PhD Research Topic

- Title of the proposed topic: Hybrid AI for sensor-based robot control
- Research axis of the 3IA: Axe 4 - AI for Smart and Secure Territories
- Supervisor : Ezio MALIS
- Research group: ACENTAURI project-team, Inria Center at Universit´e Cˆote d'Azur

Context

Hybrid AI combining data-driven and rule-driven approaches is a promising research direction that is investigated both by international research groups, academic , Stanford, Berkeley, ...) and industrial (Google, Facebook, ...), and French research groups in AI for mobile robotics. In robotics, most of the research fields (design, perception, decision, motion planning, control...) have been investigated through an hybrid approach mainly rule based where data driven algorithms are used to identify and adapt on-line parameters of dynamic systems (included in the dynamic environment). As intelligent and autonomous systems are the main concern of the ACENTAURI team a novel point of view must be taken. Our strategy for addressing our scientific objectives is to build a bridge between rule-driven and data-driven approaches to artificial intelligence with the ambition to feed the models with knowledge provided by data-driven approaches and, conversely, to constrain data-driven approaches with highly accurate prior knowledge coming from the robot task. The first problem is hard because data-driven approaches are able to capture knowledge that is not in the model, therefore we need to be able to interpret the results (e.g. conceiving explainable ANNs) in order to build physically meaningful models (probably increasing their complexity). The second problem is hard because injecting prior rule-driven knowledge into data-driven approaches implies the design and the development of new architectures.

PhD subject

Within this context, this PhD subject will focus on sensor-based (Lidar and Stereo Vision) control. In order to achieve autonomous navigation the robot must not only be able to localize itself but also be able to avoid all possible obstacles in the environment. The multi-modal sensor data may be processed by deep learning models that make up the perception module to produce the detection of navigable space. The major challenge in building the perception module is to ensure that its deep learning models function properly in every condition. While deep learning models perform well on data resembling their training sets, they are prone to errors when used in different scenarios [7].

Regardless of data variations, the underlying geometric characteristics and physical laws of our world remain constant. These unchanging laws are well-described by expert models (e.g., Newton's laws of motion, Euclidean and Projective Geometry, ...). The objective of this PhD is to study of Hybrid IA models based on the fusion of expert knowledge (rule-based approaches) with machine learning models (rule-based approaches). Such an hybrid approach should combine the adaptability of machine learning with the reliability of established geometric and physical principles.

The expert knowledge we would like to integrate into the hybrid model is the equivariance of 3D world measurements with respect to the viewpoint changes. This concept states that changes in viewpoint result in predictable transformations of the measurement of the scene. For instance, when a camera undergoes a rotation, the content in its images rotates correspondingly in a globally predictive manner. By incorporating this expert knowledge, deep learning models can disentangle geometric factors from appearance factors (such as colors and textures) that vary significantly across environments. This capability of equivariant models makes them more data efficient and generalize better to unseen data [12], resulting in more robust performance compared to their non-equivariant counterparts. Mathematically, a change in viewpoint is formalized as a rigid body transformation, which belongs to the

Special Euclidean group SE(3). Therefore, **the objective** of this PhD is to study **SE(3)-equivariant networks** for robust sensor-based control of autonomous vehicles.

As Convolutional Neural Networks (CNN) are translationally equivariant thanks to weight sharing among different spatial positions of the input, we seek an extension of CNN that is rotationally equivariant to achieve SE(3) equivariance. Theoretical developments in deep learning have shown that Group equivariant CNN (G-CNN) is the only candidate [4]. G-CNN extends the conventional CNN by additionally sharing weights among different orientations of the input. G-CNN creates multiple rotated versions of each filter, corresponding to the rotations in group $\mathcal{G}$. During convolution, these rotated filters are applied to the input signal, generating an ordered set of feature maps. The order of these feature maps reflects the order of rotations in $\mathcal{G}$.

The principle of G-CNN is illustrated in Fig.1. The original input is the faded green lizard. The original filter is small (circular) disc shown in the top row of the second column from the left of Fig.1. We consider the group $\mathcal{G}$ containing four counter-clockwise rotations of $0^o, 90^o, 180^o$, and 270^o. The result of rotating the original filter using four rotations in $\mathcal{G}$ is shown in the second column of Fig.1. Convoling the original input with these rotated filters yeilds four feature maps in last column of Fig.1. Now, rotating the original input by 90^o to make the green lizard. The result of convoling the green lizard with four rotations of the original filter is shown in the second last column of Fig.1. We can observe that the rotation of the input leads to the same rotation on the output.

