity of virtual visual servoing to noise in image measurements.

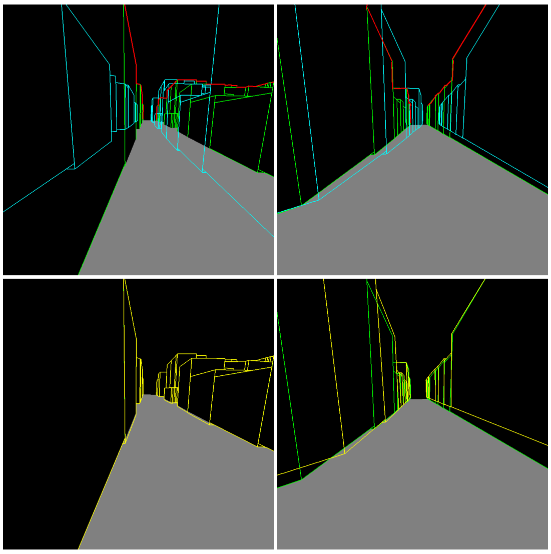

Figure 1. Example of images of two tests. Top two images are input synthetic images (green) with projected building contours by initial pose (light blue) and generated image skyline (red), bottom two images display the registration results (green for ground-truth and yellow for estimation) corresponding to top images respectively. In the registration image of the first case (bottom left image), the two projected building contours are superimposed since the estimation fits the ground-truth.

3.1 Impact of parameterization on pose estimation

In this first part of tests, the ground-truth image skyline and an approximate initial pose is provided as described before. Virtual visual servoing is applied on the matched features and initial pose to refine the camera pose. Several parameterization (either (ρ, θ) in (3) or (ρ_{px}, θ_{px}) in (5)) and normalization settings are compared:

(1) Feature on (ρ, θ), no normalization.

(2) Feature on (ρ_{px}, θ_{px}), no normalization.

(3) Feature on (ρ, θ), with normalization.

(4) Feature on (ρ_{px}, θ_{px}), with normalization.

There are 169 tests with different synthetic views and/or different initial camera poses for each group.

3.2 Impact of image noise

In this experiment, uniform noise is added on the extreme points of the segments of the ground-truth image skyline. For a segment in image plane $\overline{p_1(x_1, y_1)p_2(x_2, y_2)}$, a noise $e_{ij, i\in\{1,2\}, j\in\{x,y\}}$ is added forming new segment $\overline{p'_1(x_1 + e_{1x}, y_1 + e_{1y})p'_2(x_2 + e_{2x}, y_2 + e_{2y})}$, with e_{ij} a uniform random noise in $[-e_{\max}, e_{\max}]$.

When using RANSAC robust estimation, the threshold of maximum distance ϵ_1 between segments is 2 pixels, minimum number of inliers for a candidate is $N_{inlier} = 3$, and maximum iterations is 50.

Three levels of noise (0.5, 1 and 2 pixel(s)) have been applied, with 16900 tests launched for each level, giving out the errors of estimation with and without RANSAC.

4 Results

The configurations of the experiment to study the parametrization and the normalization are presented above. In this section, the results of the experiment are presented for each study.

4.1 Impact of parameterization on pose estimation

The results on the four different parameterization and normalization settings are shown in tables 1 and 2.

The results show that the parameterization on meter performs better than the one on pixel. For the test groups without normalization, the performance of straight line feature in pixel is much worse than in meter, both on the average estimation error on translation or on rotation, and on the standard deviation of error which describes the stability of the algorithm. This is because of the unbalanced magnitude of error on ρ_{px} and on θ_{px}. When we choose the parameterization on (ρ, θ), the errors on ρ and on θ are of the same order of magnitude. However, if we use the parameterization on (ρ_{px}, θ_{px}), the error on ρ_{px} may be of several tens of pixels while the error on θ_{px} is rather tenth of radians, which may cause the unbalance of influence of ρ_{px} and θ_{px} in iterations of virtual visual servoing.

