Using mutual information for appearance-based visual path following

Amaury Dame[1], Eric Marchand[2]

[1]*CNRS, IRISA, INRIA Rennes, France.*
A. Dame is now with the Active Vision Group, Department of Engineering Science, University of Oxford.
[2]*Université de Rennes 1, IRISA, INRIA Rennes, France*
Eric.Marchand@irisa.fr

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 maximisation 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.

Keywords:
Navigation, mutual information, visual servoing

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 nor 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 [26, 35, 31, 8, 32, 20], 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 [31], Kalman/particle filters based approaches [8], visual odometry [19, 20]. Despite the complexity of the underlying problem, SLAM has proved to be a viable solution to create accurate maps of the environment [8][32] even in large ones [18]. 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 [24, 4, 29][33][12][7][40]. 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 [29]. 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 [33][16, 3] 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 [12] homography computation wrt. the reference images allows 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 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 [40], this relative orientation is computed in the Fourrier space and the control law is directly computed from the orientation difference. Similar approach is proposed by [24] but used cross-correlation. In [10][9], it has been shown that no other information than the image intensity (the pure image signal) need 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 outdoor environment. In this paper, we propose a new approach that no longer relies on geometrical features nor on pixels intensity [9] but uses directly the information (entropy) contained in the image signal as proposed in [13, 14]. More precisely we will consider mutual information [34][39] as a similarity criterion.

MI has been introduced in the context of information theory by Shannon [34]. It has been later considered as an image similarity measure back in the mid ninety's independently by Collignon [11] for tomographic image registration, Studholme [36] for MR and CT image, an by Viola [38] for projection image. Since then MI has become a classical similarity measure especially for multi-modal registration techniques [28] (eg, 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 meters. Images are acquired at 30Hz (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 section 2 and 3 will focus on the two parts of the navigation steps which consists of the visual servoing task and the key images selection task. Finally, simulated and experimental results are presented in section 4.

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 (eg, [31]) or even on appearance-based approaches [33], 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.

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. Figure 2 shows the general control scheme used for the navigation. In [31][33][7] the considered control scheme are either pose-based control law or consider 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 originality of this work is that, rather than relying on features extraction and tracking, we build the control law directly from the information shared by $\mathbf{I}$ and $\mathbf{I}^*_k$ measured using the mutual information [34]. When the mutual information between two images is maximized, the two images are similar. We then control the robot in order to maximize the mutual information between $\mathbf{I}$ and $\mathbf{I}^*_k$. As for any visual servoing scheme it is then necessary to exhibit the Jacobian that links the variation of the mutual information to the control input of the robot (that is the steering angle ψ or the camera rotational velocity $\dot{\rho}$) needed to follow the path with a constant translational velocity v. This process is presented in the next section. In the same time, when the vehicle reaches the neighboring key image $\mathbf{I}^*_k$, a new one $\mathbf{I}^*_{k+1}$ is selected in the visual path. To achieve a seamless switching between key images, a specific process described in section 4 is proposed.

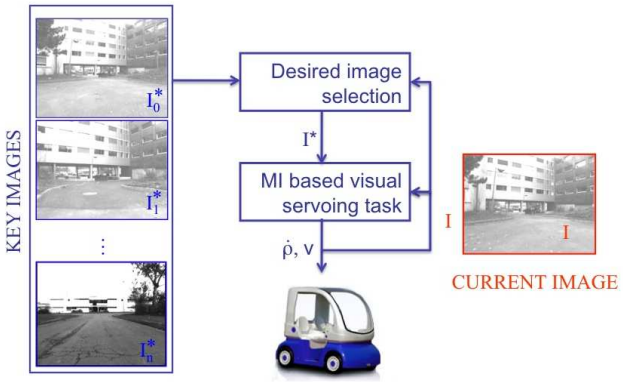

Figure 2: Navigation based on multiple visual servoing tasks.

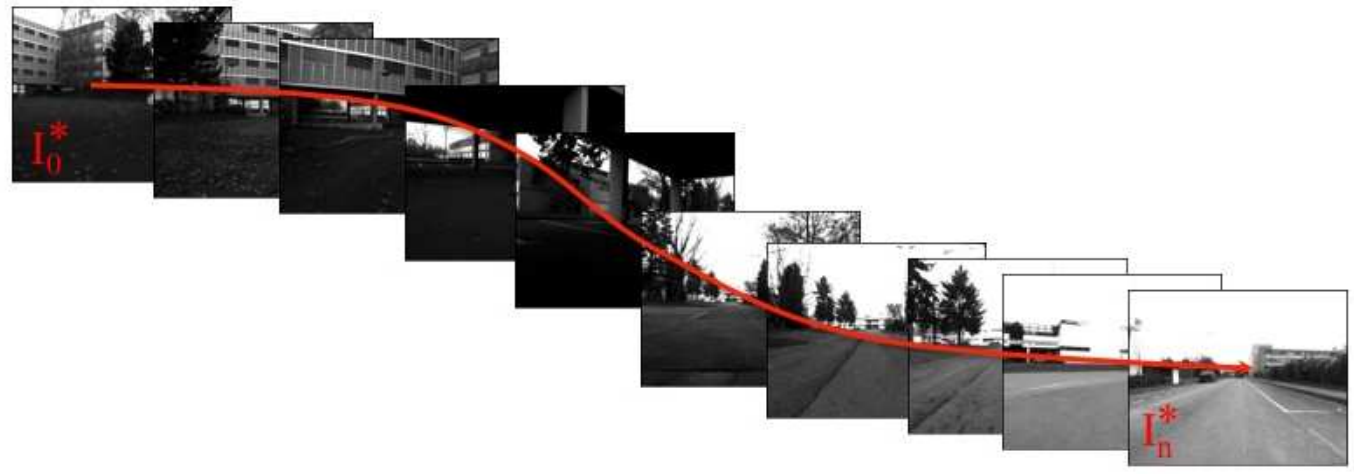

Figure 1: Key images that define the visual path. This visual path is learned prior to the navigation step.

3. Mutual information based navigation

In [14], it has been shown that it is possible to achieve 6 degrees of freedom (d.o.f) visual servoing task using only the information contained in the images acquired from the camera mounted on a robot and one reference image. The desired position of the robot is reached by maximizing the mutual information between the two images. Since mutual information is robust to illumination variations and occlusions, the use of mutual information-based visual servoing is well suited for outdoor navigation tasks.

