Spherical Vision for Mobile Robotics

Marco La Cascia[1,*,†], Liliana Lo Presti[1,†]

[1]Dipartimento di Ingegneria, Università degli Studi di Palermo, Italia

Abstract

Recent advances in artificial intelligence open the door to the development of self-aware robots, capable of autonomous and proactive perception, action and adaptation in dynamic environments. In particular, the current state of the art allows to equipe a robot with motion and path planning routines, pre-trained modern models to detect and recognize objects and pre-trained models for text and language analysis. Such a robot might be defined as an AI agent with minimal knowledge of itself and the world.

This paper mainly illustrates the current progress within the PNRR CAESAR Project with regard to the development of computer vision techniques for autonomous robotics. The goal is to understand how omnidirectional vision techniques can be used to increase over time the robot's knowledge of both the environment in which it operates and its limited interaction capabilities, allowing the emergence of self-aware behaviors. This is achieved through continuous exploration and learning of the unknown environment to adapt/expand the robot's knowledge based on accumulated experiences.

Keywords

omnidirectional vision, robotics, depth, detection

1. Introduction

Although AI is fastening the automation of several tasks, there are limitations in the development of AI agents aware both of their surroundings and of their own capabilities. This yields to the implementation of ad-hoc solutions in practical scenarios and poses several challenges when the robots must operate in dynamic environments, a scenario which, to date, constitutes one of the main pillars of the EU research programme.

Furthermore, new regulations impose explainability requirements for automated decision-making. Explainability and self-awareness of agents represent the future of robot development and will impact their social acceptance, being the basis of user trust.

In this context, the PNRR CAESAR project has two main objectives:

- Implement AI methods that support the agent's capabilities to generate self-aware behaviors.
- Investigate techniques that allow the agent to explain its motivations and actions.

To reach this goal, computer vision techniques for scene understanding and object perception with minimal apriori knowledge are required. Our main goal is the development of techniques for visual exploration of unknown areas to support visual planning, navigation, and object perception by taking advantage of spherical cameras and epipolar geometry of the scene. By visual and spatial exploration of the environment, the robot can infer what it can/cannot do also considering its body (i.e., which objects the robot can reach considering its height or what the robot cannot pick up considering the strength of its arms). In practice, the robot learns to adapt models of the scene and of its own interaction capabilities, and accumulates self-conscious knowledge.

The peculiar choice in the development of this project is the focus on spherical cameras. Indeed, in our previous study [1], we have shown that the geometry of spherical cameras can be used to infer the distance of objects from the camera in an exact way with minimal knowledge about the camera height.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.

†These authors contributed equally.

marco.lacascia@unipa.it (M. La Cascia)

0000-0002-0877-7063 (M. La Cascia); 0000-0001-7116-9338 (L. Lo Presti)

The project activities build on these results, taking into account the need to consider the changes in camera pose over time caused, for instance, by the movement of the robot in the scene. To gain a better understanding of the environment, we have systematically studied the use and adaptation of depth estimation models to automatically derive information about the scene. Finally, we are also working on object perception starting from limited a priori knowledge.

2. Related Work

Based on the literature [2, 3], AI agents become self-conscious by providing them with the ability to explore the world through movement or visual perception.

Embodied visual exploration [4] and object finding/navigation [5] require fundamental perception and navigation capabilities, which benefit from first-person representations of the world. First-person representations are linked with the agent's self-awareness while performing actions [6] and can be made easier by adopting omnidirectional/spherical cameras onboard the robot.

In past works [7, 8, 9, 10], omnidirectional cameras were mounted on robots or deployed in the environment to capture information about the scene and improve robot navigation. The adopted omnidirectional cameras consisted of a traditional video camera that recorded the scene reflected on a mirror of known shape. This last information was used to correct the distortion and reconstruct the scene. These types of cameras are not exactly 360° cameras because it is impossible to film everything above the mirror, and for this reason the cameras were placed high on the robot. Nevertheless, in several works these cameras are used for robot navigation. For instance, in [9], an agent uses images captured by omnidirectional cameras to learn how to navigate a mobile robot in its working environment. The work applies reinforcement learning techniques to develop a totally autonomous system. In [7], the work is extended using a network of calibrated omnidirectional cameras to have a complete view of the robot's surroundings.