Notice that the difference between the results of parameterization on (ρ, θ) and the one on (ρ_{px}, θ_{px}) indicates that the different parameterizations are not just a change of the unit as it may look like. In fact, the change of parameterization leads to a different interaction matrix, which means a different geometrical model of feature. As what we see in this experiment, the parameterization on (ρ_{px}, θ_{px}) causes the unbalanced magnitude of error on different coordinates. This may lead to the false termination of virtual visual servoing iteration, which brings a larger estimation error.

This also explains why the normalization has an obvious effect on the results of the test groups of geometrical feature in pixel. Comparing the result of test groups (3) and (4), which are the tests with normalization by dividing the maximum initial error in projection error vector of each component, though the group (4) (i.e. geometrical feature in pixel with normalization) is still worse than the group (3), the difference between the group of meter and the group of pixel is much smaller than the difference without normalization. For the group of 2D lines in pixel, the proper normalization can greatly improve the performance of estimation, since it reduces the difference between the magnitude of error on ρ_{px} and on θ_{px}. On the other hand, the result of normalized group in meter (group (3)) is on the same level as the no-normalization group. This indicates that the usual parametrization in meter with no normalization used in many works has an implicit normalization on ρ and on θ, so it is well adapted when using the straight line as geometrical feature.

Table 1. Estimation errors on translation with ground-truth image skyline: Mean error, standard deviation, and maximum error

	Error on translation (cm)		
Group	mean	std	max
(1)	0.24	0.39	2.16
(2)	25.0	106	617
(3)	0.24	0.42	2.65
(4)	1.7	21.2	276

Table 2. Estimation errors on rotation with ground-truth image skyline: Mean error, standard deviation, and maximum error

	Error on rotation (degree)		
Group	mean	std	max
(1)	0.0017	0.0015	0.0084
(2)	0.62	4.14	41.8
(3)	0.0017	0.0014	0.0082
(4)	0.0054	0.064	0.84

4.2 Impact of image noise

In this robustness tests, the aim is to study the impact of different levels of noise on the accuracy of estimation, and to compare the result with or without the robust estimation. We have chosen the usual 2D line in meter without normalization as the geometrical feature in virtual visual servoing.

Fig. 2 and fig. 3 show the statistical result of the tests. The result of the virtual visual servoing estimation is the left part for both two figures. In the figures, each box plot represents the results of tests with a given noise level. For each box plot, the bottom and top of the box are the first and third quartiles, and the red

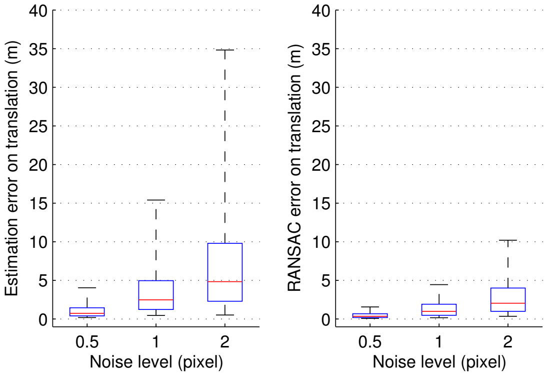

Figure 2. Estimation errors on translation with different levels of noise, without RANSAC (left) and with RANSAC(right)

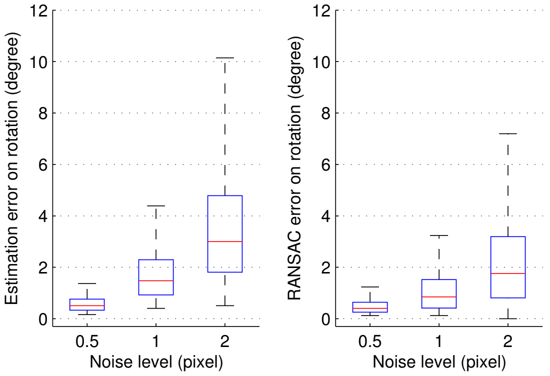

Figure 3. Estimation errors on rotation with different levels of noise, without RANSAC (left) and with RANSAC(right)

band inside the box is the median. The lower whisker is the 5% of the data, while the top whisker is 95%. The outliers are not displayed in the figure.