3.1. Mutual information

In this section, a brief reminder of the definition of mutual information is given. Mutual information is the information shared by two signals (here, images). For the two signals X and Y, mutual information is given by the following equation [34]:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where $H(X)$ denotes the entropy of the signal X, that means its variability. $H(X,Y)$ denotes the joint entropy of the signals X and Y, that is the joint variability of the system defined by the two signals. By substracting the joint variability from the variabilities, as in equation 1, we obtain the shared information of the two signals that is mutual information.

In the present work we focus on mutual information between two images. The desired image (in the navigation context, desired image is the current key image of the visual path) is noted $\mathbf{I}^*$ and the current image acquired by the vehicle camera is noted $\mathbf{I}$. To address this definitions, let us consider that $\mathbf{I}$ is now a random variable and that the actual pixel intensities are samples of this random variable ($\mathbf{I}(\mathbf{x})$ being the intensity of the pixel $\mathbf{x}$).

The entropy $H(\mathbf{I})$ is a measure of variability of a random variable $\mathbf{I}$. If i are the possible values of $\mathbf{I}(\mathbf{x})$ ($i \in [0, N_{c_\mathbf{I}}]$ with $N_{c_\mathbf{I}} = 255$) and $p_\mathbf{I}(i) = \Pr(\mathbf{I}(\mathbf{x}) = i)$ is the probability distribution function of i, then the Shannon entropy $H(\mathbf{I})$ of a discrete variable $\mathbf{I}$ is given by the following expression:

$$H(\mathbf{I}) = -\sum_{i=0}^{N_{c_\mathbf{I}}} p_\mathbf{I}(i) \log\left(p_\mathbf{I}(i)\right). \tag{2}$$

The formulation can be seen as follows: since $-\log\left(p_\mathbf{I}(i)\right)$ is a measure of the uncertainty of the event i, then $H(\mathbf{I})$ is a weighted mean of the uncertainties. $H(\mathbf{I})$ is then the variability of $\mathbf{I}$.

Since a sample of $\mathbf{I}$ is in our case given by the pixel intensities $\mathbf{I}(\mathbf{x})$, the probability distribution function can be estimated using the normalized histogram of this image. The entropy can therefore be considered as a dispersion measure of the image histogram.

Following the same principle, joint entropy $H(\mathbf{I}, \mathbf{I}^*)$ of two random variables $\mathbf{I}$ and $\mathbf{I}^*$ can be defined as the variability of the couple of variables $(\mathbf{I}, \mathbf{I}^*)$. The Shannon joint entropy expression is given by:

$$H(\mathbf{I}, \mathbf{I}^*) = -\sum_{i=0}^{N_{c_\mathbf{I}}} \sum_{j=0}^{N_{c_{\mathbf{I}^*}}} p_{\mathbf{I}\mathbf{I}^*}(i,j) \log\left(p_{\mathbf{I}\mathbf{I}^*}(i,j)\right) \tag{3}$$

where i and j are respectively the possible values of the variables $\mathbf{I}$ and $\mathbf{I}^*$, and $p_{\mathbf{I}\mathbf{I}^*}(i,j) = \Pr(\mathbf{I}(\mathbf{x}) = i \cap \mathbf{I}^*(\mathbf{x}) = j)$ is the joint probability distribution function. Here, $\mathbf{I}$ and $\mathbf{I}^*$ being images, i and j are the pixel intensities of the two images and the joint probability distribution function is a normalized bidimensional histogram of the two images. As for entropy, joint entropy measures the dispersion of the joint histogram of $\mathbf{I}$ and $\mathbf{I}^*$.

The original definition of mutual information given in [34][39] can thus be used as:

$$MI\left(\mathbf{I}, \mathbf{I}^*\right) = \sum_{i,j} p_{\mathbf{I}\mathbf{I}^*}(i,j) \log\left(\frac{p_{\mathbf{I}\mathbf{I}^*}(i,j)}{p_\mathbf{I}(i) p_{\mathbf{I}^*}(j)}\right). \tag{4}$$

The probabilities $p_{\mathbf{I}\mathbf{I}^*}$, $p_\mathbf{I}$ and $p_{\mathbf{I}^*}$ involved in the computation of MI are obtained by normalizing the histograms and joint his-

togram of the images. Their analytical formulations are given by:

$$p_{\mathbf{I}\mathbf{I}^*}(i,j) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \phi\left(i - \mathbf{I}(\mathbf{x})\right) \phi\left(j - \mathbf{I}^*(\mathbf{x})\right) \quad (5)$$

$$p_\mathbf{I}(i) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \phi\left(i - \mathbf{I}(\mathbf{x})\right) \quad (6)$$

$$p_{\mathbf{I}^*}(j) = \frac{1}{N_\mathbf{x}} \sum_\mathbf{x} \phi\left(j - \mathbf{I}^*(\mathbf{x})\right) \quad (7)$$

where $N_\mathbf{x}$ is the number of points $\mathbf{x}$ in the region of interest (the complete image in our case). $\phi(\xi)$ is the function used to fill the histogram. Typically $p_{\mathbf{I}^*}(j)$ is incremented each time $\mathbf{I}^*(\mathbf{x}) = j$. Then $\phi(\xi) = 1$ if $\xi = 0$ and null otherwise. The equation of $p_\mathbf{I}(i)$ is similar to the one of $p_{\mathbf{I}^*}(j)$.

Since the two images are typical 8 bits images with 256 gray level values, the initial definition of mutual information is given for 256 entries for i and j. This definition gives a cost function that is subject to noise, artifacts and then local extrema that may induce issues in a non-linear optimization process [27].

Several modifications on the computation allow to have a smooth cost function with a large convergence domain. The first is to consider a smaller number of entries for each histograms. The effect is to smooth the extremum and enlarge the convergence domain. To do so, image intensities used in equation (5), (6) and (7) are scaled to fit in the new number of bin N_c. Let us note $\bar{\mathbf{I}}$ and $\overline{\mathbf{I}^*}$ respectively the scaled images $\mathbf{I}$ and $\mathbf{I}^*$:

$$\bar{\mathbf{I}}(\mathbf{x}) = \mathbf{I}(\mathbf{x})\frac{N_c - 1}{255} \qquad \overline{\mathbf{I}^*}(\mathbf{x}) = \mathbf{I}^*(\mathbf{x})\frac{N_c - 1}{255}. \quad (8)$$

The intensities of this images are no more integer values. Thus, the original ϕ function has to be modified to update the histograms entries using real values (e.g. to compute $\phi\left(i - \overline{\mathbf{I}^*}(\mathbf{x})\right)$). A solution of this problem is given by the Partial Volume Interpolation [22] that defines ϕ as a first order B-spline (corresponding to a bilinear interpolation).

