Robotics and Autonomous Systems 61 (2013) 259–270

Contents lists available at SciVerse ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Using mutual information for appearance-based visual path following

Amaury Dame [a,1], Eric Marchand [b,*]

[a] CNRS, IRISA, INRIA Rennes, France

[b] Université de Rennes 1, IRISA, INRIA Rennes, France

ARTICLE INFO

Article history:
Received 17 June 2011
Received in revised form
18 October 2012
Accepted 9 November 2012
Available online 20 December 2012

Keywords:
Navigation
Mutual information
Visual servoing

ABSTRACT

In this paper we propose a new way to achieve a navigation task (visual path following) for a non-holonomic vehicle. We consider an image-based navigation process. We show that it is possible to navigate along a visual path without relying on the extraction, matching and tracking of geometric visual features such as keypoint. The new proposed approach relies directly on the information (entropy) contained in the image signal. We show that it is possible to build a control law directly from the maximization of the shared information between the current image and the next key image in the visual path. The shared information between those two images is obtained using mutual information that is known to be robust to illumination variations and occlusions. Moreover the generally complex task of features extraction and matching is avoided. Both simulations and experiments on a real vehicle are presented and show the possibilities and advantages offered by the proposed method.

1. Introduction

In recent years, robot localization and navigation have made considerable progress. Navigation can be seen as the ability for a robot to move autonomously from an initial position to a desired one (which may be far away from the initial one). Thanks to sensor based navigation, we have seen autonomous robots in various challenging areas (from highways to deserts and even on Mars). Nevertheless the design of these autonomous robots usually relies on more than one sensor (camera, stereo sensors, lidar, GPS, …). In this paper, we propose a new method that demonstrates the capability of a mobile robot to navigate autonomously using the information provided by a monocular camera. Furthermore we will show that the proposed approach does not require any tracking or matching process which is usually a bottleneck for the development of such an approach.

Most navigation approaches consider a (partial) 3D reconstruction of the environment [1–6], leading to SLAM-like techniques. Such solutions are attractive, since the navigation task will be achieved using a classical pose-based control of the robot in the metric space. Within this context, during a learning step the environment is reconstructed using bundle adjustment approaches [3], Kalman/particle filters based approaches [4], visual odometry [7,6]. Despite the complexity of the underlying problem, SLAM has proved to be a viable solution to create accurate maps of the environment [4,5] even in large ones [8]. In this context, the control of the robot during the navigation task is a well known problem and the main difficulties here are (i) the complexity of the initial reconstruction step and (ii) the matching of visual features observed during the learning step with current observations. With a monocular camera as unique sensor, these are mainly computer vision issues.

Another class of techniques relies on the definition of a visual path: the appearance-based approaches [9–15]. The trajectory is no longer described in the metric space but as a set of reference images. A 2D visual servoing step allows the robot to navigate from its current position to the next key image. When the robot gets close to this image, a new key image is selected. In this context, the environment can be modeled by a graph whose nodes are the key images. A visual path in the environment is nothing but a path in the graph [11]. Working directly in the sensor space, such approaches do not require prior 3D reconstruction step. In some cases, partial reconstruction has to be considered. In [12,16,17] a part of the epipolar geometry that links the current and key images is considered in order to predict the location of currently not visible features and ensure a robust tracking. In [13] homography computation w.r.t., the reference images allow to precisely localize the robot. In any case, the learning step of these appearance-based approaches is far less complex since it does not require any prior 3D reconstruction.

Nevertheless at navigation level, for both pose-based or most of the image-based visual navigation approaches, features have to be extracted or tracked in the image stream and matched with either the 3D database or key images to design the control

* Corresponding author. Tel.: +33 2 99 83 82 68.
E-mail address: Eric.Marchand@irisa.fr (E. Marchand).

[1] A. Dame is now with the Active Vision Group, Department of Engineering Science, University of Oxford.

0921-8890/$ – see front matter
doi:10.1016/j.robot.2012.11.004

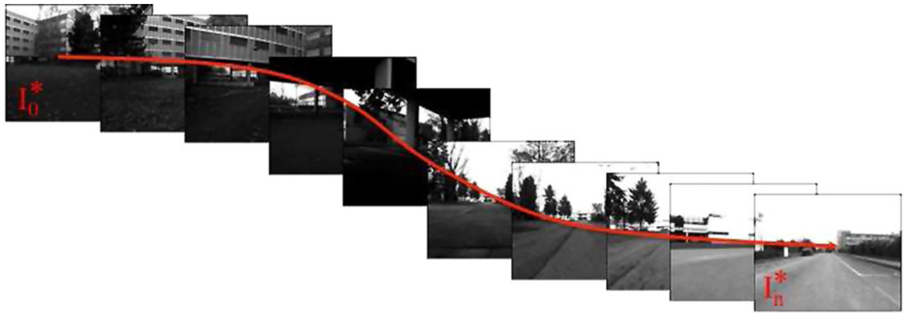

Fig. 1. Key images that define the visual path. This visual path is learned prior to the navigation step.