As shown in the figure, the result is strongly influenced by the noise on image skyline. From the level of noise at 1 pixel, the error of estimation is significantly increased. The impact of a 2 pixel noise is very significant, as the translation error and the rotation error is larger than 1 meter or 1 degree respectively for most tests.

This shows that image noise has a strong impact on estimated pose and that even a 1 pixel error on segment extreme points cannot be neglected.

4.3 Robust estimation

The right parts of fig. 2 and of fig. 3 show the results of RANSAC estimation. Compared to the estimation without RANSAC, we can see that the estimation accuracy is largely improved. The error of RANSAC estimation on noise level at 1 pixel is rather acceptable, and the result on noise level at 2 pixels has also a large improvement.

The result also shows that RANSAC estimation has a significant impact on reducing final error on both translation and rotation, and the error reduction is more notable on translation.

This test shows that even without matching errors, using a robust estimation as RANSAC has a significant benefit in the presence of noise on image features.

5 Conclusion

In this paper, we have compared several different parameterizations in applying virtual visual servoing to the estimation of camera pose in urban environments, and we have brought out the study of sensitivity to noise on 2D measures of virtual visual servoing. The comparison of parameterizations shows the importance of balance in different components of projection error vector. This explains why the usual meter/radians parameterization for polar representation of 2D lines is well adapted. The study of sensitivity towards noise of virtual visual servoing explains the difficulty of applying the virtual visual servoing in outdoor use case such as pose estimation in urban environments. However, robust estimation method, RANSAC for instance, can improve the robustness of virtual visual servoing against the noise on 2D measures. For future study, false detection of line segments in image and false matching between image skyline and model will be added for a better simulation of reality.

References

[1] A. I. Comport, E. Marchand, M. Pressigout, and F. Chaumette, "Real-time markerless tracking for augmented reality: the virtual visual servoing framework," *IEEE TVCG 06*, vol.12, no.4, pp.615–628, 2006.

[2] E. Marchand and F. Chaumette, "Virtual visual servoing: a framework for real-time augmented reality," *Computer Graphics Forum*, vol.21, pp.289–297, 2002.

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

[4] S. Zhu, L. Morin, M. Pressigout, G. Moreau, and M. Servières, "Video/GIS registration system based on skyline matching method," *ICIP 13*, pp.3632–3636.

[5] T. Colleu, G. Sourimant, and L. Morin, "Automatic initialization for the registration of GIS and video data," *3DTV Conference*, pp.49–52, 2008.

[6] U. Neumann, S. You, Y. Cho, J. Lee, and J. Park, "Augmented reality tracking in natural environments," *Int'l Symposium on Mixed Realities*, vol.24, 1999.

[7] A. J. Davison, "Real-time simultaneous localisation and mapping with a single camera," *ICCV 03*, pp.1403–1410, 2003.

[8] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit, "Instant outdoor localization and SLAM initialization from 2.5 D maps," *IEEE TVCG 15*, vol.21, no.11, pp.1309–1318, 2015.

[9] E. Royer, M. Lhuillier, M. Dhome, and T. Chateau, "Localization in urban environments: monocular vision compared to a differential GPS sensor," *CVPR 05*, vol.2, pp.114–121, 2005.

[10] M. G. Wing, A. Eklund, and L. D. Kellogg, "Consumer-grade global positioning system (GPS) accuracy and reliability," *Journal of forestry*, vol.103, no.4, pp.169–173, 2005.

[11] T. Ozyagcilar, "Accuracy of Angle Estimation in eCompass and 3D Pointer Applications," *Freescale Semiconductor Application Note*, 2015.