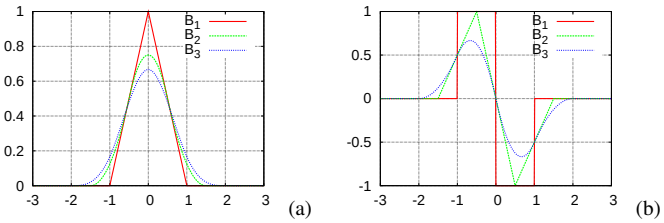

Figure 3: B-splines from order one to three (a) and their derivatives (b).

An other operation is obtained by improving this interpolation. Instead of using a simple bilinear interpolation to compute the histograms ($\phi = B_1$), B-splines of higher order are used [14]. In the present work, we consider $\phi = B_3$ (this will also be necessary to compute the second order derivatives of mutual information). B-splines functions are represented in Figure 3.

Finally, the images $\mathbf{I}$ and $\mathbf{I}^*$ can be smoothed (using for example a gaussian filter) to increase the domain of convergence by smoothing mutual information [28]. Moreover an interesting effect of the image filtering is a convexification of the cost function. The filter variance has however to be controled to keep the accuracy of the maximum.

Combining the three operations, the obtained MI is smooth with a wide and accurate maximum and thus adapted for the optimization problem that will be used in the next section. Figure 4 shows the results obtained using the comparision between the original computation ($\phi = B_1$ and 256 probability bins) and the proposed computation.

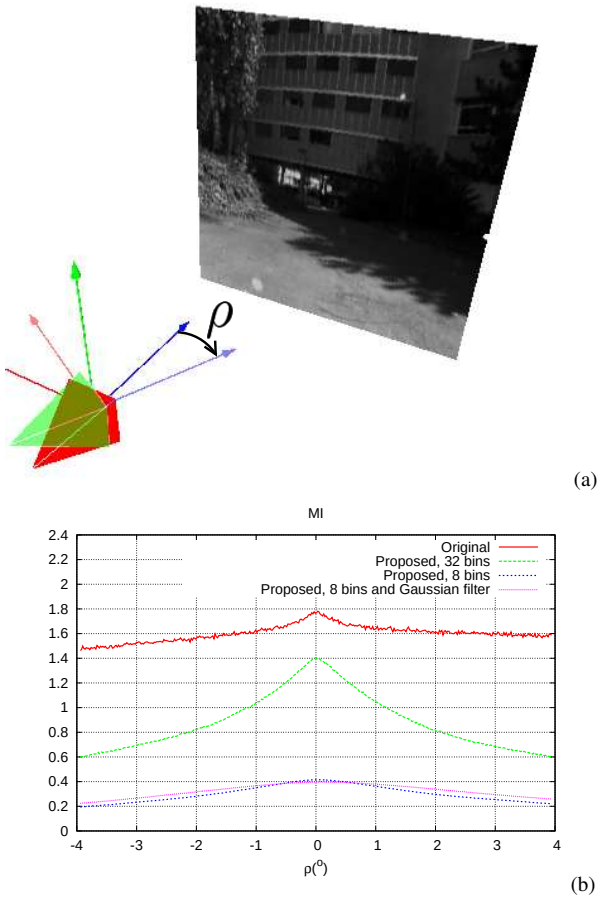

Figure 4: Method used to compute mutual information with respect to the rotation around the vertical axis ρ. (a) External view of the camera at the desired position in red (used to acquire the desired image $\mathbf{I}^*$) and at the various positions of rotation in green. (b) shows the computed mutual information between $\mathbf{I}(\rho)$ and $\mathbf{I}^*$ with respect to the rotation ρ computed using various formulation.

3.2. Navigation using Mutual information

3.2.1. Visual Servoing using Mutual information

In the general visual servoing formulation, the goal is to minimize a dissimilarity measure (generally the difference) between some desired and current features using a non linear minimization [5]. Such approaches have been already used in navigation [33][7]. Here, our goal is to propose a more direct approach that uses the image as a whole and that does not rely on geometric features and, hence, avoid the tracking and matching steps. Rather than minimizing, as usual, the error between current and desired features, the goal is to maximize the amount of information shared by the two images.

Regarding the optimization of mutual information, various approaches have been proposed: first-order gradient descent [39], multi-resolution hill climbing algorithm [36] or simulated annealing techniques [30]. Powell's [11, 22] or Simplex [25, 6] methods (which do not require function derivatives to be analytically expressed) have been very popular in MI optimization but the former is sensitive to local optima or computationally inefficient and is not adapted to our navigation problem. Considering that MI computation is evaluated from the joint image intensity histogram, an analytic derivative of the mutual information is difficult to obtain. In order to compute MI derivatives, [23] introduces partial volume interpolation for the construction of the joint histogram leading to an analytic computation of MI gradients. In [37], the authors formulate the mutual-information criterion as a continuous and differentiable function of the registration parameters using B-Spline Parzen windows. These derivatives can then be considered within a Newton like approach. Such approach has been considered here. Nevertheless we propose a dedicated control law specifically adapted to the MI cost function. We propose an inverse compositional optimization approach [2] where an important part of the required derivatives can be precomputed, resulting in small computation times. A precise, complete and efficient computation of the Hessian matrix [15] is also described.

Considering that our vehicle has a constant translational velocity $\mathbf{v}$, we will control only the vehicle steering angle ρ. The navigation task toward the next key image $\mathbf{I}^*$ can then be seen as a gradient descent optimization process where the cost function is defined as the mutual information between $\mathbf{I}$ and $\mathbf{I}^*$ wrt. the angle ρ :

$$\rho^* = \arg\max_{\rho} f(\rho) \;\text{ with }\; f(\rho) = MI(\mathbf{I}^*, \mathbf{I}(\rho)). \tag{9}$$

This maximization is performed by updating the parameter ρ to find a null derivative of mutual information using a non linear optimization. Using the parameter update $\delta\rho$, the expression to maximize can typically be rewritten as :

$$MI(\mathbf{I}^*, \mathbf{I}(\rho_{t+1})) = MI(\mathbf{I}^*, \mathbf{I}(\rho_t + \delta\rho)). \tag{10}$$

where ρ_t is the current steering angle and ρ_{t+1} is the steering angle at the next iteration. As it is demonstrated in [1], this formulation is equivalent to the inverse compositional formulation, where the expression to maximize is the following one:

$$f = MI(\mathbf{I}^*(-\delta\rho), \mathbf{I}(\rho_t)). \tag{11}$$

As it will be explained later, this formulation (also classical in visual servoing [5]) allows to precompute some terms and then have a faster computation. Using a classical Newton's method, the parameter update can be computed using:

$$\delta\rho = -\alpha \mathbf{H}_{\text{MI}}^{-1} \mathbf{L}_{\text{MI}}^{\top} \tag{12}$$

where $\mathbf{L}_{\text{MI}}$ and $\mathbf{H}_{\text{MI}}$ are respectively the interaction matrix of MI and its Hessian matrix. $\alpha \in [0, 1]$ is a scalar factor that allows to set the speed of the convergence (in our navigation task the goal is not to get to the maximum in one iteration).

