Electronic Letters on Computer Vision and Image Analysis 14(3):49-51, 2015

Modeling the environment with egocentric vision systems

Alejandro Rituerto

Dept. Informática e Ingeniería de Sistemas, I3A−Universidad de Zaragoza, Zaragoza, Spain
Advisors: Jos Jesús Guerrero, Ana C. Murillo
Date and location of PhD thesis defense: November 21, 2014, Universidad de Zaragoza

Received 23rd March 2015; accepted 27th July 2015

Abstract

More and more intelligent systems, such as robots or wearable systems, are present in our everyday life. This kind of systems interact with the environment so they need suitable models of their surrounding. Depending on the tasks that they have to perform, the information required in those models changes: from highly detailed 3D models for autonomous navigation systems, to semantic models including important information for the user. These models are created using the sensory data provided by the system (Fig. 1). Cameras are an important sensor included in most intelligent systems thanks to their small size, cheap prices and the great amount of information that they provide. This thesis studies and develops new methods to create models of the environment with different levels of precision and semantic information. There are two coomon key-points in the subsequent presented approaches:

- *The use of egocentric vision systems.* All the vision systems and image sequences used in this thesis characterize for a first-person (egocentric) point of view.
- *The use of omnidirectional vision.* This kind of vision systems provide much more information that conventional cameras thanks to their wide field of view.

This thesis studies how computer vision can be used to create different models of the environment. To test our proposals, different cameras have been used, both in robotic and wearable platforms.

Omnidirectional cameras for visual SLAM. The SLAM (Simultaneous Localization and Mapping) problem is one of the essential challenges for current robotics. We present our approach to integrate the Spherical Camera Model for catadioptric systems in a Visual-SLAM application. The Spherical Camera Model is a projection model that unifies the projection of central catadioptric and conventional cameras. The comparison of the performance of a visual SLAM system using monocular conventional and omnidirectional vision confirms that the latter produces better trajectory and orientation estimation. *Associated publications*: [2, 1].

Correspondence to: arituerto@unizar.es

Recommended for acceptance by Jorge Bernal
DOI http://dx.doi.org/10.5565/rev/elcvia.739
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

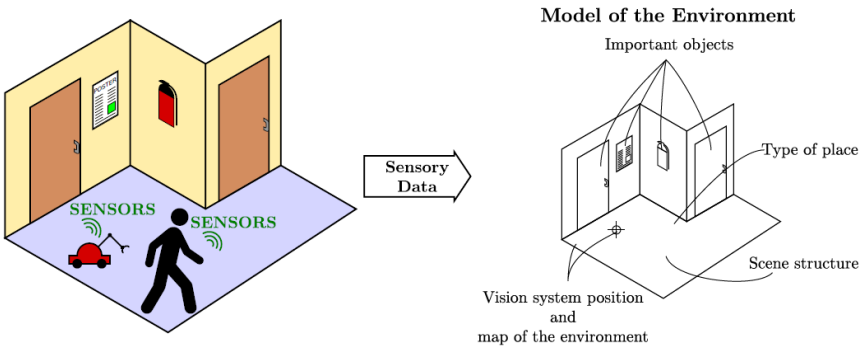

Figure 1: Intelligent systems need models of their environment to perform the tasks they are designed for. These models are the result of translating sensory data into useful information for the system.

Semantic labeling for indoor topological mapping using a wearable catadioptric system. Richer semantic representations of the environment may allow autonomous systems to perform higher level tasks and provide better human-robot interaction. We present a new omnidirectional vision based scene labeling approach for augmented indoor topological mapping. Our proposal includes novel ideas in order to augment the semantic information of a typical indoor topological map: we pay special attention to the semantic labels of the different types of transitions between places, and propose a simple way to include this information to build a topological map as part of the criteria to segment the environment. This approach is built on an efficient catadioptric image representation based on the Gist descriptor, which is used to classify the acquired views into types of indoor regions. The basic types of indoor regions considered are *Place* and *Transition*, farthest divided into more specific subclasses, e.g*Transition* into *Door*, *Stairs* or *Elevator*. Besides using the result of this labeling, the proposed mapping approach includes a probabilistic model to account for spatio-temporal consistency. All the proposed ideas have been evaluated in a new indoor data-set acquired with our wearable catadioptric vision system, showing promising results in a realistic prototype. *Associated publications*: [6].