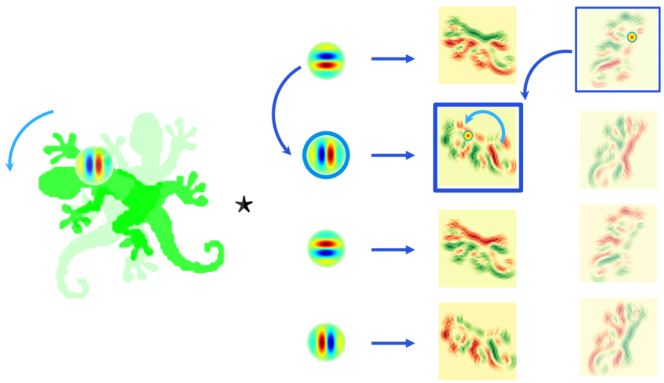

Figure 1: The equivariant property of the group convolution with respect to $\mathcal{G} = \{0, \pi/2, \pi, 3\pi/2\}$. This image is from [1].

Developments of G-CNN they have largely focused on its theoretical apsect while experimental validation has been limited to small-scale synthetic datasets (e.g., Rotated MNIST [5]). The three questions to be answered are:

1. Can G-CNN process large-scale real-world data?
2. Can G-CNN deliver on its promise of better data efficiency and robustness than convetional CNN?
3. Are they efficient enough for real-time applications?

As for the first and second question, there are a number of works that integrate variants of group convolution to CNN-based backbones to boost the performance detecting objects in in-door [17] and outdoor point clouds [13]. This strongly hints that full-fledged G-CNN is capable of at least matching the performance of convetional CNN on large-scale real-world data. Concerning the last question, G-CNN clearly have a higher computational overhead compared to convetional CNN. A possible solution to improve their inference speed is to reduce their capacity in terms of the nubmer of layers (depth) and the number of learnable parameters in each layer (width). The tradeoff of this solution is a reduction of performance. A comprehensive study on how on how the performance scales with the capacity of G-CNN is necessary to find the right tradeoff.

Another problem to be adr=dressed is the robustness to domain shifts. LiDAR-based models are notorious for their lack of robustness to domain shifts such as the change location of deployments [11] or the change of LiDAR [10]. These shifts result in the change in the appearance of objects as shown in Fig.2 and Fig.3.

Thanks to the equivariant property, G-CNN is able to disentangle geometric factors from appearance factors. As mentioned above, appearance factors are subjected to changes by domain shifts. On the other hand, geometric factor remain constant. Therefore, we expect that they have a better robustness against domain shifts.

The demonstration of robustness of G-CNN against domain shifts requires two different data distribution one for training and one for testing. It worths noticing that this is different to the traditional training and testing of deep learning models where data at both phases comes from the same distribution. We can satisfy such a requirement by taking training and testing data from datasets collected at different locations using different type of LiDAR. For example, the NuScenes dataset [3], which collected in Boston and Singapore using Velodyne HDL-32 LiDAR can be

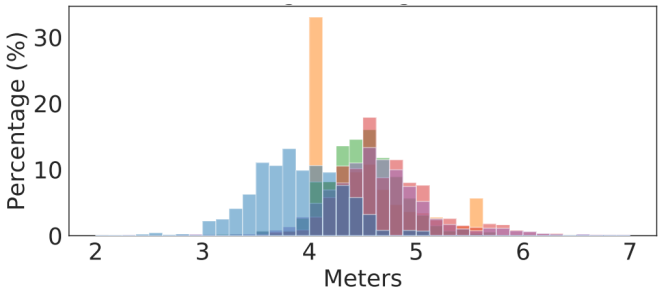

Figure 2: Changes in objects' size due to the change of location. Blue bars denote the length of cars in the KITTI dataset collected Germany, while the rest denotes length of cars in datasets collected in the USA. It is clear that cars in Germany are shorter than those in the USA. This image is from [11].

(a) Velodyne HDL-64 (b) Velodyne HDL-32

Figure 3: Changes in objects' apperances due to the change of LiDAR model. This image is from [10].

used for training. Test data can be taken from the ZOD dataset [2] which was collected in Europe using Velodyne VLS-128 LiDAR.

New domain adaptation methods will be derived to enable G-CNN to overcome the domain shifts that are challenging for them. The conventional practice to adapt CNN from a source domain, which is a dataset that has ground truth, to a new domain, which is a dataset that does not have ground truth, is to leverage self-training. This is an iterative process where the model trained at the previous iteration generates labels to train the model at the current iteration. At the first iteration, the labels are made by performing inference with the model trained in the source domain.