Since $\mathbf{I}^*$ is now depending on ρ, the expressions defined in equations (6) and (5) depends on ρ too. The derivatives of mutual information are then given by the following equations:

$$\mathbf{L}_{\text{MI}} = \sum_{i,j} \mathbf{L}_{p_{\mathbf{II}^*}} \left(1 + \log\left(\frac{p_{\mathbf{II}^*}}{p_{\mathbf{I}^*}}\right)\right) \tag{13}$$

$$\mathbf{H}_{\text{MI}} = \sum_{i,j} \mathbf{L}_{p_{\mathbf{II}^*}}^{\top} \mathbf{L}_{p_{\mathbf{II}^*}} \left(\frac{1}{p_{\mathbf{II}^*}} - \frac{1}{p_{\mathbf{I}^*}}\right) + \mathbf{H}_{p_{\mathbf{II}^*}} \left(1 + \log\left(\frac{p_{\mathbf{II}^*}}{p_{\mathbf{I}^*}}\right)\right) \tag{14}$$

Although it is classical to consider the second term of (14) as null [17][37], in this work the computation of the exact second derivative of mutual information is used. The two previous expressions depend on the joint probability derivatives. Using equation (5), the formulation of the the joint probability $p_{\mathbf{II}^*}$ between the key image $\mathbf{I}^*$ and current image $\mathbf{I}$ and its variations are given by:

$$p_{\mathbf{II}^*}(i, j, \rho) = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \phi\left(i - \bar{\mathbf{I}}(\mathbf{x}, \rho)\right) \phi\left(j - \bar{\mathbf{I}^*}(\mathbf{x})\right)$$

$$\mathbf{L}_{p_{\mathbf{II}^*}(i,j,\rho)} = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \mathbf{L}_{\phi\left(i - \bar{\mathbf{I}}(\mathbf{x},\rho)\right)} \phi\left(j - \bar{\mathbf{I}^*}(\mathbf{x})\right)$$

$$\mathbf{H}_{p_{\mathbf{II}^*}(i,j,\rho)} = \frac{1}{N_{\mathbf{x}}} \sum_{\mathbf{x}} \mathbf{H}_{\phi\left(i - \bar{\mathbf{I}}(\mathbf{x},\rho)\right)} \phi\left(j - \bar{\mathbf{I}^*}(\mathbf{x})\right)$$

where $N_{\mathbf{x}}$ is the number of pixels considered in the images $\mathbf{I}$ and $\mathbf{I}^*$. ϕ is a B-spline function differentiable twice (see the MI definition in the previous section). The interaction matrix and Hessian of ϕ are given by:

$$\mathbf{L}_{\phi\left(i - \bar{\mathbf{I}}(\mathbf{x},\rho)\right)} = -\frac{\partial \phi}{\partial i} \mathbf{L}_{\bar{\mathbf{I}}} \tag{15}$$

$$\mathbf{H}_{\phi\left(i - \bar{\mathbf{I}}(\mathbf{x},\rho)\right)} = \frac{\partial^2 \phi}{\partial i^2} \mathbf{L}_{\bar{\mathbf{I}}}^{\top} \mathbf{L}_{\bar{\mathbf{I}}} - \frac{\partial \phi}{\partial i} \mathbf{H}_{\bar{\mathbf{I}}} \tag{16}$$

and:

$$\mathbf{L}_{\bar{\mathbf{I}}} = \nabla \bar{\mathbf{I}}\, \mathbf{L}_{\mathbf{x}} \tag{17}$$

$$\mathbf{H}_{\bar{\mathbf{I}}} = \mathbf{L}_{\mathbf{x}}{}^{\top}\, \nabla^2 \bar{\mathbf{I}}\, \mathbf{L}_{\mathbf{x}} + \nabla \bar{\mathbf{I}} \mathbf{H}_{\mathbf{x}} \tag{18}$$

where $\nabla \bar{\mathbf{I}}$ and $\nabla^2 \bar{\mathbf{I}}$ are respectivelly the gradient and the second order gradient of the image. Since the only degree of freedom considered is the rotation around the vertical axis (the y axis of the camera), the interaction and Hessian matrices are then:

$$\mathbf{L}_{\mathbf{x}} = \begin{bmatrix} -(1 + x^2) \\ -xy \end{bmatrix} \tag{19}$$

and

$$\mathbf{H}_{\mathbf{x}} = \begin{bmatrix} 2x(1 + x^2) \\ y(1 + 2x^2)] \end{bmatrix} \tag{20}$$

The final update $\delta\rho$ values have been computed on the example presented in Figure 4. Figure 5 shows the value of the derivatives of mutual information and the corresponding value $\delta\rho$ depending on the rotation between the desired and the current position. The relation between the real rotation and the computed update is quasi linear. The result of this proposed update will then cause a quasi exponential decreasing of the error, that is the ideal goal of typical visual servoing tasks [5].

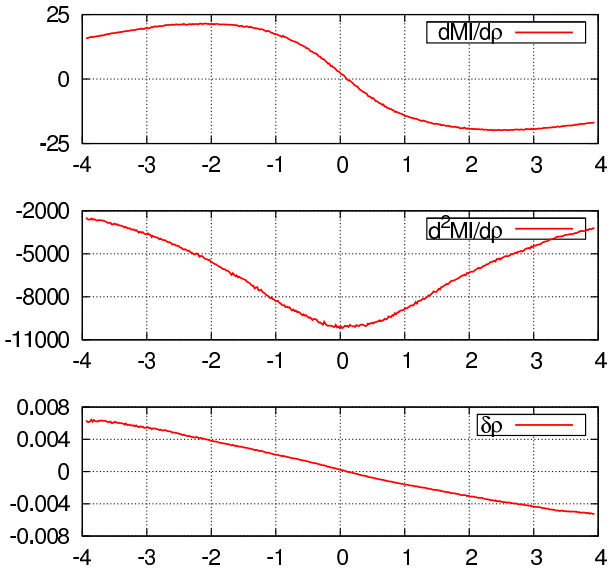

Figure 5: Derivatives of mutual information and corresponding rotational update. First row: derivative of mutual information with respect to the rotational error ρ (^{o}), second row: second derivative and third row: update $\delta\rho$.