In our research, we mounted a 360° camera on top of the robot to collect spherical images of its surroundings, and avoiding dealing with a camera network. The robot will use information derived from the spherical vision system to refine its location in the environment and to gather awareness about its capabilities. Only recently spherical cameras are starting to be used in robotics. In [11] a guiding robot using 360° cameras is proposed. The robot aims to bring people to and from specific places within the environment. The robot detects and tracks the people it is guiding, trying to never lose sight of them. This is a very recent work indicating the potentialities of 360° cameras for service robots. Several works show that pedestrian detection and tracking can be facilitated by using equirectangular images. In [12, 13], images from the spherical camera are used to simulate virtual PTZ cameras, which allows not only testing the algorithms for PTZ cameras (which are known to suffer from the problem of reproducibility of experimental results [14]) but also to track pedestrians by combining particle filtering and Camshift algorithms in [13] or by predicting the movements of the target through a motion model based on a deep neural network [12] . In [15], spherical images were used to develop a multi-object tracker that uses estimates of the object distance from the camera [1] to track pedestrians on the ground-plane. Knowledge of the distances of pedestrians to the camera has been used successfully to human activity understanding, for example to detect whether or not a person is within an area of interest [16, 17].

However, at our best knowledge, no work deals with the use of spherical cameras in a mobile setting, which introduces new issues such as the need to stabilize captured videos and account for camera pose changes. These issues are relevant when using spherical cameras on-board robotic platforms.

3. Smart Spherical Cameras and Spherical Vision

A vision system integrates software and hardware technologies and methods to provide image analysis in specific fields of application. In a spherical vision system the analyzed images are spherical and acquired by 360° cameras. To ensure real-time image processing, the vision system is independent of

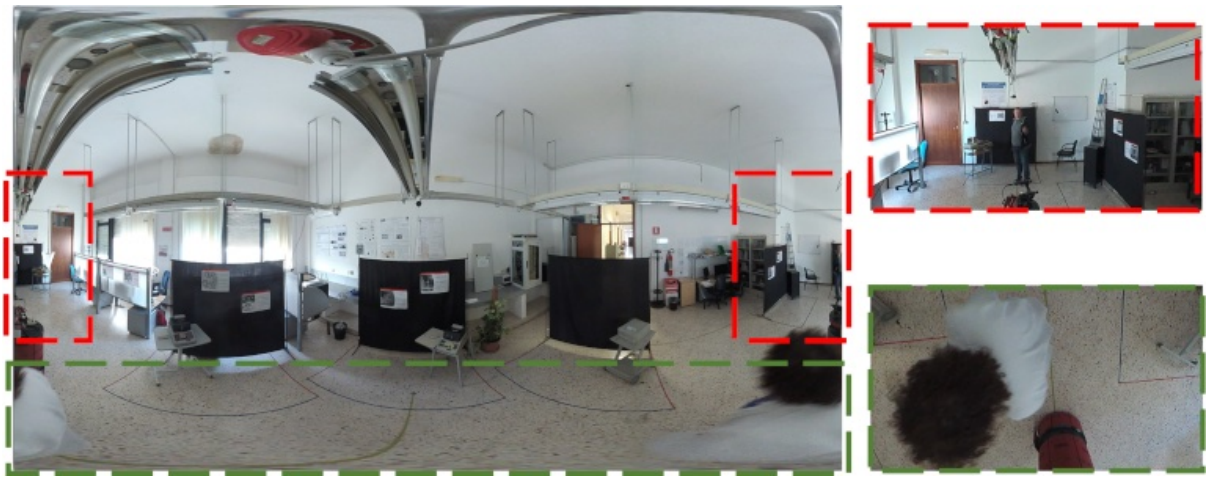

Figure 1: Example of equirectangular image (on the left) and planar projections (on the right). The camera was mounted on top of the scene.

the robot, i.e. image processing does not consume the robot's computational resources. The vision system enhances the spherical cameras with software that makes the camera intelligent in the sense that the software will allow the extraction of information useful for the robot to perform a specific task.

