Visual Gyroscope: Combination of Deep Learning Features and Direct Alignment for Panoramic Stabilization

Bruno Berenguel-Baeta, Antoine André, Guillaume Caron, Jesus Bermudez-Cameo, Josechu Guerrero

Visual Gyroscope: Combination of Deep Learning Features and Direct Alignment for Panoramic Stabilization

Bruno Berenguel-Baeta[1], Antoine N.Andre[2], Guillaume Caron[23], Jesus Bermudez-Cameo[1], Jose J.Guerrero[1]
[1] I3A, Universidad de Zaragoza, Spain
[2] CNRS-AIST JRL (Joint Robotics Laboratory), IRL, Japan
[3] Universite de Picardie Jules Verne, MIS laboratory, France

Abstract

In this article we present a visual gyroscope based on equirectangular panoramas. We propose a new pipeline where we take advantage of combining three different methods to obtain a robust and accurate estimation of the attitude of the camera. We quantitatively and qualitatively validate our method on two image sequences taken with a 360° *dual-fisheye camera mounted on different aerial vehicles.*

1. Introduction

Autonomous navigation of Unmanned Aerial Vehicles (UAV) require the knowledge of pose and orientation. Among the different sensors we can use, cameras provide information of the surroundings of the UAV which can be used for several applications, such as object detection [8], while being a lightweight sensor. One of these applications is the use of the camera as a Visual Gyroscope (VG), estimating the 3D orientation of the camera for a given image with respect to a reference.

VG algorithms consider two kinds of information, direct or indirect. Indirect VG (IVG) leverage image features, either handcrafted, as patches around image points [6], or learned, *e.g.* estimating first the optical flow and then the 3D orientation of the camera [9]. Instead, Direct VG (DVG) consider pixel brightness of the whole image as input of a 3D rotation optimization method [1]. Usually, this pixel information is transformed into different domains before estimating the orientation, such as spherical Fourier transform [10] or Mixture of photometric potentials [3].

Previous methods present high accuracy on 3D orientation estimation on narrow domains [1], or the possibility of a wide estimation domain but with lower accuracy [10]. In this paper, we propose a hybrid pipeline combining three different methods to obtain a robust VG within a very large estimation domain and with high orientation accuracy. We leverage the strong points of each method, overcoming the weak points and limitations of each of them. In the next section we present the different methods that form our pipeline while in Sec. 3 we evaluate and compare our pipeline against state-of-the-art methods on two 360 equirectangular panoramas data-sets taken from UAV.

Work supported by PID2021-125209OB-I00 and TED2021-129410B-I00 (MCIN/AEI/10.13039/501100011033 and FEDER/UE and NextGenerationEU/PRTR)

2. Visual Gyroscope

Our proposed pipeline[1] is composed by three different methods that act sequentially to obtain the Roll-Pitch-Yaw angles of the camera in outdoor environments (see Fig. 1). The first block is a IVG composed by a convolutional neural network called HoLiNet (**Ho**rizon **Li**ne **Net**work) where we obtain a first approximation of Roll-Pitch angles. The second block is a DVG algorithm that uses MPP (**M**ixture of **P**hotometric **P**otentials) [3] to retrieve the Yaw angle taking as initial solution the orientation from HoLiNet. The third and final block is second DVG algorithm defined as PVG (**P**hotometric **V**isual **G**yroscope) where we refine the previous solutions.

2.1. HoLiNet

Our proposed HoLiNet follows an encoder-decoder structure taking as input information an equirectangular panorama in any orientation and providing two heat-maps, one for the horizon line and another for the vertical vanishing points (see Fig. 1).

For the encoder and feature extractor we use the convolutional part of ResNet-50 [7]. We use the available pretrained weights, leveraging the performance of this network for feature extraction. In the decoder we propose a set of convolutional layers followed by up-scaling layers and skip connections from the encoder part. The output of the decoder is a two-channel heat-map with the horizon line and