Controlling both steering angle and translational velocity.. Increasing the number of d.o.f to be controlled is quite straightforward. In [14], experiment with 6 controlled d.o.f have been reported. In our case if two d.o.f have to be considered (translational velocity v and steering angle ρ the main difference w.r.t the presented approach is the computation of the control law that is then given by:

$$\begin{pmatrix} v \\ \delta\rho \end{pmatrix} = -\alpha \mathbf{H}_{\mathrm{MI}}^{-1} \mathbf{L}_{\mathrm{MI}}^{\top} \tag{21}$$

Such a control law would allow forward and backward motions of the vehicle (although not point stabilization). Formulation of $\mathbf{L}_{\mathrm{MI}}$ and $\mathbf{H}_{\mathrm{MI}}$ is not modified and equation (13) to (18) remain valid (although size are modified, $\mathbf{L}_{\mathrm{MI}}$ is now a 2×1 vector and $\mathbf{H}_{\mathrm{MI}}$ is a 2×2 matrix). Nevertheless it is necessary to modify equation (19) and (20). $\mathbf{L}_{\mathbf{x}}$ given by

$$\mathbf{L}_{\mathbf{x}} = \begin{bmatrix} x/Z & -(1+x^2) \\ y/Z & -xy \end{bmatrix} \tag{22}$$

and the complete derivation of the Hessian matrices can be found in [21]. Let us note that this formulation of the problem required a knowledge of depth information Z in (22). Furthermore since Z can be large in the case of outdoor scene this may lead to instabilities in the control law. This is why we have considered only the steering angle in this paper.

3.2.2. Navigation using visual servoing

For every acquisition of an image $\mathbf{I}$, an update $\delta\rho$ is computed in order to move the camera and increase the mutual information between $\mathbf{I}$ and $\mathbf{I}^*$. Using the model of the vehicle it is possible to go back to the steering angle that will give the estimated update.

Firstly the update of the rotation of the camera is linked to the camera rotational velocity $\dot{\rho}$ (around the y axis) by the following equation:

$$\dot{\rho} = \frac{\delta\rho}{\Delta t} \tag{23}$$

where Δt is the processing time (30Hz in our case).

The velocity is directly linked to the steering angle ψ of the wheels. Using the model of the non-holonomic vehicle used in our experiments (See the car-like model Figure 6) the general steering angle is computed as follows:

$$\psi = \arctan\left(\frac{L\,\dot{\rho}}{v}\right) = \arctan\left(\frac{L\,\delta\rho}{v\,\Delta t}\right) \tag{24}$$

where v is the translational velocity of the vehicle (along the z axis) and L is the distance between the front and rear wheels.

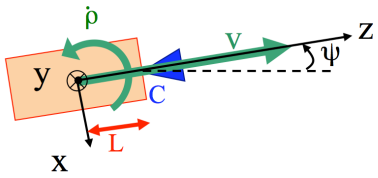

Figure 6: Model of the Cycab vehicle (Cycab can be seen on Figure 10).

4. Key images switch in the visual path

The previous section shows how to control a vehicle toward one key image using information shared by the current and key images. To be able to follow the learned trajectory a switching process between the key images of the visual path has to be defined.

4.1. Various switching solutions

Several solutions can be considered. One could propose to simply analyse the cost function evolution and check if the function (equation (9)) is increasing since mutual information is supposed to increase in the visual servoing task. If it is no longer the case it could mean that the key image is outdated. This solution is unfortunately limited to nominal or simple cases. In an outdoor environment such conditions are unpractical: if the illumination is changing then the cost function value will be affected and the desired image selection will fail.

Another solution could be to consider only the rotation required (given by the parameter update $\delta\rho$) to reach the alignment position. If the computed rotation is smaller than a given threshold it could mean that the vehicle is next to the desired position, then the next key image can be loaded. But such a simple solution will obviously fail when the tracked trajectory is a straight line.

4.2. Proposed key image selection process

The proposed approach is based on this solution coupled with a translation estimation. The key image is updated each time the remaining rotational error is low and that the translation

to reach the desired position is small (so that there is no more problem in straight lines).

The rotational error is directly given by the parameter update $\delta\rho$. However no estimation is given on the translational error (that is the remaining distance between the current position and the position corresponding to the key image).

The approach is to consider the variation of the mutual information between the key image and the current image depending on the variation of translational position t_z of the vehicle. If the variation of MI is null, it means that the maximum of MI is reached and that the robot is at the desired position. If the variation is positive (respectively negative) it means that mutual information is increasing (respectively decreasing) and that the robot is getting close to (respectively moving away from) the desired position.

This variation is simply computed as the derivation of mutual information wrt. the translational velocity v_z of the camera. The formulation of the problem is similar to the one proposed in section 3.2.1 as the difference that the current image is now depending on v_z. The derivative of the mutual information is now expressed with the interaction matrix corresponding to the translational degree of freedom that is $\mathbf{L}_{t_z} = [x/Z \quad y/Z]^t$ with Z the depth of each points. Since an accurate estimation of the translation is not needed, Z is approximated to be constant with $\bar{Z} = 20$ meters.

To validate the proposed approach, some simulations have been performed using a strongly rough environment. Figure 7 illustrates the performed experiment. The value of the derivative of mutual information is shown depending on the translation between the current and the key positions along the z axis. We can see that the choice of the depth value is not critical (in fact using the previous equations, it can be seen that changing Z is only modifying the derivative by a scale factor). Considering a strongly non flat scene, mutual information derivative with respect to the translation remains accurate with a null value when the robot reaches the desired translation.

Using two given thresholds on both the parameter update and the translation estimation allows to update the key image each time the robot is close to the current desired position.

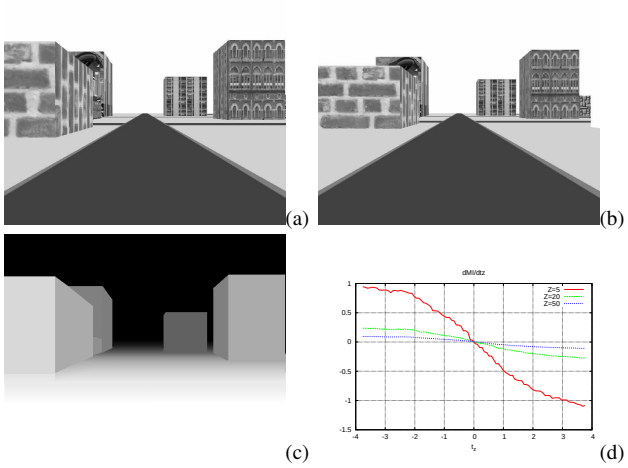

Figure 7: Translation estimation between the current and desired image. (a) Desired image, (b) acquired image with a 4 meter translation, (c) scene depth and (d) derivative of mutual information with respect to the translation along the z axis (in meters) with various fixed scene's depth Z.