A spherical image is stored as a rectangular multichannel array (like traditional images) by mapping the sphere to a rectangle via an equirectangular projection [1] An example of an equirectangular image is shown in Fig. 1.

As shown in Fig. 1, the equirectangular image captures the entire scene with one shot. Analysis of the environment can be done, for instance, by processing the equirectangular image. Due to the projection used, the areas of the scene are represented in the equirectangular image with different resolutions, i.e. there are more pixels available to represent the areas around the poles of the sphere and the projection introduces distortions into these areas. Spherical views are continuous, but the equirectangular projection cuts the sphere along a meridian to map the spherical surface into a rectangle. There is therefore a form of circularity in the resulting projected image. Thus, it is sometimes more convenient to use planar projections of the image to deal with these issues [13, 12].

The circularity of the equirectangular image, the non-uniform distribution of the resolution with which the details of the scene are represented, the deformation and the high dimensionality of the equirectangular images are all aspects to be considered in the analysis of such images.

3.1. Understanding the Camera-Scene Geometry

In [1], 3D information of points on the ground can be easily recovered by pure geometry under simple assumptions. In particular, the method assumes that the camera is in a canonical pose, namely the acquired images are gravity-aligned, and the height h_c of the camera is known. Under these assumptions, the method exploits the fact that, as a result of the acquisition and projection processes, all the points of the 3D world equidistant from the camera and belonging to a plane parallel to the ground are projected onto the same horizontal line in the equirectangular image. Therefore, given a point P on the equirectangular image, there is a mathematical relationship between the image line on which the point P lies and the distance of the corresponding point P_w in 3D world coordinates from the camera. This fact has been exploited for pedestrian tracking and activity understanding [16, 17, 15].

However, when the camera is mounted on top of the robot, for instance on the head of an anthropomorphic robot such as the Pepper robot, and more in general in a mobile setting, additional image processing techniques are required. For example, since the robot is moving, the assumption that the camera plane is horizontal to the ground plane may not be satisfied. More in general, any assumption on the camera pose can be problematic and may lead to inaccuracy of the results.

Within the CAESAR project, we developed methods for spherical camera pose estimation and

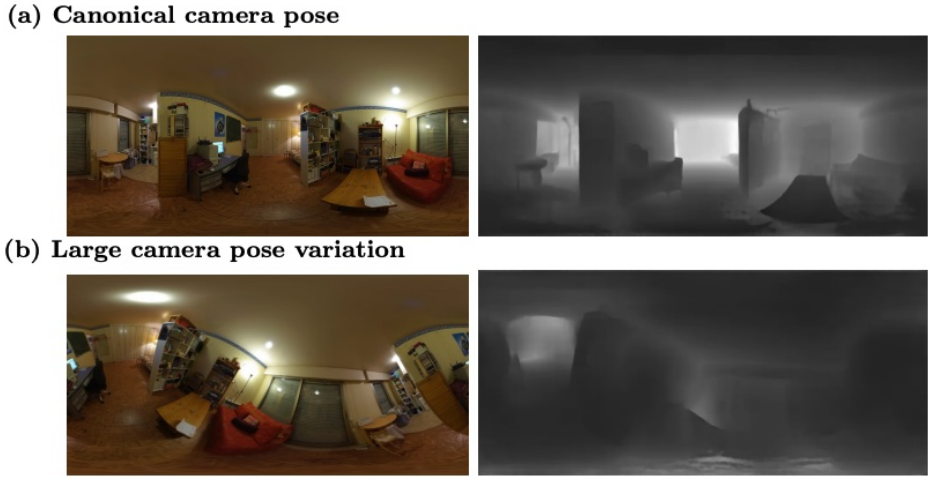

Figure 2: Examples of depth maps estimated from a gravity aligned image (top line) and an image acquired with a rotated camera (bottom line) using [19].

