Master internship proposal – 2025

Unsupervised language-aided landmark discovery and matching for visual localization in complex environments

Contact: Dr. Vincent Gaudillière (vincent.gaudilliere@loria.fr)

1 General information

Position: M2 internship

Duration: 5 to 6 months, starting in February 2025 (flexible)

Location: Loria, Nancy, France

Affiliation: TANGRAM team (Inria-Loria)

Supervisors: Vincent Gaudillière, Marie-Odile Berger and Gilles Simon

2 Context, description and objectives

This internship will deal with the problem of visual localization, which involves determining a camera's viewpoint by automatically matching features in an image with elements from a known 3D model of the environment. These features are referred to as landmarks.

Object-based localization [1, 7, 3] uses "high-level" landmarks, such as objects (*e.g.*, chairs, tables, cupboards), as opposed to the more commonly used "low-level" keypoints (*e.g.*, SIFT [2], ORB [5]). This approach offers the advantage of relying on fewer, more discriminative landmarks but is currently limited to environments that are rich in common objects, often artificially created for research purposes. Moreover, creating the 3D model requires manually matching objects detected across multiple images, a process that can be time-consuming and tedious.

For this internship, we will focus on complex industrial environments (*e.g.*, factories, power plants, ships) where the concept of an object is not always clearly defined. The goal is to automatically identify high-level landmarks in each image and ensure automatic matching of the detected landmarks across different images. To achieve this, we will employ "unsupervised" methods, which do not require environment-specific training, and explore the role of language in describing objects.

The first part of the internship will involve a literature review on unsupervised object localization in images [6] and the use of vision-language models (e.g., CLIP [4]). The second part will involve applying some of these methods to images of industrial environments and analyzing their results in terms of relevancy and repeatability. The final part will focus on proposing methods for automatically matching detected landmarks across different images.

3 Candidate profile

- The candidate is pursuing his/her last year of Master's or engineering's degree in Computer Vision, Electrical Engineering, Computer Science, Applied Mathematics or a related field.
- A strong background in image processing or/and in computer vision is required.
- A strong level of Python programming is required.
- An interest in deep learning frameworks (Pytorch) is also required.
- Commitment, team working and a critical mind.
- Good oral and written communication skills in English.

4 How to apply

Interested candidates are encouraged to send their applications (detailed CV, transcripts and a brief motivation letter) as soon as possible to the following address: vincent.gaudilliere@loria.fr. Applications will be processed upon reception.

References

[1] V. Gaudillière, G. Simon, and M.-O. Berger, "Perspective-2-Ellipsoid: Bridging the Gap Between Object Detections and 6-DoF Camera Pose," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5189–5196, Oct. 2020, conference Name: IEEE Robotics and Automation Letters. [Online]. Available: https://ieeexplore.ieee.org/document/9127873

[2] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[3] S. Matsuzaki, T. Sugino, K. Tanaka, Z. Sha, S. Nakaoka, S. Yoshizawa, and K. Shintani, "CLIP-Loc: Multi-modal Landmark Association for Global Localization in Object-based Maps," in *2024 IEEE International Conference*

on Robotics and Automation (ICRA), May 2024, pp. 13 673–13 679. [Online]. Available: https://ieeexplore.ieee.org/document/10611393

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[5] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 2564–2571, iSSN: 2380-7504. [Online]. Available: https://ieeexplore.ieee.org/document/6126544

[6] O. Siméoni, Zablocki, S. Gidaris, G. Puy, and P. Pérez, "Unsupervised Object Localization in the Era of Self-Supervised ViTs: A Survey," *International Journal of Computer Vision*, Aug. 2024. [Online]. Available: https://doi.org/10.1007/s11263-024-02167-8

[7] M. Zins, G. Simon, and M.-O. Berger, "OA-SLAM: Leveraging Objects for Camera Relocalization in Visual SLAM," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2022, pp. 720–728, iSSN: 1554-7868. [Online]. Available: https://ieeexplore.ieee.org/document/9995573