[1] Code is available in https://github.com/Sbrunoberenguel/HoLiNet

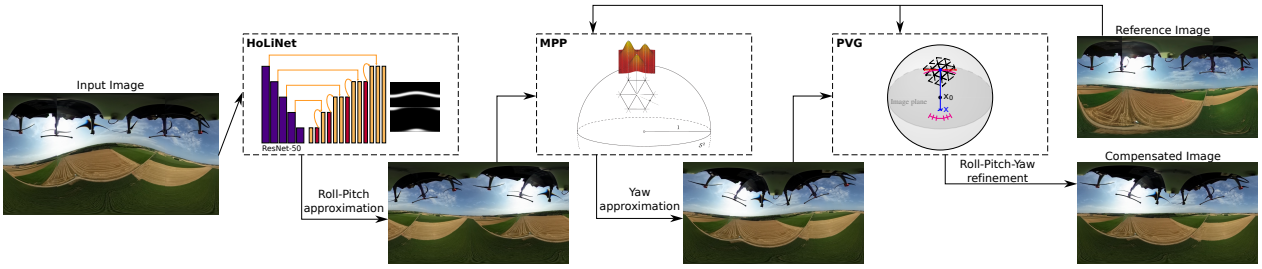


Figure 1: Overview of our proposed pipeline. The input equirectangular panorama first goes through HoLiNet, obtaining a first approximation of Roll-Pitch angles with respect to the horizontal plane. The output is feed to the MPP, obtaining an approximation of the Yaw angle with respect to a reference image. Finally, the PVG refines the three angles such that the rotation compensated image is closer to the reference one.

vanishing points prediction. To adapt the neural network to the distortion, we use in all the convolutional layers convolutions adapted to the equirectangular projection. These convolutions, EquiConvs, were first presented in [4] and, in this work, are re-implemented pre-computing the kernel's distortion for a faster inference time (i.e. the original implementation runs at 0.2fps while ours can run at 10fps).

The network is trained in a collection of gravity-oriented equirectangular panoramas from indoor and outdoor environments [5]. We apply rotations to these panoramas and generate the ground truth labelling according to these rotations. We also perform several data augmentations such as: horizontal flip, color permutation, color jitter and random cropping. At training, we obtain intermediate representations at different resolutions from the decoder of the network. These intermediate representations allow to evaluate the network on early stages, making the intermediate features of the network aim for the final task, improving the overall performance of the network.

From the output information of the network, we compute the most probable horizon plane, obtaining a first approximation of the camera's Roll-Pitch angles. To compute the horizon plane, we use the information of both outputs of the network. First, we use the vanishing point heat-map to make a rough estimation of the vertical direction. We project the vanishing points heat-map in the unit sphere and compute the average vertical direction. This is a rough estimation since the output is really sparse and can be noisy. Then, we project the horizon line heat-map into the unit sphere. Using a random sample consensus algorithm, we compute the 3D plane that better fits the horizon line heat-map in the unit sphere. Assuming that both network's outputs are coherent with each other, the first estimation of the vertical direction and the normal vector of the horizon plane should be close. So, to maintain this coherence, in the iterative process we discard any possible solution which differs from the first estimation more than a threshold angle (i.e. we assume a threshold of $\pm 30^\circ$).

Once obtained the plane that better fits the output of the network, we can compute the Roll and Pitch angles of the camera (Yaw is the rotation around the plane's normal vector, which cannot be computed with this method). The main advantage of this method is that we can obtain 2 angles from a 360° domain. The main limitation is the angle precision, which is limited by the uncertainty of the output of the network, and the impossibility to obtain the Yaw angle with the proposed representation.

2.2. MPP-based visual gyroscope

The MPP is a full spherical image representation of image brightness as a mixture of Gaussians. The MPP is structured as an icosahedron subdivided $n \in \mathbb{N}$ times, of which each vertex is the center of a Gaussian function, weighted by the normalized intensity of the pixel where it projects in the captured image. Thus, there are as many Gaussians as vertices of the sphere but all Gaussians of the MPP share the same variance $\lambda_g \in \{\mathbb{R} > 0\}$ that acts as a smoothing parameter. In the seminal work introducing the MPP-based visual gyroscope [3], captured dual-fisheye images are first transformed as MPPs: G^* for a reference image, G for a request image. Thus, λ_g acts as a smoothing parameter of the cost function, that is the sum of squared differences between G and G^*, to be minimized by optimizing the three angles of the 3D rotation that aligns the best both MPPs with a Newton-like algorithm.

In this work, the differences w.r.t. [3] are motivated by the HoLiNet-MPP-PVG pipeline we propose (Fig. 1):

- *input images are equirectangular* instead of dual-fisheye for both the sake of uniformity with respect to HoLiNet and more generality since the equirectangular format is much more common than dual-fisheye.
- *the single Yaw angle is estimated* since not only HoLiNet already outputs estimates of Roll and Pitch angles that can serve as good initial guesses for PVG (see Sec. 2.3) but it reduces a little the computation load.

The main advantage of the MPP-based visual gyroscope is its very large convergence domain, observed as being

both globally convergent indoor for a single pure rotation angle and almost as large for 3D rotations, even in the presence of translations, in [3]. However, its main drawback is the required trade-off between processing time and estimation accuracy: to reach a processing rate of 3 images per second, the icosahedron maximum subdivision level is $n = 3$ and the average estimation error was reported about 3 degrees with $\lambda_g = 0.325$. Thus, PVG is required as a final step to improve the accuracy of estimations.