To augment the quality of labels at any iteration, several test-time augmentation techniques are used. The most popular among them is the gemoetric augmentation. This technique first to apply several rigid body transformation to the original input. Next, the model produces its predictions for each transformed input. Then, the prediction of each transformed input undergo the inverse of the transformation applied to its corresponding input, yeilding a set of predictions. Finally, the set of predictions is aggregated into a single prediction using heuristics such as choosing the prediction that is the most confident or average all predictions [9]. This test-time augmentation technique is based on the heuristic that neural networks provide different prediction given different transformation applied to an input. By aggregating the set of predictions made at various transformation, we can improve the precision of the final prediction.

In contrast to CNN, transforming the input of a G-CNN leads to the application of the same transformation to the G-CNN's output. Therefore, the need for such a geometric test-time augmentation is dismissed. Instead, we would like to explore learning techniques that leverage similarity among points belong to the same object (e.g., colors [8] or deep features [16]) to address the domain adaptation challenge.

We will consider the following experimental scenario. A calibrated multi-sensor system (Lidar and Stereo-Vision) will be mounted on a ground robot and manually driven in an unknown and dynamic environment. The acquired data will be processed off-line to produce the multi-layer representation of the environment. This multi-layer representation will allow to define a desired trajectory (e.g. way-points, topological graph, ...) to be executed by the robot. Given the desired trajectory the robot will localize itself in the multi-layer representation and navigate autonomously towards the goal. The global framework will be implemented in C/C++ under ROS2 and evaluated using datasets acquired by our instrumented robots.

Work plan

The work will be decomposed with incremental steps as follows:

1. Bibliography on hybrid AI (gemetric informed networks and physic informed networks)

2. Choice and setup of a simulation environment and databases selection
3. Design of the hybrid approach
4. Simulation and tuning of the hybrid approach
5. Comparison with the state of the art techniques
6. Experimental results on real data
7. Writing of reports and conference papers
8. Improvement on the hybrid approach
9. Experimental results on real data acquired on ACENTAURI robots
10. Writing Phd Thesis and journal papers

Skills

The candidate is expected to have a Master in Robotics or in Computer Science, as well as solid skills in software development (LINUX, ROS2, Git, MATLAB, C/C++, Python, Pytorch). He/she must also be highly motivated for multidisciplinary studies and all aspects of research ranging from fundamental to experimental work. A good level of written/spoken English is also important.

How To Apply

Interested candidates must send to Ezio Malis at ezio.malis@inria.fr the following documents:

- Motivation letter
- Curriculum vitæ including the list of the scientific publications
- Bachelor and Master's transcript
- Letter of recommendations (at least the Master thesis supervisor)
- Letter of recommendation of the PhD thesis supervisor

All the requested documents must be gathered and concatenated in a single PDF file named in the following format: {LAST NAME of the candidate}_{Last Name of the supervisor}_June_2025.pdf.

References

[1] Tutorial on group convolutional networks - ammi geometric deep learning course. https://colab.research.google.com/drive/1h7U15-qFC2yy6roRIfLPk5TSlo6sONsm. Accessed: 2025-02-12.

[2] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson. Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20178–20188, 2023.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[4] R. Kondor and S. Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International conference on machine learning*, pages 2747–2755. PMLR, 2018.

[5] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, pages 473–480, 2007.

[6] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021.

[7] Z. Liu, E. Malis, and P. Martinet. Adaptive learning for hybrid visual odometry. *IEEE Robotics and Automation Letters*, 9(8):7341–7348, 2024.

[8] T.-Y. Pan, C. Ma, T. Chen, C. P. Phoo, K. Z. Luo, Y. You, M. Campbell, K. Q. Weinberger, B. Hariharan, and W.-L. Chao. Pre-training lidar-based 3d object detectors through colorization. In *Proceedings of 2024 International Conference on Learning Representations*, 2024.

[9] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021.

[10] D. Tsai, J. S. Berrio, M. Shan, S. Worrall, and E. Nebot. See eye to eye: A lidar-agnostic 3d detection framework for unsupervised multi-target domain adaptation. *IEEE Robotics and Automation Letters*, 7(3):7904–7911, 2022.

[11] Y. Wang, X. Chen, Y. You, L. E. Li, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020.

[12] M. Weiler, F. A. Hamprecht, and M. Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018.

[13] H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang. Transformation-equivariant 3d object detection for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2795–2802, 2023.

[14] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma. Bridging the domain gap for multi-agent perception. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6035–6042. IEEE, 2023.

[15] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022.

[16] J. Yin, D. Zhou, L. Zhang, J. Fang, C.-Z. Xu, J. Shen, and W. Wang. Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *European conference on computer vision*, pages 17–33. Springer, 2022.

[17] H.-X. Yu, J. Wu, and L. Yi. Rotationally equivariant 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1456–1464, 2022.