5. Experimental results

This section presents navigation experiments performed with the vehicle represented in Figure 6 using the mutual information-based navigation process.

5.1. Simulation

The first experiment is a simulation that describes the behavior of the proposed navigation task in nominal conditions. Since this is a simulation, the acquired trajectory and the resulting one are perfectly known. The simulation is performed in an urban-like environment that is shown in Figure 8(a). The ground is flat and there are no occlusions nor illumination variations between the environment in the learning and in the navigation phases. The buildings of the environment have various type of textures with low, high and repetitive textures. The experiment has been done using 320×240 images. Figure 8(b) shows the trajectory used to acquire the learned sequence. The learned sequence contains 400 images on the whole trajectory and the navigation task is performed using 2500 images.

The results have been obtained using a number of histogram bins set to 8. The standard deviation of the Gaussian filter applied on the images is set to 11. The resulting trajectory is represented on Figure 8(b) overlaid with the learned trajectory.

Considering the steering angle sent to the vehicle (red plot on Figure 9), we can see that the control law is obviously not continuous. Each time the key image is changed, the computed rotation is abruptly changed and then decreases exponentially. The effect on a real vehicle may be hard to endure. To solve this issue, we propose to filter the previous result to have smoother changes of the steering angle. A simple Kalman filter with a constant velocity model has been applied to the computed angle. The result of the Kalman filter is shown on previous computed steering angles (see Figure 9). This result is adapted for the control of the vehicle that keeps on following properly the path with smoother changes in its direction.

5.2. Navigation in natural environment

The experiments have been carried out in real-time using a camera mounted on an electric car-like robot named Cycab (see Figure 10) . The presented experiments consider realistic physical paths for which no common landmarks are visible from the initial and the desired position. The navigation process has been achieved with the same vehicle/camera that were used during the learning step (construction of the visual path). Therefore potential camera distortions were not an issue in the computation of MI between two images. All computation are performed in the normalized space. Nevertheless, if two cameras are considered, then it will be necessary to undistort both images (using known/estimated calibration parameters).

The mutual information navigation scheme has been tested on a non-holonomic vehicle (see Figure 6) in an outdoor envi-

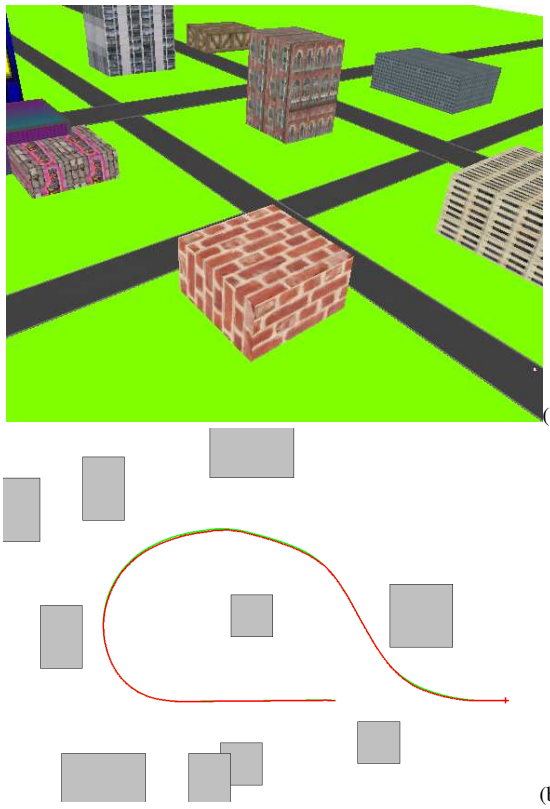

Figure 8: Simulation experiment. (a) Aerial view of the environment, (b) 2D representation of map with the learned trajectory in green and the resulting path in red (gray rectangles are buildings).

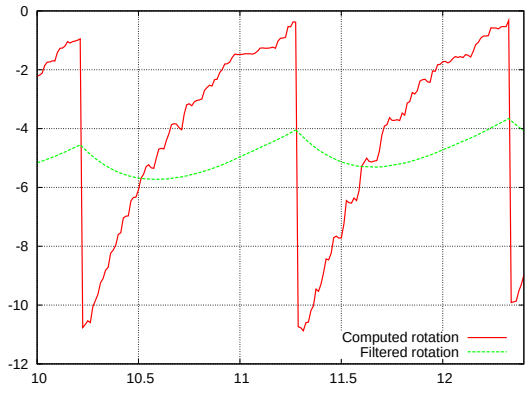

Figure 9: Simulation experiment: steering angle (in degrees) in the first turn of the path with respect to the time (in second). The computed steering angle in red shows two exponential decreasement corresponding to two visual servoing tasks, the filtered steering angle is in green.

Figure 10: The Cycab vehicle considered in this experiment.

ronment. The final approach presented in the previous paragraph has been used. Let us emphasize that the vehicle is equipped with a monocular camera and that no other sensor such as GPS, radar or odometry are considered in these experiments. Furthermore, the 3D structure of the scene remains fully unknown during the learning and navigation steps.

Aerial views of the environment, where the navigation task takes place, are shown in Figure 11 along with the considered trajectory (about 400 meters). As seen on the pictures, the environment is semi-urban with both trees and buildings (with windows acting as repetitive textures). Let us note that the vehicle crosses a covered parking lot (green part of the trajectory in Figure 11) and that the ground is no longer perfectly flat (mainly in the first 100 meters of the trajectory).

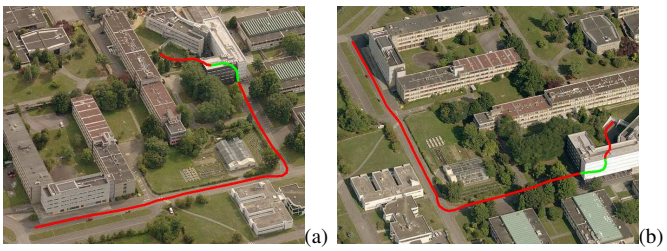

Figure 11: Outdoor environment with an approximation of the learned trajectory (the green part is the trajectory executed under a covered parking lot). (a) First aerial view, (b) second aerial view.