2.3. Photometric Visual Gyroscope (PVG)

To improve both the processing time and the accuracy with respect to the MPP-based visual gyroscope, the PVG [1] uses a spherical image representation directly from image brightness without a transformation to a MPP. The PVG is built on the same principle as the MPP-based visual gyroscope and also relies on the intensities of the image mapped to a subdivided icosahedron. However, instead of computing a MPP, the direct pixel brightness are assigned to each vertices of the icosahedron, allowing a much faster computation of the Jacobians in the Newton-like optimization of the orientations. When mapping full resolution images (typically 1280×720) to the subdivided icosahedron, a higher precision in the orientation estimation can be obtained. The spherical gradient is computed in the image plane (as highlighted in the PVG thumbnail of Fig. 1) and not with neighbors vertices as in the MPP-based gyroscope.

The main difference of this work with respect to the source work [1] is that *equirectangular input images* are used, instead of dual-fisheye images, allowing a consistency of input images all along the orientation estimation pipeline.

PVG has shown estimation errors below $0.1°$ in the single axis rotation case, both pure and with translations [1]. To reach such a precision, the icosahedron is subdivided $n = 5$ times and the captured image at full resolution is used for a computation time $2\%$the one of the MPP-based gyroscope. But inversely to the MPP-based gyroscope, the low processing load and the high accuracy come at the price of a narrow convergence domain that was observed at $\pm 12.5°$ for the single rotation angle case. Hence, PVG needs good initial guesses that HoLiNet (Roll, Pitch) and the MPP-based visual gyroscope (Yaw) bring.

3. Experiments

3.1. Experimental setup

We evaluate our pipeline with two datasets. The first is the Sequence 8 of PanoraMIS [2] that was taken with a Ricoh Theta S mounted on a Parrot Disco fixed-wing drone. Since this sequence is captured as raw dual-fisheye, we first transform the images to the equirectangular format using the software shared with the dataset[2]. With this sequence of 12000+ images, we make a qualitative evaluation by compensating the estimated rotations with respect to image 8890 in order to compare the success rate with [3].

To allow the quantitative evaluation, we introduce a new dataset of 2700+ equirectangular images, SVMIS+[3]. SVMIS+ is acquired with a Ricoh Theta S too but embedded on a Matrice 600 Pro hexarotor manually flown in a countryside area during $4'36"$. The use of the hexarotor has two interests over the Disco drone. First, it allows to embed an extra computer connected to the onboard electronics in order to save the orientation information of the drone's Inertial Measurement Unit synchronously with the image capture. Second, the different type of motion of a hexarotor compared to a fixed-wing drone provides an additional variety in the evaluation, together with different meteorological conditions (sunny versus cloudy). With this dataset, in addition to the quantitative evaluation, we also perform a qualitative evaluation by compensating the estimated rotations of all the images with respect to a selected frame. Frame 1074 is selected such that the image is roughly at the center of the flown area, with attitude angles near zero (horizon line near straight and horizontal in the image).

3.2. Results

With PanoraMIS' Sequence 8, we evaluate qualitatively the success rate of the computed orientations by studying the misalignment of stabilized images with respect to the reference image. Over the 12000+ images, 564 were misaligned. This represents a success rate of 95.3%, thus outperforming previously shown results of 88% [3]. The stabilization results with the whole first dataset are shared as a video[4] highlighting each intermediate result of our pipeline for orientation estimation.

With the synchronized and reliable ground truth provided with SVMIS+, a quantitative evaluation of the pipeline is put in practice. HoLiNet is only evaluated over 2 angles. To avoid ambiguities in the evaluation of the two angles, we compute the predicted normal vector, n, of the horizon plane and compare with the ground truth measure, $\hat{n}$, computing the angle between both vectors as: $\arccos(n \cdot \hat{n})$. The quantitative results are shown in Fig. 2, obtaining a mean error of $3.04°$.

The evaluation for the 3 angles is made computing the angle difference between the angle-axis rotation of ground truth with respect to the reference image, R_{gt}, and our estimation, R_{pred}, as: $\arccos\left((trace(R_{gt} \cdot {R_{pred}}^T) - 1)/2\right)$. The results are presented in Fig. 3, obtaining a mean error of $17.6°$ for HoLiNet + MPP, $16.7°$ for the whole pipeline over the whole sequence, and $4.6°$ for 900 images (33% of SVMIS+) around the reference image (Fig.3b), corresponding to a difference of altitude with respect to the reference