law. Robust extraction and real-time spatio-temporal tracking or matching of these visual cues are non trivial tasks and also one of the bottlenecks of the expansion of visual navigation. It could be then very interesting to consider directly a comparison with current image and next keyframe to control the robot. In [15], this relative orientation is computed in the Fourier space and the control law is directly computed from the orientation difference. A similar approach was proposed by [9] but used cross-correlation. In [18,19], it has been shown that no other information than the image intensity (the pure image signal) needs to be considered to control the robot motion and that these difficult tracking and matching processes can be totally removed. Although very efficient, this approach is sensitive to light variations and, thus, can hardly be considered in an outdoor environment. In this paper, we propose a new approach that no longer relies on geometrical features nor on pixel intensity [19] but directly uses the information (entropy) contained in the image signal as proposed in [20,21]. More precisely we will consider mutual information [22,23] as a similarity criterion.

MI has been introduced in the context of information theory by Shannon [22]. It has been later considered as an image similarity measure back in the mid ninety's independently by Collignon [24] for tomographic image registration, Studholme [25] for MR and CT images, an by Viola [26] for projection images. Since then MI has become a classical similarity measure especially for multi-modal registration techniques [27] (e.g., for medical or remote sensing applications). Being closer to the signal, we will show that this approach is robust to very important illumination variations and robust to large occlusions.

Our goal is then to propose a control law that allows the robot to maximize the mutual information between the current acquired image and the next image in its visual path. This is an optimization process. We show that it is possible to compute the interaction matrix that relates the variation of the mutual information to the vehicle velocity leading to the definition of the control law. Let us emphasize that since mutual information is computed from the whole images (current and key images) it is possible to directly control the motion of the vehicle along a given path without any feature extraction or matching. Furthermore no 3D reconstruction of the environment is necessary.

We will demonstrate the efficiency of this new approach on a navigation task carried out at 0.5 m/s over 400 m. Images are acquired at 30 Hz (nearly 25.000 images were acquired and processed in real-time during this navigation task).

In this paper we will first present a general overview of the method with the learning and the navigation steps. Then Sections 2 and 3 will focus on the two parts of the navigation steps which consist of the visual servoing task and the key images selection task. Finally, simulated and experimental results are presented in Section 4.

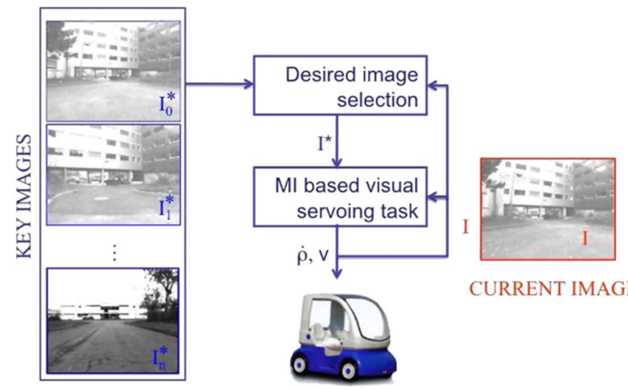

Fig. 2. Navigation based on multiple visual servoing tasks.

2. Navigation process overview

In this work, we consider a non-holonomic robot with a camera mounted on the front. Our goal is not to localize the robot within its environment (visual odometry) but only to ensure that it is able to reproduce a visual path defined as a set of images previously acquired by the camera.

2.1. Learning step: definition of the visual path

With respect to previous approaches that rely on 3D reconstruction (e.g., [3]) or even on appearance-based approaches [12], the learning step of the approach is simple. It does not require any feature extraction nor scene reconstruction: no image processing is done, only raw images are stored. The vehicle is driven manually along a desired path. While the vehicle is moving, the images acquired by the camera are stored chronologically thus defining a trajectory in the image space. Let us call $\mathbf{I}_0^*, \ldots, \mathbf{I}_N^*$ the key images that define this visual path. (See Fig. 1.)

2.2. Navigation step: following the visual path

The vehicle is initially positioned close to the initial position of the learned visual path (defined by the image $\mathbf{I}_0^*$). The navigation is performed using a visual servoing task. Fig. 2 shows the general control scheme used for the navigation. In [3,12,14] the considered control scheme is either a pose-based control law or a classical visual servoing process based on the use of visual features extracted from the current and key images ($\mathbf{I}$ and $\mathbf{I}_k^*$).

In this work the definition of a new control law is proposed. One of the original features of this work is that, rather than relying on feature extraction and tracking, we build the control law directly from the information shared by $\mathbf{I}$ and $\mathbf{I}_k^*$ measured using mutual information [22]. When the mutual information between two images is maximized, the two images are similar. We then control