When learning the path the vehicle is manually driven at a roughly constant velocity. For this experiment we consider 1200 key images (that is around three key images per meter). The navigation task itself is carried out at 0.5 m/s. Images are acquired at 30Hz (nearly 25.000 images are acquired and processed in real-time during this navigation task).

Some pictures of one navigation task are shown in Figure 12. By comparing the current and key images (and the image error on the third row), we can see that the robot is qualitatively (as defined in [29]) following the same path. The navigation task has been tested with both cloudy and sunny weather using the same learned visual path). Since time had passed between the acquisition of the visual path and the navigation task, there have been very large illumination changes between the current and the key images as it is highlighted in Figure 14(a). The

task has even been tested with a ground recovered with snow still using the same initial visual path (See Figure 14(b)). Despite those illumination variations the navigation task was still converging. That shows the robustness of the proposed control law to illumination variations and the efficiency of considered mutual-information similarity criterion to perturbation. Let us also note that the road on which experiments have been conducted is not completely flat. As can be seen on the video, just entering under the cover parking, the road features a quite important slope (at least 1 meter down in 5 meters), the camera is then pitching at this point.

Since no obstacle avoidance process is considered, the navigation task has been performed in quiet conditions. Nevertheless several vehicles have overtaken our experimental vehicle and appeared in the camera view. Despite this perturbation, and thanks to the robustness of the similarity criterion, the navigation task has never failed showing the robustness of mutual information to occlusions. One of these moments is represented in Figure 14(c) (the van in the current image was not present in the key image). The complete video is provided with this paper. Let us note that MI between the current image and the next key image can be monitored in order to detect failures (deviation from the visual path). In nominal conditions (few scene modification), if this MI is smaller than the MI between two successive key frames it is possible that the robot deviates from its visual path and security actions has to be considered.

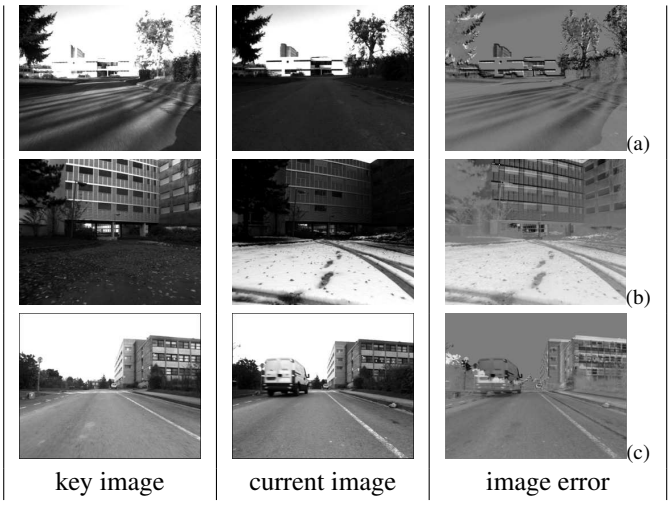

Figure 14: Mutual information robustness. (a) robutness to illumination variations, (b) illumination variations and snow on the ground, (c): robustness to occlusions. First column: desired image, second column: current image and third column: difference of the current and desired images.

6. Conclusion

In this paper, we have presented a new way to achieve image based navigation task. We show that our vehicle is able to track a previously learned trajectory using only the information provided by a monocular camera. The navigation task can be achieved despite important variation in the lighting condition and possible perturbations. This can be achieved thanks to various elements related to the use of mutual information which is new in vision-based control:

- Our approach does not rely on features extracted from the image. Therefore we do not need to track or match features (eg, keypoints) which has proved to be a difficult and not always reliable process. Furthermore no 3D information related to the scene structure is required.
- To avoid this tracking and matching processes, the vision-based control law of the non-holonomic vehicle is directly linked to the optimization of a similarity criterion based on the information shared by two images.
- We propose a control law that directly links the vehicle motion to the variation of the mutual information. The proposed approach that considers a derivation of the MI, up to the second order, allows a large convergence domain along with fast (video rate) computation.
- Considering information contained in the images and not features extracted from the image or the image intensities induces a natural robustness to perturbation that is essential in our navigation context.

We also considered a key images selection process which is efficient regarding the considered navigation taks. Future work is planned on the post processing of the steering angle update. For the moment, it is simply filtered, while using the knowledge on the vehicle kinematics and smoothness of the trajectory could improve the navigation task. Moreover, future work will devoted to the definition of efficient navigation task that require more degrees of freedom and more complex control models such as aerial drones.

Acknowledgment

This work was supported by DGA under contribution to student grant.

References

[1] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition, CVPR'01*, pages 1090 – 1097, December 2001.

[2] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *Int. Journal of Computer Vision*, 56(3):221–255, 2004.

[3] O. Booij, B. Terwijn, Z. Zivkovic, and B. Krose. Navigation using an appearance based topological map. In *IEEE Int. Conf. on Robotics and Automation, ICRA 2007*, pages 3927 –3932, april 2007.

[4] D. Burschka and G. Hager. Vision-based control of mobile robots. In *IEEE Int. Conf. on Robotics and Automation, ICRA'2001*, volume 2, pages 1707–1713, 2001.

[5] F. Chaumette and S. Hutchinson. Visual servo control, Part I: Basic approaches. *IEEE Robotics and Automation Magazine*, 13(4):82–90, December 2006.

[6] H. Chen, P.K. Varshney, and M.-A. Slamani. On registration of regions of interest (roi) in video sequences. In *IEEE Conf. on Advanced Video and Signal Based Surveillance.*, pages 313 – 318, July 2003.

[7] Z. Chen and S.T. Birchfield. Qualitative vision-based path following. *IEEE Transactions on Robotics*, 25(3):749–754, 2009.

Figure 12: Outdoor navigation experiment, begining of the experiment. First row: current image acquired by the vehicle, second row: desired image and third row: difference between the current and desired image. The first top column and the last bottom column show respectively the first images and the last images used in the navigation task.

Figure 13: Outdoor navigation experiment, end of the experiment.

[8] L.A. Clemente, A.J. Davison, I.D. Reid, J. Neira, and J.D. Tardós. Mapping large loops with a single hand-held camera. In *Robotics: Science and Systems*, Atlanta, Georgia, June 2007.

[9] C. Collewet and E. Marchand. Photometric visual servoing. *IEEE Trans. on Robotics*, 27(4):828–834, August 2011.