epipolar geometry exploitation to enhance the use of spherical cameras in mobile settings. In particular, we developed numerical methods capable of recovering extrinsic camera parameters from minimal knowledge of the scene.

3.2. Depth estimation from 360° images

Within the CAESAR project, we have the goal of implementing computer vision methods to help the robot gather knowledge about the spatial layout of the environment and enable self-perception, enhancing a first-person representation of its capabilities/limits.

To this end, we focused on depth estimation models from spherical images [18, 19, 20, 21, 22] in order to support the robot in immersive navigation and understanding of the scene. As already mentioned, the spherical camera integrated on the robotic platform can undergo variations in the camera pose. Thus, our first task was to study how such variations in the camera pose affect the performance of depth estimation models. In this sense, we conducted a systematic and reproducible study of the robustness of depth estimation models from equirectangular images. Our study showed that:

- *Perspective-based models*, such as [18], are designed to handle planar images and can hardly be effective on equirectangular images, even when using cubemap reprojection techniques;
- *Spherical-based models*, such as [19, 20, 21, 22], although specifically designed to process equirectangular images, they are only able to handle gravity-aligned images to a certain extent and are sensitive to changes in camera orientation.

Fig. 2 shows examples of depth maps estimated from a gravity aligned image (top line) and an image acquired with a rotated camera (bottom line). The depth maps are obtained by using the model [19].

Our results highlight the need to account for camera pose variations in depth estimation models for equirectangular images. We are currently investigating the design of an encoder-decoder architecture to estimate camera pose-invariant depth maps.

3.3. Unknown Object Detection

Within the CAESAR Project, the robot is equipped with an initial knowledge base that includes pre-trained object detection models. It is well-known that these models are generally trained on a limited set of categories. Models trained on the COCO-dataset can recognize among 80 categories, while ImageNet-21K includes 21,841 categories but it is not very much used for pre-training/training vision models.

Considering that the robot should be able to work on unknown, dynamic environments, it is necessary to deal with situations where unknown objects or out-of-distribution objects are in the environment. Withing the CAESAR Project, our main goal is that of devising a model that helps the robot to detect objects that are in the scene. In practice, we are designing CNN-based models for class-agnostic object detection. Post-processing techniques will then use prior knowledge or external information source to help the robot understanding the class of the detected object and its functionality.

4. Conclusions and Future Work

This paper summarizes the development of computer vision algorithms and models to address new problems in spherical signal processing and mobile robotics. Most of the presented work has been carried on within the PNRR CAESAR Project. The findings of our research activities are of interest in fields beyond mobile robotics. Indeed, the developed methods could be extended and be useful for building other applications such as: the monitoring of large touristic environments or industrial environments, in the retail field to enhance customer interactions, in the context of security and smart cities to better patrol the city.