Line Image Signature, global descriptor for Scene Understanding. Pursuing similar objectives, we propose a novel line-based image global descriptor that encloses the structure of the scene observed. This descriptor is designed with omnidirectional imagery in mind, where observed lines are longer than in conventional images. The descriptor has been tested with two omnidirectional systems: a catadioptric camera and panoramic images. Experiments show how the proposed descriptor can be used for indoor scene recognition with comparable results to state-of-the-art global descriptors. Additional advantages of the new descriptor are higher robustness to rotation, compactness, and easier integration with other scene understanding steps using the observed lines. *Associated publications*: [3, 8].

Building a hierarchical vocabulary from an image sequence. Vision based recognition approaches frequently use quantized feature spaces, commonly know as Bag of Words (BoW) or vocabulary representations. A drawback using standard BoW approaches is that conceptual or semantic information is not considered as criteria to group the visual features into words. To solve this challenging task, we study how to leverage the standard vocabulary construction process to obtain a more meaningful visual vocabulary for particular applications using image sequences. We take advantage of spatio-temporal constraints and prior-knowledge about the position of the camera. The key contribution of our approach is to define a new method to incorporate tracking information to the process of vocabulary construction, and to add geometric cues to the appearance descriptors. Motivated by long-term indoor robotic appli-

cations, we focus in a robot camera pointing to the ceiling, which facilitates the capture of more stable regions of the environment, improving long term operation and the discovery of repetitive and representative elements. The experimental validation shows how our vocabulary models the environment in more detail than standard vocabulary approaches, while keeping comparable recognition performance. We show different robotic tasks that could benefit of the use of our visual vocabulary approach, such as place recognition or object discovery. *Associated publications*: [4].

3D Spatial layout propagation. Indoor scene understanding from monocular images has been widely studied, and a common initial step to solve this problem is the estimation of the layout of the scene, the basic 3D structure. Many previous approaches obtain the layout from a single image, however, we address the problem of scene layout propagation along a video. Our approach uses a Particle Filter framework to propagate the scene layout obtained using a single image technique on the initial frame. We propose how to generate, evaluate and sample new layout hypotheses for the scene on each of the frame. The intuition we follow is that we can obtain better layout estimation at each frame through propagation than running separately at each image. The experimental validation is run on a publicly available indoor data-set and shows promising results for the layout computed using our approach, without the need of estimating accurate 3D maps. Additionally, they demonstrate how this layout information can be used to improve detection tasks. *Associated publications*: [5, 7].

References

[1] Rituerto, A., Puig, L., and Guerrero, J. Comparison of omnidirectional and conventional monocular systems for visual SLAM. In *Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras (OMNIVIS) (Best paper award)*.

[2] Rituerto, A., Puig, L., and Guerrero, J. Visual SLAM with an omnidirectional camera. In *International Conference on Pattern Recognition (ICPR)*, pages 348–351.

[3] Rituerto, A., Murillo, A. C., and Guerrero, J. Line image signature for scene understanding with a wearable vision system. In *International SenseCam & Pervasive Imaging Conference*, pages 16–23.

[4] Rituerto, A., Andreasson, H., Murillo, A., Lilienthal, A., and Guerrero, J. (2014a). Building a hierarchical vocabulary from an image sequence. *Pattern Recognition (Under review)*.

[5] Rituerto, A., Manduchi, R., Murillo, A. C., and Guerrero, J. 3D spatial layout propagation in a video sequence. In *International Conference on Image Analysis and Recognition (ICIAR)*.

[6] Rituerto, A., Murillo, A., and Guerrero, J. Semantic labeling for indoor topological mapping using a wearable catadioptric system. *Robotics and Autonomous Systems, Special Issue Semantic Perception, Mapping and Exploration*, 62(5):685–695.

[7] Rituerto, A., Murillo, A. C., and Guerrero, J. 3D layout propagation to improve object recognition in egocentric videos. In *Assistive Computer Vision and Robotics (ACVR)*.

[8] Rituerto, A., Murillo, A. C., and Guerrero, J. Line-based global descriptor for omnidirectional vision. In *IEEE International Conference on Image Processing (ICIP)*.