[10] C. Collewet, E. Marchand, and F. Chaumette. Visual servoing set free from image processing. In *IEEE Int. Conf. on Robotics and Automation, ICRA'08*, pages 81–86, Pasadena, CA, May 2008.

[11] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multi-modality image registration based on information theory. In *Int. Conf. on Information Processing in Medical Imaging, IPMI'95*, Ile de Berder, France, June 1995.

[12] J. Courbon, Y. Mezouar, and P. Martinet. Autonomous navigation of vehicles from a visual memory using a generic camera model. *IEEE Trans. on Intelligent Transportation Systems*, 10(3):392–402, 2009.

[13] A. Dame and E. Marchand. Entropy-based visual servoing. In *IEEE Int. Conf. on Robotics and Automation, ICRA'09*, pages 707–713, Kobe, Japan, May 2009.

[14] A. Dame and E. Marchand. Mutual information-based visual servoing. *IEEE Trans. on Robotics*, 27(5):958–969, October 2011.

[15] A. Dame and E. Marchand. Second order optimization of mutual information for real-time image registration. *IEEE Trans. on Image Processing*, 21(9):4190–4203, September 2012.

[16] A. Diosi, A. Remazeilles, S. Segvic, and F. Chaumette. Outdoor visual path following experiments. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS'07*, pages 4265–4270, San Diego, CA, October 2007.

[17] N.D.H. Dowson and R. Bowden. A unifying framework for mutual information methods for use in non-linear optimisation. In *European Conference on Computer Vision, ECCV'06*, volume 1, pages 365–378, June 2006.

[18] U. Frese and L. Schröder. Closing a million-landmarks loop. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems, IROS'06*, pages 5032–5039, Beijing, China, October 2006.

[19] P. Furgale and T. Barfoot. Stereo mapping and localization for long-range path following on rough terrain. In *IEEE Int. Conf. on Robotics and Automation, ICRA 2010*, pages 4410 –4416, May 2010.

[20] P. Furgale and T. Barfoot. Visual path following on a manifold in unstructured three-dimensional terrain. In *IEEE Int. Conf. on Robotics and Automation, ICRA 2010*, pages 534 –539, May 2010.

[21] J.-T. Lapresté and Y. Mezouar. A Hessian approach to visual servoing. In *IEEE/RSJ Int. Conf. on Intelligent Robots and System, IROS'04*, pages 998–1003, Sendai, Japan, October 2004.

[22] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE trans. on Medical Imaging*, 16(2):187–198, 1997.

[23] F. Maes, D. Vandermeulen, and P. Suetens. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Medical Image Analysis*, 3(4):373–386, 1999.

[24] Y. Matsumoto, M. Inaba, and H. Inoue. Visual navigation using view-sequenced route representation. In *IEEE Int. Conference on Robotics and Automation, 1996*, pages 83–88, 1996.

[25] C. Meyer, J. Boes, B. Kim, P. Bland, K. Zasadny, P. Kison, K. Koral, K. Frey, and R. Wahl. Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations. *Medical Image Analysis*, 1(3):195 – 206, 1997.

[26] T. Ohno, A. Ohya, and S. Yuta. Autonomous navigation for mobile robots referring pre-recorded image sequence. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, IROS'96*, volume 2, pages 672–679, November 1996.

[27] G. Panin and A. Knoll. Mutual information-based 3D object tracking. *Int. Journal of Computer Vision*, 78(1):107–118, 2008.

[28] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Trans. on Medical Imaging*, 22(8):986–1004, August 2003.

[29] A. Remazeilles and F. Chaumette. Image-based robot navigation from an image memory. *Robotics and Autonomous Systems*, 55(4):345–356, April 2007.

[30] N. Ritter, R. Owens, J. Cooper, R.H. Eikelboom, and P.P. Van Saarloos. Registration of stereo and temporal images of the retina. *IEEE Transactions on Medical Imaging*, 18(5):404–418, May 1999.

[31] E. Royer, M. Lhuillier, M. Dhome, and J.M. Lavest. Monocular vision for mobile robot localization and autonomous navigation. *International Journal of Computer Vision*, 74(3):237–260, 2007.

[32] S. Se, D. Lowe, and J. Little. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *The Int. Journal of Robotics Research*, 21(8):735, 2002.

[33] S. Segvic, A. Remazeilles, A. Diosi, and F. Chaumette. A mapping and localization framework for scalable appearance-based navigation. *Computer Vision and Image Understanding*, 113(2):172–187, February 2009.

[34] CE. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27:379–423, 623–656, 1948.

[35] R. Sim and G. Dudek. Mobile robot localization from learned landmarks. In *IEEE/RSJ Conference on Intelligent Robots and Systems, IROS'98*, Victoria, BC, 1998.

[36] C. Studholme, D.L.G. Hill, and D. J. Hawkes. Automated 3D registration of truncated mr and ct images of the head. In *British Machine Vision Conference, BMVC'95*, pages 27–36, Birmingham, Surrey, UK, September 1995.

[37] P. Thévenaz and M. Unser. Optimization of Mutual Information for Multiresolution Image Registration. *IEEE Trans. on Image Processing*, 9(12):2083–2099, 2000.

[38] P. Viola and W. Wells. Alignment by maximization of mutual information. In *Int. Conf. on Computer Vision, ICCV'95*, Washington, DC, 1995.

[39] P. Viola and W. Wells. Alignment by maximization of mutual information. *Int. Journal of Computer Vision*, 24(2):137–154, 1997.

[40] A. Zhang and L. Kleeman. Robust appearance based visual route following for navigation in large-scale outdoor environments. *The International Journal of Robotics Research*, 28(3):331–356, March 2009.

Amaury Dame. graduated from the Institut National des Sciences Appliquées de Rennes in 2007. He received the Master's degree in signal and image processing and a PhD from the Université de Rennes 1 respectively in 2007 and december 2010. Since 2011, he is now with the Active Vision Group, Department of Engineering Science, University of Oxford. His research interests include computer vision, robotics, visual servoing and SLAM.

Eric Marchand. is professor of Computer Science at Université de Rennes 1 in France and a member of the INRIA/IRISA Lagadic team at IRISA-INRIA Rennes. He received the Ph.D degree and the "Habilitation à Diriger des Recherches" in Computer Science from the University of Rennes in 1996 and 2004 respectively. He spent one year as a Postdoctoral Associates in the AI lab of the Dpt of Computer Science at Yale University in 1997. He has been an INRIA research scientist from 1997 to 2009. His research interests include robotics, visual servoing, real-time object tracking and augmented reality. Since 2010 he is an associate editor for IEEE Trans. on Robotics.