[2] http://github.com/PerceptionRobotique/dualfisheye2equi

[3] http://mis.u-picardie.fr/~g-caron/pub/data/SVMISplus_er_pano.zip

[4] http://mis.u-picardie.fr/~g-caron/videos/2023OmniCV_svmis.mp4

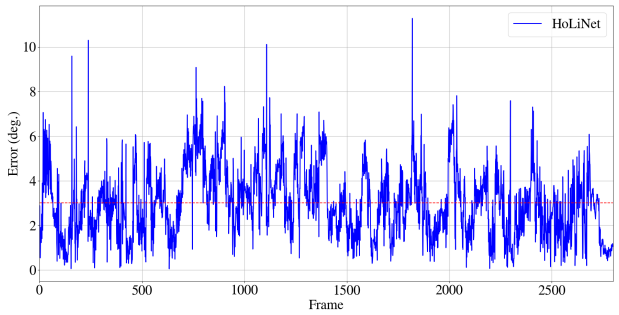
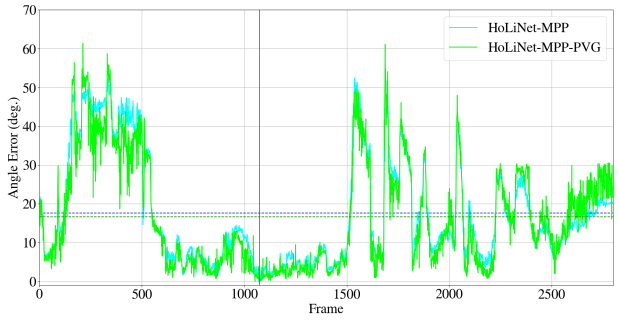
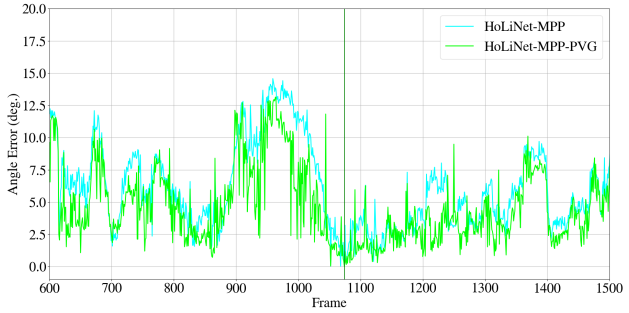


Figure 2: Quantitative results of HoLiNet for 2angles. Dashed lines show mean error.



(a)



(b)

Figure 3: a) Quantitative results of HoLiNet+MPP and HoLiNet+MPP+PVG in the SVMIS+ dataset. b) (Zoom) Quantitative results of MPP and PVG around the reference image. Vertical green line defines the reference image and horizontal dashed lines show mean error of each method.

image lower than 10 m. We also present a video[5] of qualitative results with SVMIS+.

4. Conclusions

We have presented a novel pipeline that leverages the strong points of three methods, HoLiNet using neural networks and two using direct alignment, for panoramic image rotation compensation in outdoor environments. The quantitative results on two datasets of airborne panoramic images show HoLiNet is robust to any translation and improves the visual alignment success rate. But since the direct alignment methods following HoLiNet in the pipeline assume pure rotation between the reference and current images, the presence of a large translation decreases the rotation estimation accuracy, as shown with the newly introduced public dataset SVMIS+. However, the rotation estimation appears accurate within a reasonable translation to the reference image that further study with SVMIS+ will allow to quantify during future works.

[5] http://mis.u-picardie.fr/~g-caron/videos/2023OmniCV_svmis_p.mp4

References

[1] Antoine N André and Guillaume Caron. Photometric visual gyroscope for full-view spherical camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5232–5235, 2022.

[2] H. Benseddik, F. Morbidi, and G. Caron. Panoramis: An ultra-wide field of view image dataset for vision-based robot-motion estimation. *The International Journal of Robotics Research*, 39(9):1037–1051, 2020.

[3] Guillaume Caron and Fabio Morbidi. Spherical visual gyroscope for autonomous robots using the mixture of photometric potentials. In *International Conference on Robotics and Automation*, pages 820–827. IEEE, 2018.

[4] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cedric Demonceaux, Javier Civera, and Jose J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 5(2):1255–1262, 2020.

[5] Eduardo Fraguas Bordonaba and Jesus Bermúdez Cameo. Estimación de línea del horizonte en imágenes de 360 grados con aprendizaje automático profundo. *Master thesis*, 2020.

[6] Hicham Hadj-Abdelkader, Omar Tahri, and Houssem-Eddine Benseddik. Closed form solution for rotation estimation using photometric spherical moments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 627–634. IEEE, 2018.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A review of yolo algorithm developments. *Procedia Computer Science*, 199:1066–1073, 2022.

[9] Dabae Kim, Sarthak Pathak, Alessandro Moro, Atsushi Yamashita, and Hajime Asama. Self-supervised optical flow derotation network for rotation estimation of a spherical camera. *Advanced Robotics*, 35(2):118–128, 2021.

[10] Ameesh Makadia and Kostas Daniilidis. Rotation recovery from spherical images without correspondences. *IEEE transactions on pattern analysis and machine intelligence*, 28(7):1170–1175, 2006.