Acknowledgments

The work has been partially supported by PNRR MUR CAESAR (grant n. U-GOV PRJ-1637).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] G. Mazzola, L. Lo Presti, E. Ardizzone, M. La Cascia, A dataset of annotated omnidirectional videos for distancing applications, Journal of Imaging 7 (2021) 158.
[2] D. Legrand, Pre-reflective self-consciousness: on being bodily in the world, Janus head 9 (2007) 493–519.
[3] V. Gallese, The inner sense of action. agency and motor representations, Journal of Consciousness studies 7 (2000) 23–40.
[4] S. K. Ramakrishnan, D. Jayaraman, K. Grauman, An exploration of embodied visual exploration, International Journal of Computer Vision 129 (2021) 1616–1649.
[5] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, E. Wijmans, Objectnav revisited: On evaluation of embodied agents navigating to objects, arXiv preprint arXiv:2006.13171 (2020).
[6] L. Farina, Artificial intelligence systems, responsibility and agential self-awareness, in: Conference on Philosophy and Theory of Artificial Intelligence, Springer, 2021, pp. 15–25.
[7] E. Menegatti, G. Cicirelli, C. Simionato, T. D'Orazio, H. Ishiguro, Explicit knowledge distribution in an omnidirectional distributed vision system, in: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566), volume 3, IEEE, 2004, pp. 2743–2749.
[8] E. Menegatti, T. Maeda, H. Ishiguro, Image-based memory for robot navigation using properties of omnidirectional images, Robotics and Autonomous Systems 47 (2004) 251–267.
[9] E. Menegatti, G. Cicirelli, C. Simionato, A. Distante, E. Pagello, et al., Reinforcement learning based omnidirectional vision agent for mobile robot navigation, in: Workshop Robotica del IX Convegno della Associazione Italiana Intelligenza Artificiale, 2004.

[10] E. Menegatti, A. Pretto, E. Pagello, A new omnidirectional vision sensor for monte-carlo localization, in: RoboCup 2004: Robot Soccer World Cup VIII 8, Springer, 2005, pp. 97–109.
[11] A. Bacchin, F. Berno, E. Menegatti, A. Pretto, People tracking in panoramic video for guiding robots, in: International Conference on Intelligent Autonomous Systems, Springer, 2022, pp. 407–424.
[12] L. Lo Presti, M. La Cascia, Deep motion model for pedestrian tracking in 360 degrees videos, in: International Conference on Image Analysis and Processing, Springer, 2019, pp. 36–47.
[13] V. Monteleone, L. Lo Presti, M. La Cascia, Particle filtering for tracking in 360 degrees videos using virtual ptz cameras, in: Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part I 20, Springer, 2019, pp. 71–81.
[14] G. Chen, P.-L. St-Charles, W. Bouachir, G.-A. Bilodeau, R. Bergevin, Reproducible evaluation of pan-tilt-zoom tracking, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE, 2015, pp. 2055–2059.
[15] L. Lo Presti, G. Mazzola, G. Averna, E. Ardizzone, M. La Cascia, Depth-aware multi-object tracking in spherical videos, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 362–374.
[16] L. Lo Presti, G. Mazzola, M. La Cascia, Activity monitoring made easier by smart 360-degree cameras, in: European Conference on Computer Vision, Springer, 2022, pp. 270–285.
[17] F. Becattini, G. Becchi, A. Ferracani, A. D. Bimbo, L. L. Presti, G. Mazzola, M. L. Cascia, F. Cunico, A. Toaiari, M. Cristani, et al., I-mall an effective framework for personalized visits. improving the customer experience in stores, in: Proceedings of the 1st Workshop on Multimedia Computing towards Fashion Recommendation, 2022, pp. 11–19.
[18] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, H. Zhao, Depth anything v2, Advances in Neural Information Processing Systems 37 (2024) 21875–21911.
[19] C. Zhuang, Z. Lu, Y. Wang, J. Xiao, Y. Wang, Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation, in: Proceedings of the AAAI conference on artificial intelligence, volume 36, 2022, pp. 3653–3661.
[20] N.-H. A. Wang, Y.-L. Liu, Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation, Advances in Neural Information Processing Systems 37 (2024) 127739–127764.
[21] F.-E. Wang, Y.-H. Yeh, Y.-H. Tsai, W.-C. Chiu, M. Sun, Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation, IEEE transactions on pattern analysis and machine intelligence 45 (2022) 5448–5460.
[22] G. Pintore, M. Agus, E. Almansa, J. Schneider, E. Gobbetti, Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11536–11545.